Investigation of X-ray induced radiation damage in proteins, nucleic acids and their complexes



Charles Simon Bury

Linacre College

University of Oxford

A thesis submitted in partial fulfilment of the degree of

Doctor of Philosophy

Trinity 2017

supervised by

Prof. Elspeth F. Garman

Abstract

Macromolecular X-ray crystallography (MX) is currently the dominant technique for the structural elucidation of macromolecules at near atomic resolution. However, the progression and deleterious effects of radiation damage remains a major limiting factor in the success of diffraction data collection and subsequent structural solution at modern third generation synchrotron facilities. For experiments conducted at 100 K, protein specific damage to particular amino acids has been widely reported at doses of just several MGy, before any observable decay in average diffraction intensities. When undetected, such artefacts of X-ray irradiation can lead to significant modelling errors in protein structures, and ultimately the failure to derive the correct biological function from a model. It is thus vital to develop tools to help MX experimenters to detect and correct for such damage events. This thesis presents the development of an automated program, *RIDL*, which is designed to objectively quantify radiation-induced changes to electron density at individual atoms, based on $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps between different dose states for a single crystal. The high-throughput *RIDL* program developed in this work provides the ability to systematically investigate a wide range of macromolecular systems.

To date, damage to the broad class of nucleic acids and nucleoprotein complexes has remained largely uncharacterised, and it is unclear how radiation damage will disrupt the validity of such models derived from MX experiments. This thesis presents the first systematic investigations on a range of nucleic acid, protein-RNA and protein-DNA complex case studies. In general, it is concluded that nucleic acids are highly robust to radiation damage effects at 100 K, relative to control protein counterparts across the tested systems.

For protein crystals at 100 K, cleavage of the phenolic C−O bond in tyrosine has disseminated through the MX radiation damage literature as a dominant specific damage event at 100 K, despite the absence of any energetically favourable cleavage mechanism. To clarify the radiation susceptibility of tyrosine, this thesis presents a systematic investigation on radiation damage to tyrosine in a wide range of MX protein radiation damage series retrieved from the Protein Data Bank. It is concluded that the tyrosine C−O bond remains intact following X-ray irradiation, however the aromatic side-group can undergo radiation-induced displacement.

This thesis also presents further applications of the *RIDL* program. A protocol is introduced to calculate explicit half-dose values for the electron density at individual atoms to decay to half of their initial value at zero absorbed dose. In addition, a methodology is developed to detect radiation-induced changes to electron density occurring over the course of the collection of a single MX dataset of diffraction images, all of which are required for structural solution. These protocols aim to advise experimenters of when previously-undetected site-specific damage effects may have corrupted the quality of their macromolecular model.

Overall, the work in this thesis is highly applicable to both the future understanding of radiation damage in macromolecular structures, as well as of interest to the wider crystallographic community.

Acknowledgements

First and foremost, I would like to thank my D.Phil supervisor Prof. Elspeth Garman, without whom I don't think I would have reached the position I am in today. I am ever so grateful for her supervision over the past few years; she has always stuck by me, and through her mentoring I now feel much more confident in my own ability.

Jonny Brooks-Bartlett has been my academic role model in these past few years. I have thoroughly enjoyed working at the desk besides him and the continuous sharing of ideas that this has entailed. I do not think I have ever met someone quite as charismatic when it comes to public speaking - I will always think "what would Jonny do?" when it comes to writing my own presentations. Whereas Jonny has been an academic father, Markus Gerstel has most definitely been the grandfather figure. From my first days in the Garman group, his own research and coding brilliance were what initially inspired me to work on my programming skills and has played a major contributing factor in the direction in which I have progressed my research. Kathryn Shelley has always known how to truly get me thinking when it comes to science. She is one of the brightest, yet sincere, people that I know, and I am truly grateful that I have had the opportunity to work with her at Oxford. I would also like to thank two great collaborators: Prof. John McGeehan, University of Portsmouth, and Prof. Ian Carmichael, University of Notre Dame. They have both given me so much of their time to help with my own scientific progression, and I owe any success to them.

Connor Sharp has been a best friend at Oxford. Since our first days at the Doctoral Training Centre (DTC), Connor has been a major influence, both academically and recreationally. We must surely have drank over a thousand Biochemistry cafe coffees by this point whilst talking about *Python* alone. I also want to especially thank Alexandra Pons for always supporting me throughout my time in Oxford. Thank you for always believing in me; you have truly given me the happiness and laughter to drive me to succeed. I would also like to give an enormous thanks to my family: my parents, Katrina and Simon, and my sister, Natasha - you guys are seriously the best! You have always been there me for through the good times and the bad, and I cannot express how much I am grateful for your support and love. I could not have done this without you all.

Finally, I am grateful to the UK's Engineering and Physical Sciences Research Council (EPSRC) for funding through a studentship at the Systems Biology programme at the University of Oxford's Doctoral Training Centre. Further funding has been provided by the Ninth International Workshop on X-ray Damage to Biological Crystalline Samples, the Keith Prout Crystallography Fund, the Oxford University Department of Biochemistry Conference Fund, and the annual CCP4 Study Weekend fund.

Associated publications and presentations

The following is a comprehensive list of publications and oral presentations associated with work presented in this thesis:

- A subset of the work presented in Section 2.4 and Section 3.2 has been published in Bury *et al.* (2015).
- A subset of the work presented in Section 2.5 and Chapter 4 has been published in Bury *et al.* (2016b).
- A review based on work presented in Chapter 3 and Chapter 4 has been published in Bury *et al.* (2016a).
- A subset of the work presented in Chapter 6 has been published in Bury *et al.* (2017).
- Work presented in Chapter 2 and 4 has been presented by the author in an oral presentation at the 66th annual American Crystallographic Association (ACA) meeting, Denver, 2016. The author was awarded the Student Etter award for Light Sources for this talk, presented by Prof. Gerd Rosenbaum, University of Georgia.
- Work presented in Chapter 2 and 7 has been presented by the author in a short oral presentation at the 67th annual ACA meeting, New Orleans, 2017.

Poster presentations presented by the author can be found in Appendix B. Journal front covers designed by the author have been presented in Appendix C. Each of the aforementioned publications are also located in Appendix D.

Abbreviations

CI ..... Confidence Interval

COD ..... Crystallography Open Database

DEA ..... Dissociative Electron Attachment

DFT ..... Density Functional Theory

DLS ..... Diamond Light Source

DNA ..... Deoxyribonucleic acid

DSB ..... Double Strand Breakage

dsDNA ..... Double stranded Deoxyribonucleic Acid

DWD ..... Diffraction-Weighted Dose

EDS ..... Electron Density Server

END ..... Electron Number Density (map)

ESR ..... Electron Spin Resonance

ESRF ..... European Synchrotron Radiation Facility

FWHM ..... Full Width Half Maximum

GFP ..... Green Fluorescent Protein

GUI ..... Graphical User Interface

HEWL ..... Hen/Chicken Egg White Lysozyme

KDE ..... Kernel Density Estimation

KS ..... Kolmogorov-Smirnov

LEE ..... Low Energy Electron

LUMO ..... Lowest Unoccupied Molecular Orbital

MAD ..... Multi-wavelength Anomalous Dispersion/Diffraction

MES ..... 2-(N-morpholino)ethanesulfonic acid

microPIXE ..... micro-Particle-Induced X-ray Emission

MIR ..... Multiple Isomorphous Replacement

MX ..... Macromolecular X-ray Crystallography

NCS ..... Non-Crystallographic Symmetry

PDB ..... Protein Data Bank

RIP ..... Radiation-damage Induced Phasing

RIPAS ..... Radiation-damage Induced Phasing with Anomalous Scattering

R-M ..... Restriction-Modification System

RNA ..... Ribonucleic Acid

RSCC ..... Real-Space Correlation Coefficient

RSD ..... Relative Standard Deviation

RSZD ..... Real-Space Difference Density Z-score

RT ..... Room Temperature

SAD ..... Single-wavelength Anomalous Dispersion

SIR ..... Single Isomorphous Replacement

SSB ..... Single Strand Breakage

***Tc*AChE** ..... *Torpedo californica* acetylcholinesterase

TRAP ..... Tryptophan RNA-binding Attenuation Protein

TRX ..... Time-Resolved X-ray crystallography

VDW ..... Van der Waals

XANES ..... X-ray Absorbance Near-Edge Spectrum

XAS ..... X-ray Absorption Spectroscopy

XFEL ..... X-ray Free Electron Laser

Contents

1 Introduction **27**

1.1 The macromolecular X-ray crystallographic pipeline 28

1.1.1 Crystallogenesis 29

1.1.2 Diffraction data collection 30

1.1.3 Data reduction and radiation damage 39

1.1.4 Phasing and radiation damage 47

1.1.5 Model building and refinement 50

1.2 Specific damage in protein crystallography 52

1.2.1 Specific damage observations in proteins 52

1.2.2 Specific damage mechanisms 55

1.2.3 Factors influencing specific damage events 59

1.2.4 Exploitation of specific radiation damage 63

1.3 The aims of this thesis 64

2 Development of tools to investigate site-specific radiation damage **65**

2.1 Introduction 66

2.2 Outline of chapter 69

2.3 Glycoside hydrolase family 7 data collection 71

2.3.1 Crystallisation 71

2.3.2 X-ray data collection 72

2.3.3 Diffraction data processing 72

2.3.4 Dose calculations 77

2.4 Facilitated manual inspection of difference maps 77

2.4.1 Generation of Fourier difference maps 80

2.4.2 Difference peak detection 82

2.4.3 Filtering of peaks 83

2.4.4 Facilitated manual inspection 85

2.4.5 Assessment of manual inspection quality 91

2.4.6 Further filtering of peaks 94
2.4.7 The output of *PEAKS* 94
2.5 Automated detection of damage in difference maps 95
2.5.1 The availability of other programs and motivation for *RIDL* 97
2.5.2 Construction of the *RIDL* pipeline 99
2.5.3 Data preparation for the *RIDL* pipeline 101
2.5.4 $F_{obs,n} - F_{obs,1}$ difference map generation 106
2.5.5 Assignment of a search radius per-atom 109
2.5.6 Calculation of per-atom summary metrics 114
2.5.7 Density-weighted metrics 116
2.5.8 The C_α normalisation scheme 122
2.5.9 Testing of the *RIDL* pipeline 127
2.6 Discussion 140

3 Radiation damage to nucleic acids in complex with proteins. Part I: C-protein DNA complexes 143

3.1 Introduction 144
3.1.1 Aims of the current chapter 147
3.2 The C.Esp1396I-DNA complex 148
3.2.1 Crystallisation 150
3.2.2 X-ray data collection 151
3.2.3 Dose calculation with *RADDOSE-3D* 153
3.2.4 Diffraction data processing 154
3.2.5 $F_{obs,n} - F_{obs,1}$ map analysis 157
3.2.6 Site specific damage observations using $F_{obs,n} - F_{obs,1}$ maps 162
3.2.7 Mean isotropic *B*-factor analysis 174
3.3 The C.Csp 21 bp DNA complex 178
3.3.1 Expression, purification and crystallisation 178
3.3.2 X-ray data collection 179

3.3.3 Dose calculation with *RADDOSE-3D* . . . 180
3.3.4 Diffraction data processing . . . 181
3.3.5 Site-specific damage observations using $F_{obs,n} - F_{obs,1}$ maps . . . 183
3.4 Conclusion . . . 190

4 Radiation damage to nucleic acids in complex with proteins. Part II: The TRAP-RNA complex 195

4.1 Introduction . . . 196
4.2 Materials and methods . . . 198
4.2.1 Crystallisation and X-ray data collection . . . 198
4.2.2 Dose calculation . . . 202
4.2.3 Data processing . . . 203
4.2.4 Generation of Fourier difference maps . . . 205
4.3 Results . . . 207
4.3.1 Protein site-specific damage observations . . . 207
4.3.2 RNA is less susceptible to electron-density loss than protein . . . 208
4.3.3 Distribution of damage within the TRAP-RNA complex: investigation of Glu and Asp residues . . . 210
4.3.4 RNA protects residues at the RNA-binding interfaces . . . 214
4.3.5 The role of crystal contacts . . . 223
4.4 Discussion . . . 226

5 Characterisation of site-specific radiation damage in DNA 231

5.1 Introduction . . . 232
5.2 Materials and methods . . . 235
5.2.1 Crystallisation . . . 235
5.2.2 X-ray data collection . . . 236
5.2.3 Diffraction data processing . . . 237
5.2.4 Dose calculation . . . 245
5.2.5 MicroPIXE experiment . . . 246

5.3 Radiation damage analysis . . . 248
5.3.1 Global radiation damage effects with dose . . . 248
5.3.2 Site-specific damage observations . . . 253
5.3.3 Relationship between phosphate disordering and DNA geometry . . . 262
5.4 Conclusion . . . 267

6 Characterisation of site-specific radiation damage to Tyrosine 271
6.1 Introduction . . . 272
6.2 Reanalysis of the myrosinase damage series (Burmeister, 2000) . . . 275
6.2.1 Acquisition of the MX damage series data . . . 275
6.2.2 Evaluation of myrosinase model building quality . . . 276
6.2.3 Dose recalculation . . . 277
6.2.4 Evaluation of recalculated $F_{obs,n} - F_{obs,1}$ maps . . . 279
6.3 Investigation of radiation-induced trends between proteins . . . 282
6.3.1 Acquisition of MX damage series data from *PDB_REDO* . . . 284
6.3.2 Damage comparison strategies between deposited structures . . . 287
6.4 A hypothesis for the observed Tyr electron density loss . . . 295
6.4.1 Comparison of damage signatures between residues . . . 295
6.4.2 Tyr O_η versus C_ζ correlation analysis . . . 299
6.4.3 Occupancy refinement of Tyr groups in myrosinase . . . 300
6.4.4 Simulation of Tyr damage with $F_{calc} - F_{calc}$ maps . . . 306
6.5 Automated approaches to detection of radiation-induced motion . . . 307
6.5.1 Automatic detection with *RINGER* . . . 308
6.5.2 Automatic detection of electron density shifts with *RIDL* . . . 311
6.6 Factors affecting the Tyr −OH group . . . 313
6.6.1 Availability of hydrogen bonds . . . 313
6.6.2 Dependence on disulphide bond damage . . . 317
6.7 Conclusion . . . 317

7 Further methods for site-specific damage detection 323

7.1 Introduction . . . 324

7.1.1 Specific damage half-doses . . . 324

7.1.2 Specific damage detection in single MX datasets . . . 325

7.2 Specific damage half dose model . . . 325

7.2.1 The protocol . . . 326

7.3 How much diffraction data are required to detect site-specific damage? . . . 347

7.3.1 The test proteins . . . 347

7.3.2 A protocol to screen for damage in a single MX dataset . . . 349

7.3.3 Damage detection to individual atoms . . . 352

7.3.4 Assessment of damage to overall structure . . . 359

7.3.5 Extension to other protein datasets . . . 362

7.4 Discussion . . . 362

8 Conclusion 365

A Bovine insulin data collection 397

B Poster presentations 401

C Journal covers 407

D Publications 411

List of Figures

1.1 The primary X-ray interaction processes 32

1.2 The cross sections for elastic, Compton and photoelectric effect 33

1.3 The number of MX structures solved at 100 K and RT each year 37

1.4 The approximate time for a typical protein crystal to reach the 30 MGy experimental dose limit (Owen *et al.*, 2006) for varying dose rates 40

1.5 An example change to a diffraction pattern with dose 43

1.6 The I_D/I_1, $B_{rel}(D)$ and $R_d(\Delta D)$ global damage metrics 46

1.7 Classical protein specific damage illustrated with $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps 54

1.8 The decomposition pathway for the radiolysis of H_2O 56

1.9 Schematic for electron and hole transfer along protein backbones 57

2.1 Collection of a 'typical' MX damage series from a single protein crystal 67

2.2 The variation in the number of detected peaks with dose for the De la Mora *et al.* (2011) lysozyme damage series 69

2.3 GH7 enzyme crystal in loop at DLS beamline I02 73

2.4 A flow chart illustrating assisted manual peak inspection using *PEAKS* 81

2.5 The dependence of the *PEAKMAX* peak detection on the lower density threshold cut-off 83

2.6 The definition of the *PEAKS* per-atom search radius 84

2.7 The distribution of peaks filtered at each stage of the *PEAKS*protocol 86

2.8 Damaged atom types detected by *PEAKS* for the GH7 family protein 89

2.9 The number of damaged atoms identified by *PEAKS* for the GH7 family protein 90

2.10 The determination of the onset of noise in *PEAKS* 92

2.11 The distribution of atom-peak distances detected by *PEAKS* 93

2.12 The top 10 recurrent damage sites determined by *PEAKS*for the GH7 family protein 95

2.13 An illustration of $F_{\text{obs,2}} - F_{\text{obs,1}}$ difference map behaviour localised to example disulphide bonds in the NCS-related chains of the GH7 family protein 96

2.14 The dependence of the RSZD− metric in *EDSTATS* on the refinement strategy for high dose diffraction datasets 98

2.15 The dependence of the RSZD metric in *EDSTATS* on the quality of the refined coordinate model 100

2.16 A flow chart illustrating the *RIDL* pipeline 102

2.17 An example `.txt` input file for the *RIDL* pipeline 103

2.18 A schematic illustrating the per-atom search radius defined for *RIDL* 104

2.19 The dependence of the C_α-normalised D_{loss} metric on the chosen *SCALEIT* scaling protocol 108

2.20 Schematic illustrating the role of *SFALL* in the *RIDL* pipeline 110

2.21 The effect of variation of the *SFALL* program α parameter on atomic *B*-factor . 111

2.22 The dependence of the C_α-normalised D_{loss} metric on the *SFALL* α parameter . 112

2.23 A comparison of the *PEAKS* output with the *RIDL* D_{loss} metric 117

2.24 The variation of *RIDL* metric values with dose for the top 10 damage sites in the GH7 family protein 118

2.25 A comparison of metric values output from *RIDL* for the GH7 family protein . . . 119

2.26 A schematic to illustrate the effect of weighting each $F_{obs,n} - F_{obs,1}$ map voxel by the magnitude of the ρ_{calc} density map at each voxel 120

2.27 A comparison of the standard and *density-weighted* metric values output from *RIDL* for the GH7 family protein 123

2.28 A comparison of the metrics output from *RIDL* for selected atoms in the GH7 family protein 124

2.29 The distribution of D_{loss} and $D_{loss}^{C_\alpha}$ metric values for the GH7 family protein case study 128

2.30 A comparison of D_{loss} and $D_{loss}^{C_\alpha}$ metric values for the GH7 family protein case study 129

2.31 The dependence of the C_α-normalised D_{loss} metric on the $F_{obs,n} - F_{obs,1}$ map resolution 131

2.32 The dependence of *RIDL* metrics on the unit cell dimensions used to calculate each $F_{obs,n} - F_{obs,1}$ map 132

2.33 The dependence of *RIDL* metrics on the chosen higher dataset refinement strategy 136

2.34 The dependence of the C_α-normalised D_{loss} metric on the choice of phase set used to calculate each $F_{obs,n} - F_{obs,1}$ map 138

2.35 The dependence of the C_α-normalised D_{loss} metric on the $F_{obs,n} - F_{obs,1}$ map phase quality 139

3.1 The mechanism by which C.Esp1396I protein regulates the *Esp1396I* bacterial R-M system 149

3.2 The C.Esp1396I-DNA tetrameric complex 150

3.3 The C.Esp1396I-DNA complex crystal in a cryoloop at beamline ID29, ESRF . . 152

3.4 The dose distributions calculated with *RADDOSE-3D* for the C.Esp1396I-DNA complex crystal 153

3.5 The I_n/I_1 mean intensity decay with dose plot for the C.Esp1396I-DNA damage series 155

3.6 An illustration of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map over the full C.Esp1396I-DNA complex across the collected dose range 160

3.7 The reduction in detected $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks for the C.Esp1396I-DNA damage series following facilitated manual inspection using *PEAKS* 162

3.8 The spatial distribution of specific damage sites detected with *PEAKS* throughout the C.Esp1396I-DNA complex 163

3.9 Examples of protein damage sites in the C.Esp1396I-DNA complex detectable in $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps 164

3.10 The frequency of protein specific damage sites for the C.Esp1396I-DNA complex detected by *PEAKS* 165

3.11 An illustration of the unexplained $F_{\text{obs,n}} - F_{\text{obs,1}}$ map negative density near the C8 atom of A-10 in the C.Esp1396I-DNA complex 166

3.12 The frequency of DNA specific damage sites for the C.Esp1396I-DNA complex detected by *PEAKS* 167

3.13 DNA damage sites in the C.Esp1396I-DNA complex detectable in $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps 168

3.14 Further DNA damage sites in the C.Esp1396I-DNA complex detectable in $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps 170

3.15 The distribution of C_α-normalised D_{loss} values for protein and DNA components of the C.Esp1396I-DNA complex 172

3.16 The *asymmetry* score calculated for the C_α-normalised D_{loss} metric distributions in Figure 3.15 as a function of dose 174

3.17 The distribution of C_α-normalised D_{loss}(atom) values across the C.Esp1396I-DNA complex 175

3.18 The major and minor groove widths calculated for the 35 bp DNA sequence in the C.Esp1396I-DNA complex 176

3.19 The dose-dependence of the mean isotropic *B*-factor for each chain of the C.Esp1396I-DNA complex 177

3.20 The C.Csp231I-DNA complex crystal in a cryoloop at beamline ID14-4, ESRF . . 180

3.21 I_n/I_1 versus dose for the C.Csp231I-DNA complex damage series 182

3.22 An illustration of the $F_{obs,n} - F_{obs,1}$ map over the full C.Csp231I-DNA complex across the collected dose range 186

3.23 Examples of suspected damage sites to protein in the C.Csp231I-DNA complex detectable in $F_{obs,n} - F_{obs,1}$ difference maps 187

3.24 Examples of suspected damage sites to the DNA phosphate backbone in the C.Csp231I-DNA complex detectable in $F_{obs,n} - F_{obs,1}$ difference maps 189

3.25 Examples of unexplainable $F_{obs,n} - F_{obs,1}$ map difference density located on the DNA component of the C.Csp231I-DNA complex 190

3.26 The crystal packing for the C.Esp1396I-DNA complex 192

4.1 A rendering of the TRAP-RNA complex asymmetric unit 197

4.2 The transcription attenuation in *B. subtilis* involving TRAP 199

4.3 The TRAP-RNA complex crystal in a cryoloop at beamline ID14-4, ESRF 201

4.4 The intensity decay with dose for the TRAP-RNA complex radiation damage series 202

4.5 The homogeneous dose distribution calculated by *RADDOSE-3D* across the TRAP-RNA complex crystal 203

4.6 The top electron-density loss sites as indicated by the D_{loss} metric for the TRAP-RNA complex asymmetric unit 208

4.7 The average D_{loss} for each residue/nucleotide type with respect to DWD 209

4.8 Manual inspection of $F_{obs,n} - F_{obs,1}$ maps for protein and RNA components of the TRAP-RNA complex 211

4.9 The distribution of D_{loss} metric values calculated for RNA P, O_5‘ and O_3‘ atoms 212

4.10 The distribution of D_{loss} metric values calculated for all phosphorus atoms for each RNA nucleotide types 213

4.11 A schematic illustrating the distribution of Glu and Asp groups dispersed around the RNA-bound TRAP ring 214

4.12 The D_{loss} metric against dose for Glu and Asp carboxyl side-groups in TRAP . . 215

4.13 The Density-weighted D^{-}_{mean} metric against dose for Glu and Asp carboxyl side-groups in TRAP 216

4.14 The D_{loss} and ΔB-factor metrics versus nucleobase hydrogen-bonding distance for Asp-39 and Glu-36 in TRAP 218

4.15 The variation in binding interactions between Asp-39 and G1 nucleobase for different structures of TRAP deposited in the PDB 220

4.16 The D_{loss} metric versus dose for specific atoms at the RNA-binding interfaces of TRAP . . . 222

4.17 The crystal packing arrangement for the TRAP-RNA complex . . . 224

4.18 The crystal contacts exhibited by TRAP in the crystal . . . 225

4.19 The D_{loss} and density-weighted D^{-}_{mean} metrics against dose for Tyr-62 in TRAP . 226

4.20 The isotropic *B*-factor and real-space Z-score values attained by each Glu and Asp group around TRAP. . . . 227

5.1 A schematic illustrating the semiintercalating and intercalating interactions of the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand with the double-stranded TCGGCGCCGA DNA sequence . . . 234

5.2 A schematic illustrating the CGCAAATTTGCG DNA duplex and intercalated four-stranded CCCT DNA . . . 235

5.3 The DNA crystals mounted in loops at DLS beamline I02 . . . 238

5.4 $2mF_{\text{obs}}-DF_{\text{calc}}$ map density consistent with an ordered cobalt hexammine molecule for the CCCT tetrameric DNA crystal . . . 239

5.5 A screenshot of the PDBe ligand summary page indicating the distorted ligand geometry for the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand . . . 240

5.6 Elemental areal concentration microPIXE maps for P and Ba for the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal . . . 248

5.7 PIXE and RBS spectra for the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal used for determination of Ba^{2+} content . . . 249

5.8 The decay in average intensity, I_n, relative to the initial diffraction dataset, I_1, as a function of DWD, for DNA crystal types. . . . 251

5.9 Crystal packing interactions for the TCGGCGCCGA duplex, CGCAAATTTGCG duplex and CCCT tetramer . . . 254

5.10 Site-specific damage analysis of the TCGGCGCCGA duplex crystal using *RIDL* . 256

5.11 Site-specific damage analysis of the CGCAAATTTGCG duplex crystal using *RIDL* 257

5.12 Site-specific damage analysis of the CCCT tetrameric crystal using *RIDL* . . . 258

5.13 Examples of phosphate electron density loss in $F_{\text{obs,11}} - F_{\text{obs1}}$ maps for the CGCAAATTTGCG duplex and CCCT tetramer crystals . . . 259

5.14 The atomic scattering factor f as a function of resolution for Ru, Ba, P and C atoms 261

5.15 Analysis of Fourier ripples in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps for the TCGGCGCCGA duplex crystal . . . 263

5.16 D_{loss}(atom) metric values for phosphate backbone P atoms as a function of sequence for the TCGGCGCCGA crystal 264

5.17 Parameters derived by *3DNA* defining the geometry and orientation between bases in each base pair, and between adjacent base pairs for the TCGGCGCCGA crystal 265

5.18 Correlation between D_{loss}(atom) metric and isotropic *B*-factor values attained in the first dataset coordinate model for P atoms in the TCGGCGCCGA duplex crystal 266

6.1 The original observation of Tyr-330 −OH electron density loss in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps reported in Burmeister (2000) 273

6.2 The fit of Tyr-215 residue to electron density for successive stages of the *PDB_REDO* model rebuilding pipeline 278

6.3 Reproduction of specific radiation damage to Tyr-330 and Glu-409 in the Burmeister (2000) myrosinase damage series 280

6.4 Specific radiation damage to Tyr-215 in the myrosinase damage series (Burmeister, 2000) 281

6.5 The B_{Damage} change for each Tyr O_η atom in the highest dose myrosinase structure relative to the lowest-dose model 282

6.6 The D_{loss} and density-weighted D^{-}_{mean} metrics for the top three Tyr −OH damage sites in myrosinase 283

6.7 The distribution of C_α-normalised D_{loss}(atom) values for side-chain oxygen atoms in MX damage series retrieved from *PDB_REDO* 289

6.8 Electron density loss rankings for distinct side-chain oxygen atom types occurring in each protein MX damage series retrieved from *PDB_REDO* 293

6.9 Combined electron density loss rankings for distinct side-chain oxygen atom types occurring across protein MX damage series retrieved from *PDB_REDO* 294

6.10 C_α-normalised D_{loss} damage signature histogram plots over all atoms of specific residue types in proteins retrieved from *PDB_REDO* 296

6.10 C_α-normalised D_{loss} damage signature histogram plots over all atoms of specific residue types in proteins retrieved from *PDB_REDO* 297

6.11 Kolmogorov-Smirnov (KS) test statistics quantifying similarity of D_{loss} damage signatures for different residue types for large protein structures (> 60 kDa) . . . 298

6.12 A comparison of C_α-normalised D_{loss}(atom) behaviour for covalently-bound atoms in the side-chains in residues of myrosinase 301

6.13 A comparison of C_α-normalised D_{loss}(atom) behaviour for covalently-bound atoms in the side-chains in residues of *Trp* RNA-binding attenuation protein (TRAP) . . 302

6.14 The linear correlation between C_α-normalised D_{loss} metric values for selected covalently bound atom pairs for large protein structures (> 60 kDa) 303

6.15 Manual building of secondary Tyr-330 conformation in myrosinase prior to occupancy refinement . . . 304

6.16 Occupancy refinement for a selected subset of Tyr residues in myrosinase . . . 305

6.17 A simulation of radiation-induced conformational change of Tyr-330 in myrosinase 307

6.18 The variation in electron density (σ) sampled around the χ^1 angle of Tyr-330 and Tyr-215 in myrosinase . . . 309

6.19 A schematic to demonstrate the partitioning of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxels assigned to an atom by cluster analysis . . . 311

6.20 The D_{shift} metric for O_η atoms of Tyr residues in the myrosinase damage series (Burmeister, 2000) . . . 314

6.21 The effect of carboxylate hydrogen bonding on the distribution of C_α-normalised D_{loss} values for Tyr −OH groups in MX damage series retrieved from *PDB_REDO* 315

6.22 Linear R^2 coefficients calculated for each MX damage series for correlation between Tyr −OH D_{loss} metric values and solvent accessibility of each Tyr aromatic ring . 316

6.23 Linear R^2 coefficients calculated for each MX damage series for correlation between Tyr −OH D_{loss} metric values and proximity to closest disulphide bond . . . 318

7.1 The decay in $2mF_{\text{obs}} - DF_{\text{calc}}$ map electron density localised around specific atoms, quantified by the D_{median} and D_{max} metrics . . . 329

7.2 The decay in $2mF_{\text{obs}} - DF_{\text{calc}}$ map electron density localised around specific atoms, quantified by the D_{median} and D_{max} metrics for an alternative map calculation strategy . . . 331

7.3 A schematic illustrating the steps performed to generate an END map from a $2mF_{\text{obs}} - DF_{\text{calc}}$ map . . . 331

7.4 The decay in END map electron density localised around specific atoms, quantified by the D_{median} and D_{max} metrics . . . 332

7.5 The effect of using C_α-normalised metrics to monitor $2mF_{\text{obs}} - DF_{\text{calc}}$ map electron density localised around specific atoms . . . 334

7.6 The decay in the C_α-normalised D_{median} metric as a function of dose for the GH7 family protein, fitted with an exponential decay function assuming that density decays to the bulk solvent level as dose $d \rightarrow \infty$ . . . 337

7.7 The decay in the C_α-normalised D_{median} metric as a function of dose for the GH7 family protein, fitted with an exponential decay function assuming density decay to a variable end point as dose $d \rightarrow \infty$ . . . 341

7.8 Half-dose values for $2mF_{\text{obs}} - DF_{\text{calc}}$ map electron density to decay to halfway between the initial density state (at 0 MGy) and high dose limit (as dose $d \rightarrow \infty$) 343

7.9 A simulation of electron density decay against dose for two hypothetical atoms A and B that have the same half-dose value, yet significantly different overall extent of electron density decay. 345

7.10 Initial decay rate values for $2mF_{obs} - DF_{calc}$ map electron density to decay to halfway between the initial density state (at 0 MGy) and high dose limit (as dose $d \rightarrow \infty$) 346

7.11 $F_{obs,D_2} - F_{obs,D_1}$ Fourier difference maps generated from single MX dataset for insulin 350

7.12 A schematic illustrating a protocol to systematically generate Fourier difference maps from a single MX dataset 351

7.13 Heat map analysis of the radiation-susceptibility of individual atoms in insulin . . 353

7.14 The data reduction statistics output by *AIMLESS* for the scaling and merging of each sub-dataset for insulin and the GH7 family protein 354

7.15 Heat map analysis of the radiation-susceptibility of individual atoms in the GH7 family protein 356

7.16 A visualisation of Fourier difference maps generated from a single MX dataset for the GH7 family protein 357

7.17 The rate of dose accumulation (in units of MGy/°) for the GH7 family protein data collection 358

7.18 Heat map analysis of the radiation-susceptibility over all atoms in insulin 360

7.19 Heat map analysis of the radiation-susceptibility over all atoms in the GH7 family protein 361

B.1 Poster presentation at the 29th annual European Crystallographic Meeting (ECM29), Rovinj, Croatia, taking place on 23-28th August, 2015 402

B.2 Poster presentation at the 9th International Workshop on X-ray Radiation Damage to Biological Crystalline Samples (RD9), Lund, Sweden, taking place on 9-11th March 2016 403

B.3 Poster presentation at the Gordon Research Conference (GRC) on Diffraction Methods in Structural Biology, Lewiston, ME, USA, taking place on 17-22th July 2016 404

B.4 Poster presentation at the 67th annual American Crystallographic Association meeting (ACA), New Orleans, ME, USA, taking place on 26-30th May 2017 . . . 405

C.1 Front cover image for Journal of Synchrotron Radiation 2015 March edition . . . 408

C.2 Front cover image for Journal of Synchrotron Radiation 2017 Jan edition 409

C.3 Front cover image for ACA refleXions Summer 2016 edition 410

List of Tables

2.1 Demonstration of the increase in geometry weighting in *REFMAC* when refining higher dose datasets of the *Dp*Cel7B protein test series 75

2.2 GH7 family protein data processing and refinement statistics 78

2.3 Geometry validation for the GH7 family protein at the lowest reported dose . . . 79

2.4 Van der Waals radii for atoms typically found in proteins and nucleic acids 84

2.5 Number of peaks filtered at each stage of *PEAKS* 85

2.6 Possible *RIDL* summary metrics to quantify site-specific damage 115

3.1 Diffraction data collection details for the C.Esp1396I-DNA complex at 100 K . . 152

3.2 Data processing statistics for the C.Esp1396I-DNA complex damage series 158

3.3 Geometry validation for the C.Esp1396I-DNA complex at the lowest reported dose 159

3.4 Statistical properties associated with the distribution of C_α-normalised D_{loss} values for protein and DNA components of the C.Esp1396I-DNA complex 173

3.5 Data processing statistics for the C.Csp231I-DNA complex damage series 184

3.6 Geometry validation for the C.Csp231I-DNA complex at the lowest reported dose 185

4.1 Data processing statistics for the TRAP-RNA complex damage series 206

4.2 Geometry validation for the TRAP-RNA complex at the lowest reported dose . . 207

5.1 TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ DNA data processing and refinement statistics 241

5.1 TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ DNA data processing and refinement statistics 242

5.2 CGCAAATTTGCG duplex DNA data processing and refinement statistics 243

5.3 CCCT tetramer DNA data processing and refinement statistics 244

5.4 Concentrations of each identified element derived from the PIXE spectra taken at two sample positions on the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal . . . 250

6.1 Distribution of real-space Z_{obs} scores calculated by *EDSTATS* for the *1dwa* structure retrieved directly from the PDB and retrieved from *PDB_REDO* 277

6.2 Selection of MX damage series (all conducted at 100 K) previously deposited in the PDB 285

6.2 Selection of MX damage series (all conducted at 100 K) previously deposited in the PDB 286

6.3 R_{work} and R_{free} statistics corresponding to the lowest dose structures for each MX damage series located in the PDB and *PDB_REDO* 288

6.4 Number of conformations detected by *RINGER* per Tyr side-group for each dataset in the myrosinase damage series (Burmeister, 2000) 310

A.1 MX data collection information for a bovine insulin dataset supplied by Dr. J. C. Brooks-Bartlett 398

A.2 Data processing statistics for bovine insulin dataset supplied by Dr. J. C. Brooks-Bartlett 399

CHAPTER 1

Introduction

The attainment of the three-dimensional (3D) atomic structure of any macromolecule is an essential hurdle that must be overcome to achieve a correct understanding of its function in the cellular environment. Without precise knowledge of the relative positions of the atoms comprising a macromolecule at near-atomic resolution, quite how it achieves this function, through an intricate interplay with other proteins, nucleic acids and other biomolecules, cannot be unravelled.

Macromolecular X-ray crystallography (MX) is currently the most widely utilised technique available to achieve a near-atomic resolution viewpoint of macromolecules. As of January 2017, over 110,000 macromolecular structures have been solved by MX techniques and deposited in the Protein Data Bank (PDB), comprising 89.4 % of the total deposited structures. However, the ionising radiation used in MX to solve a macromolecular structure also inevitably deposits unwanted energy directly within the irradiated crystal throughout diffraction data collection. The resulting chemical and physical changes, termed *radiation damage*, occur at both the whole crystal and individual macromolecular scale, and ultimately disrupt the success of the crystallographic pipeline to output a valid atomic structure.

1.0.0.1 Aim of the thesis

The aim of this thesis is to present research into radiation damage in MX, with a focus on the development of novel methods to enable high-throughput, objective detection and quantification of radiation-induced changes to macromolecular structures.

1.1 The macromolecular X-ray crystallographic pipeline

The crystallographic pipeline for macromolecular structure solution describes the full workflow from initial crystal growth through to final specification of 3D atomic coordinates. Each stage in this pipeline has inherent difficulties that must be overcome, with manifestations of radiation damage known to contribute to many of these difficulties. In this section, the full pipeline will be presented in the context of radiation damage, to illustrate where radiation damage is detrimental to the pipeline's success.

1.1.1 Crystallogenesis

Prior to X-ray irradiation, a well-ordered crystalline lattice (a crystal) of the target macromolecule is required, and the achievement of this state remains a major bottleneck in the crystallographic pipeline. A target protein must be successfully cloned, expressed and purified, such that a concentrated pure (typically $5 - 15\,\mathrm{mg\,ml^{-1}}$) soluble sample is available in µl quantities (Chayen and Saridakis, 2008). However, this alone does not guarantee the desired crystal phase transition. The screening for successful crystallisation conditions is a multiparametric process (e.g. pH, buffer, temperature, precipitating agents are some of the variable parameters) with no clear correlations between viable conditions and protein family. Furthermore, the method of crystallisation can be varied, with vapour diffusion (either hanging or sitting drop) most widely utilised, but with other methods available (e.g. microbatch, dialysis or free interface diffusion) in cases where vapour diffusion is deemed unsuitable. Efforts are ongoing to optimise high-throughput crystallisation screens that are applicable to general proteins (Skarina *et al.*, 2014).

Radiation damage effects that emerge later in the MX pipeline must be considered during crystal growth. The inclusion of heavy atom compounds in the crystallisation buffer increases the X-ray absorption cross-section of the resulting crystal sample, and should be omitted if possible (Paithankar and Garman, 2010). In cases where heavy metal derivatization is required when phasing by isomorphous replacement, backsoaking can be performed to remove non-bound heavy atoms that do not contribute to the anomalous diffraction signal (Pike *et al.*, 2016). The addition of small molecule compounds, known to have free radical scavenging or radioprotectant properties for proteins irradiated in solution and *in vivo* (Koukourakis, 2012), can also be introduced by either co-crystallisation or crystal soaking (Murray and Garman, 2002). However, there have been conflicting results on the efficacy of particular small molecule candidates at mitigating radiation damage, suspected to be due to the use of differing metrics to assess radiation damage in each study (Southworth-Davies and Garman, 2007; Nowak *et al.*, 2009; Kmetko *et al.*, 2011; De la Mora *et al.*, 2011). Only one study has reported an enhancement of crystal *life-dose* exceeding a factor of two (a nine-fold increase with 1,4-benzoquinone), but only for room-temperature (RT) data collection (Barker *et al.*, 2009). The benefit of scavenger inclusion is thus unclear; a detailed review is provided in Allan *et al.* (2013).

1.1.2 Diffraction data collection

Once a suitable high-quality crystal has been grown, the aim is to expose the crystal to a coherent X-ray beam and to record, with high accuracy, the intensities of elastically scattered photons on a position-sensitive detector. For a monochromatic X-ray beam of wavelength λ incident on a crystal, Bragg peaks (otherwise known as reflections) are detectable at well defined detector positions at which elastically scattered photons constructively interfere (from sets of crystalline lattice planes at interplanar distances $d > 0$). Such positions satisfy Bragg's law (Bragg and Bragg, 1913):

$$n\lambda = 2dsin(\theta) \tag{1.1.1}$$

for positive integer $n \in \mathbb{N}_0$ and scattering angle 2θ. With a single fixed crystal orientation relative to the incident beam, Bragg's law is satisfied for only a subset of the possible sets of crystalline lattice planes. Rotation of the crystal within the beam allows Bragg's law to be satisfied for additional distinct sets of lattice planes, consequently allowing more unique reflections to be observable within the physical limits of the detector size. For modern MX data collection at a third generation synchrotron radiation source, this is typically achieved by irradiating the crystal whilst suspended in a thin film of its own mother liquor in a mylar loop (Teng, 1990) and typically held at 100 K in a gaseous nitrogen stream. Most commonly the loop is mounted on a goniometer, which permits precise rotation of the loop, and thus the crystal, about an axis whilst exposed to the X-ray beam. More complicated data collection strategies are now also possible, such as helical scanning (rotation combined with translation) and the use of multiple rotation axes (with kappa-based goniometers (Brockhauser *et al.*, 2013)). Numerous synchrotron beamlines now also facilitate *in situ* methods to irradiate crystals directly from plates at room temperature without the requirement of crystal harvesting (Huang *et al.*, 2015; Axford *et al.*, 2015a) and from micro-patterned chips (Roedig *et al.*, 2016).

1.1.2.1 The interaction processes of X-rays with macromolecular crystals

Only a small proportion of X-ray photons interact with the crystal during irradiation. For the case of a 12.4 keV (1 Å) beam incident on a 100 µm thick protein crystal, it has previously been calculated that only 2 % of the photons interact with the crystal (Garman, 2010). Unfortunately

the majority of these interacting photons do not elastically scatter as desired (Figure 1.1a), but instead deposit energy directly within the crystal through two processes: (i) Compton (inelastic) scattering and (ii) the photoelectric absorption, at 8 % and 84 % of the interacting photons respectively.

When a photon interacts via Compton scattering (Figure 1.1b), it transfers a fraction of its energy to an atomic electron, to be re-emitted at a lower energy (and hence a higher wavelength by the Planck-Einstein relation $E = hc/\lambda$, with Planck's constant $h = 6.63 \times 10^{-34}\,\mathrm{m^2\,kg\,s^{-1}}$ and speed of light $c = 3 \times 10^{8}\,\mathrm{m\,s^{-1}}$). Furthermore, the excited electron (termed a *recoil electron*) may be ejected from the atom (Nave, 1995). The energy E_{recoil} transferred to the recoil electron obeys:

$$E_{\mathsf{recoil}} = \frac{E_{\mathsf{init}}^2(1 - cos(\theta))}{m_{\mathsf{e}}c^2 + E_{\mathsf{init}}(1 - cos(\theta))} \tag{1.1.2}$$

where E_{init} and E_{recoil} are photon energies prior and post scattering, $m_{\mathsf{e}} = 9.11 \times 10^{-31}\,\mathrm{kg}$ is the electron rest mass, and $\theta > 0$ is the photon scattering angle (Paithankar and Garman, 2010).

When a photon interacts via the photoelectric effect (Figure 1.1c), it transfers all of its energy to an atom and is completely absorbed in the process. This results in the ejection of a high energy photoelectron from an inner electron shell (typically the K shell), which is ejected with the majority of the incident photon energy, as follows:

$$E_{\mathsf{pe}} = E_{\mathsf{init}} - E_{\mathsf{bind}} \tag{1.1.3}$$

where E_{pe} is the ejected photoelectron energy and E_{bind} is the electron binding energy*. The excited atom may then release the excess potential energy through transition of a higher shell electron to fill the resulting inner shell vacancy. In order for the atom to return to its ground state, the transition is accompanied with either the emission of a fluorescent X-ray (also termed a characteristic X-ray) or ejection of a low energy Auger electron from an outer shell (typically with energy between a few hundred eV for C, N and O to $\approx 2\,\mathrm{keV}$ for S and P). Fluorescence only becomes significant for atoms of high atomic number, and thus for the majority of atoms consti-

*The K-shell electron binding energies E_{bind} are of the order of hundreds of eV for the low atomic number elements (C, N and O) typically constituting proteins: `http://www.chembio.uoguelph.ca/educmat/atomdata/bindener/elecbind.htm`. They are consequently negligible relative to the incident photon energy E_{init}, but become more significant with increasing atomic number Z

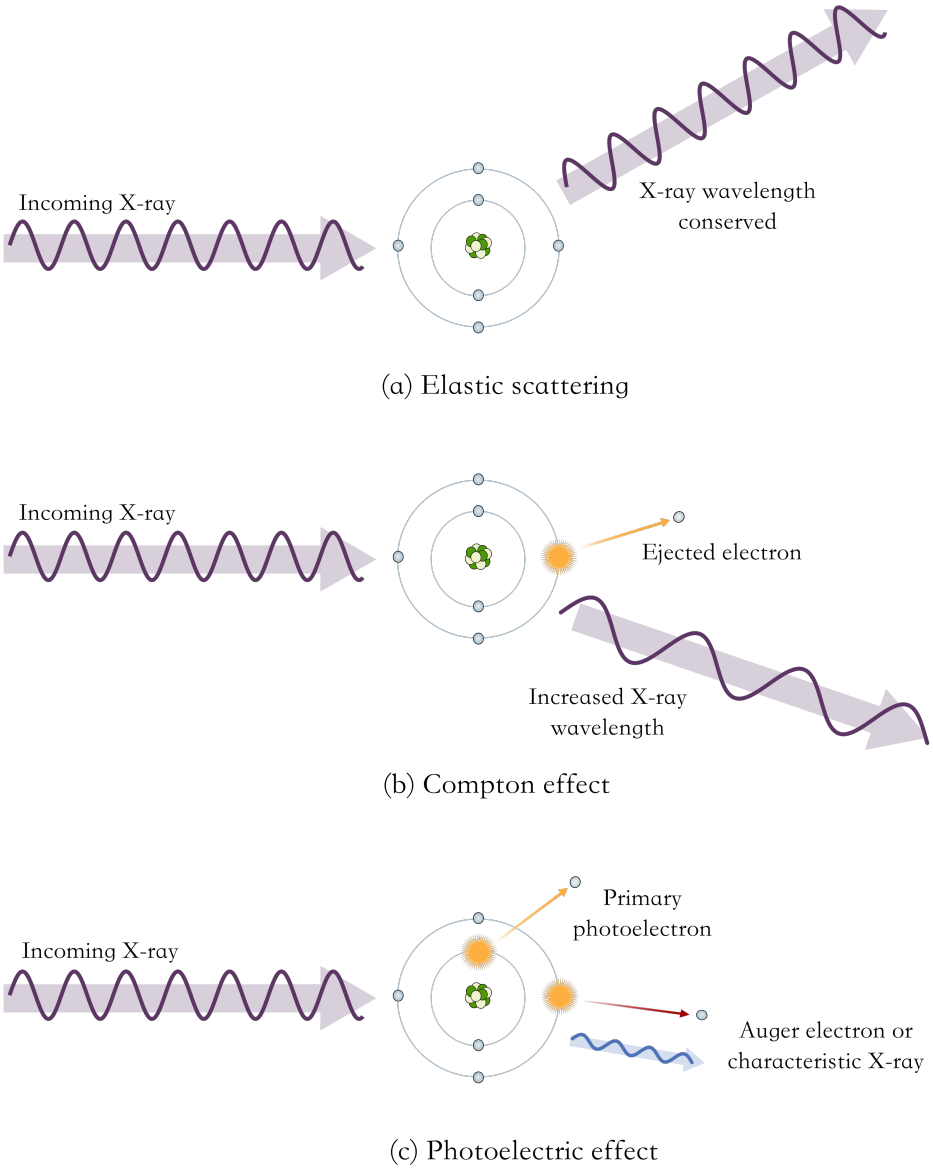


Figure 1.1: The three primary processes by which X-rays can interact with crystals. (a) Elastic scattering (also known as Thomson or coherent scattering): the incident X-ray photon is scattered, however its kinetic energy is left unaltered. (b) Compton scattering: the incident photon deposits a fraction of its kinetic energy in an atomic electron, before scattering at a higher wavelength. The excited electron may be ejected from its atom following the interaction. (c) Photoelectric absorption: the incident photon transfers all its energy to an inner shell electron, which is consequently ejected (termed a photoelectron). The resulting orbital vacancy is then filled by a higher shell electron. This shell transition is accompanied by either emission of a characteristic X-ray or the ejection of a second electron (termed an Auger electron).

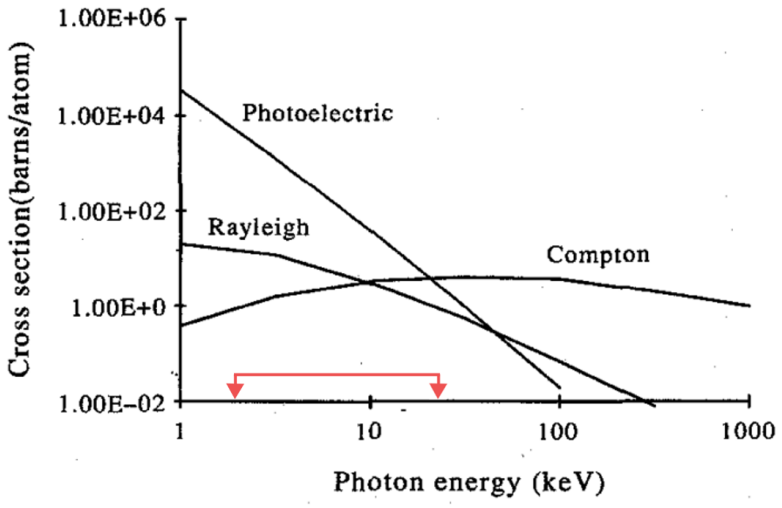


Figure 1.2: The dependence of incident photon energy on cross sections for elastic scattering, Compton scattering and photoelectric absorption. Figure is taken from Nave (1995) for the example of carbon. It has been modified to indicate the typical range of incident photon energies at MX beamlines provided at the DLS and ESRF third generation synchrotrons (in red), which coincides with the region at which the ratio of diffractive to absorption cross-sections $\mu_{\text{el}}/\mu_{\text{abs}}$ is also maximised. Both axes have base-10 logarithmic scales.

tuting proteins (with $Z \leq 8$) Auger emission is the dominant pathway (Nave, 1995). However, for metalloproteins and heavy metal derivatised crystals, the relative yield of fluorescent X-rays becomes significant.

For an interacting X-ray photon, the probability of an elastic or inelastic scattering or photoelectron absorption event is dependent on the incident photon wavelength. For the lighter elements typically constituting proteins (with $Z \leq 8$), and the wavelengths that are achievable at modern third generation synchrotrons[†], the photoelectron absorption cross-section (in units of barns/atom, where $1\,\text{barn} = 10^{-28}\,\text{m}^2$) varies as a function of $1/\lambda^3$. In comparison, the elastic and inelastic scattering cross sections are approximately constant as a function of λ at these wavelengths (Nave, 1995). For each interaction process, the variation of cross section is shown for the example of carbon in Figure 1.2. The wavelengths most commonly used at synchrotrons ($\lambda \approx 0.98\,\text{Å}$[‡]) fall into the approximate domain for which the ratio of cross sections for elastic scattering to Compton scattering and photoelectric absorption is maximised.

[†]typically tuned between $0.5 - 2.5\,\text{Å}$, quoted for MX beamlines at DLS (`http://www.diamond.ac.uk/Beamlines/Mx.html`) and ESRF (`http://www.esrf.eu/UsersAndScience/Experiments/MX/About_our_beamlines`), but as high as 4 Å at the long wavelength I23 beamline at DLS (Wagner *et al.*, 2016).

[‡]coincident with the Selenium K-shell absorption edge at $\lambda = 0.98\,\text{Å}$ to facilitate Se-SAD phasing (Taylor, 2010).

1.1.2.2 Deposition of energy and absorbed dose

The direct ionisation of atoms due to interaction with the incident photons as discussed above is termed *primary* radiation damage (Garman, 2010). The ejected photoelectrons, Auger electrons and recoil electrons (following Compton scattering) are able to migrate from the initial site of photon energy deposition, and distribute their kinetic energy along tracks throughout the crystal through a series of ionisation events with other atoms. Ionisation events result in ejection of further lower energy electrons, eventually leading to cascades of ionisation events (O'Neill *et al.*, 2002). The mean path length for photoelectrons has been found to be $\approx 4\,\mu m$ for incident X-ray beam energies under 20 keV (both through simulations of electron trajections using Monte Carlo program *CASINO* (Nave and Hill, 2005) and experimentally at 100 K (Sanishvili *et al.*, 2011)). Consequently, for microcrystals with dimensions $< 4\,\mu m$, energy initially deposited by the interacting photon may escape from the crystal entirely. It has been estimated that for an incident 12 keV photon, an ejected ~ 12 keV photoelectron will induce approximately 500 further ionisation events, assuming a cost of 25 eV per ionisation event (O'Neill *et al.*, 2002). Due to the lower energy of Auger and recoil electrons, their respective path lengths are anticipated to be considerably shorter (Nave, 1995). In the alternative case in which a fluorescent X-ray is produced following photoelectric absorption, the fluorescent X-ray is either reabsorbed or may also escape the crystal.

The absorbed dose is defined as the energy deposited in the crystal per unit mass of crystal (with SI units of grays, $Gy := J\,kg^{-1}$), and is a direct indicator for the prevalence of primary radiation damage events. For typical MX data collection at third generation synchrotrons with high fluxes of the order of $10^{10} - 10^{12}$ ph/s, the dose is usually of the order of $1 - 10$ MGy per complete dataset. Accumulated dose is directly dependent on the yield of Compton scattering and photoelectric absorption events throughout the whole crystal during X-ray irradiation, and as discussed above is thus dependent on the X-ray beam wavelength and atomic composition of the crystal (both ordered macromolecules and bulk solvent) (Murray *et al.*, 2004). The program *RADDOSE* (Murray *et al.*, 2004), along with subsequent versions (Paithankar *et al.*, 2009; Paithankar and Garman, 2010), was developed to simulate the irradiation of crystal in the beam, whilst taking into account the properties of the crystal (e.g. dimensions, chemical composition, solvent content) and beam (e.g. energy, incident flux, beam size, collimation, intensity profile),

that must be considered to provide an accurate estimate of the dose absorbed by the crystal. The original *RADDOSE* program only reported a *maximum* dose value across a crystal volume, and was unable to account for crystal rotation in the X-ray beam. *RADDOSE* was eventually superseded by *RADDOSE-3D* (Zeldin *et al.*, 2013b), which additionally takes into account both the 3D crystal geometry and crystal rotation, to calculate the spatial and temporal distribution of dose throughout the crystal.

Several dose metrics have been proposed to suitably represent the damage state of the crystal during irradiation, including the *maximum* dose and *whole-crystal average* dose (AD-WC), however the *diffraction-weighted* dose (DWD), offered by *RADDOSE-3D*, appears to be most suitable (Zeldin *et al.*, 2013a). DWD weights the accumulated dose at each point $\underline{x}$ in the crystal, $D(\underline{x}, t)$, by the X-ray fluence (photons per unit area) through that point, $F(\underline{x}, t)$, at each time point $t > 0$ of data collection. For a given image number $i \in \mathbb{N}$ collected during data collection, the diffraction-weighted-dose is given by:

$$\text{DWD}^i = \frac{\int_{t_{i-1}}^{t_i} \int_{cryst} D(\underline{x}, t) \times F(\underline{x}, t) \; d\underline{x}dt}{\int_{t_{i-1}}^{t_i} \int_{cryst} F(\underline{x}, t) \; d\underline{x}dt} \tag{1.1.4}$$

where $[t_{i-1}, t_i]$ is the time interval over which image number i is collected. For a given image, the dose in regions of the crystal that are more pertinent to the recorded diffraction pattern are thus upweighted. The average DWD accumulated across a complete dataset of diffraction images is now typically quoted to quantify the deposited energy throughout the duration of data collection.

1.1.2.3 Radiation-induced changes to the crystal

It is crucial to determine what physical changes occur to the crystal during irradiation, in order to elucidate how the accumulation of dose affects the later processing steps in the MX pipeline.

On the scale of the whole crystal, upon X-ray irradiation, visible discolouration of the crystal is often observable (Ravelli and Garman, 2006), notably a characteristic blue colouring due to trapped solvated electrons (McGeehan *et al.*, 2009). It has been reported in Holton (2009) that regions of crystal can physically expand at 100 K when exposed to sufficiently high X-ray beams due to increasing unit cell dimensions (see below). An increase in crystal mosaicity, as in the

non-isomorphism and misalignment of discrete mosaic blocks within the crystal, has also been reported as a function of increasing dose (Ravelli and McSweeney, 2000; Meents *et al.*, 2009; Nave *et al.*, 2016). However, this is not consistently true for all protein crystals, and thus mosiacity change is an unsuitable metric for predicting the level of damage (Garman, 2010).

On the scale of the unit cell, the smallest grouping of atoms for which the entire crystal symmetry can be reconstructed by only translations along three principal axes a, b and c, the unit cell dimensions themselves have been widely reported to expand with accumulated dose (Ravelli and McSweeney, 2000; Burmeister, 2000; Meents *et al.*, 2010). It has been demonstrated that crystals of the identical protein and comparable dimensions expand at varying rates with dose and the unit cell dimensions are thus also deemed an unsuitable indicator for implicitly measuring radiation damage (Murray and Garman, 2002; Ravelli *et al.*, 2002).

A consensus on the origins of the aforementioned changes to crystalline integrity has not been reached, however it has been suggested that radiation-induced hydrogen abstraction, for instance from protein backbone C_α atoms (Garrison, 1987), may contribute (Meents *et al.*, 2009). A resulting accumulation of molecular hydrogen gas in the crystal would exert a disruptive force on the crystalline lattice, leading to a deterioration in crystalline order (Meents *et al.*, 2010).

1.1.2.4 Experimental strategies to mitigate radiation damage

Temperature

Initial observations of a temperature dependence on the diffracting life-time of protein crystals were made as early as 1970, where it was shown that holding a lactate dehydrogenase crystal at 100 K reduced the decay of reflection intensities with dose, relative to at RT (Haas and Rossmann, 1970). As demonstrated by Figure 1.3, by the late 1990s, cryo-crystallography (diffraction data collection whilst holding the crystal at approximately 100 K) had consequently become the dominant technique to mitigate the effects of radiation damage (Hope, 1988; Garman and Schneider, 1997; Garman and Owen, 2006). It has recently been reported that of all structures deposited in the PDB that do report information on data collection temperature, 84 % were collected at 100 K (Garman and Weik, 2017a).

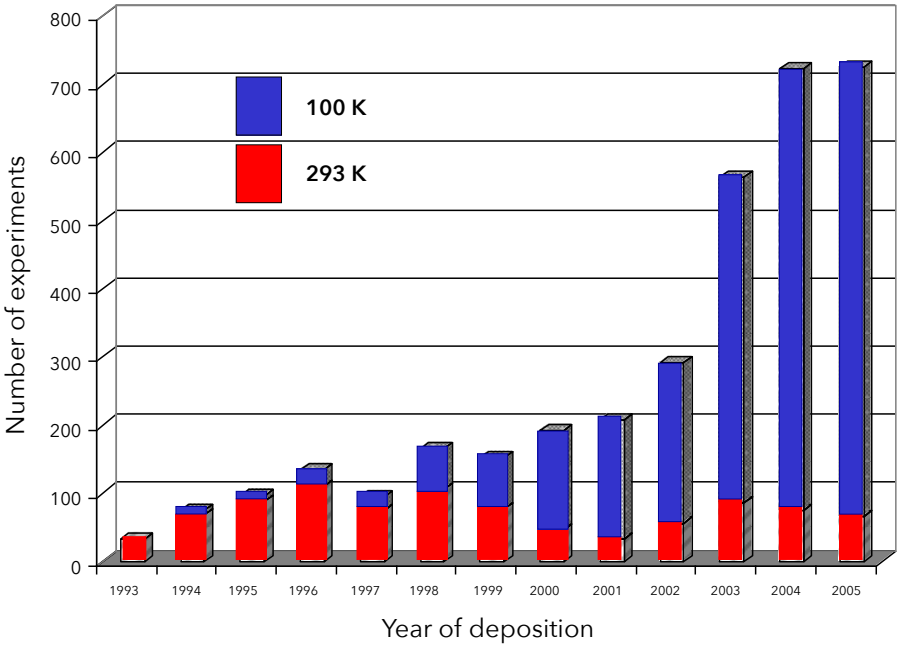


Figure 1.3: The number of MX structures that were reported in Acta Crystallographica. D, between January 1993 and December 2005. Structures have been divided into those reported to be solved at cryocooled temperatures ($\sim$ 100 K) and those at room temperature (293 K). Figure adapted from Garman (2009).

The molecular basis for damage reduction at 100 K is believed to be due to the immobility of most damaging radical species (e.g. hydroxyl radicals, see Owen *et al.* (2012)). However, electrons have been shown to still be mobile at 77 K (Jones *et al.*, 1987), and thus radiation damage is not entirely eliminated at 100 K[§]. Observations of damage at 100 K were first made at second generation synchrotrons using focused white beams (Gonzalez *et al.*, 1992), and are now widely reported at undulator beamlines at third generation synchrotrons (Garman, 2010). However, the dose tolerance for crystals held at 100 K has been predicted to be on average around a factor of 70 times that at RT (Nave and Garman, 2005). This has since been experimentally verified for lysozyme crystals, which exhibited 26-113 times increase in crystal lifetime at 100 K (Southworth-Davies *et al.*, 2007).

Cryo-cooling is typically achieved by first soaking the crystal in an adequate cryosolution (a combination of cryoprotectant and mother liquor), before flash-cooling the looped crystal in liquid nitrogen. Selection of cryoprotectant agent needs to be optimised to best mitigate crystalline ice formation within the crystal solvent channels (Garman and Owen, 2006). During X-ray irradiation, the crystal must be held at cryo-temperatures through use of a cryostat to deliver a stream of

[§]A more detailed discussion on how such species induce damage is presented in Section 1.2.2

nitrogen gas at a precise temperature of 100 ± 0.1 K (Cosier and Glazer, 1986).

Further marginal increases in diffraction data quality have been demonstrated for crystals held below using open flow helium cryostats 100 K (Borek *et al.*, 2007; Meents *et al.*, 2010). Radiation-induced photoreduction in metalloprotein crystals has also been reported to be substantially lower at 40 K compared to 100 K (Corbett *et al.*, 2007). However, the necessity of using an open flow or closed-cycle helium cryostat, both costly and non-standard on MX beamlines, has severely limited the desirability for data collection below 100 K.

Despite the significant enhancement in crystal lifetime using cryo-crystallographic techniques, there are reported instances where cryo-cooling introduces non-biologically relevant conformational artefacts into structures and reduces conformational diversity (Fraser *et al.*, 2011). With the recent availability of faster detectors (Owen *et al.*, 2014) and *in situ* plate handling methods implemented on synchotron beamlines (Axford *et al.*, 2015b), it is now possible to collect sufficient RT diffraction data for structure solution, but with the compromise of typically merging data from a large number of isomorphous crystals. Recent studies have indicated that the conformational landscape of structures is highly sensitive to temperature, and there is significant benefit in collecting data at multiple distinct temperatures between $100 - 310$ K to characterise multiconformer models of structures (Keedy *et al.*, 2015).

Dose rate

The original MX radiation damage study conducted at RT on sperm-whale myoglobin predicted that radiation damage was linearly proportional to absorbed dose (Blake and Phillips, 1962), with no anticipated role in the rate at which dose was deposited (in units of $\mathrm{Gy\,s^{-1}}$). In modern cryo-crystallographic data collection at high flux-density MX beamlines, the presence of a dose rate effect is unclear, and at higher dose rates, studies have reported between a 10 % reduction in crystal lifetime (Owen *et al.*, 2006) to none at all (Sliz *et al.*, 2003).

In contrast, at RT there have been multiple studies indicating that at high dose rates it is possible to outrun the timeframe of radiation damage processes (Southworth-Davies *et al.*, 2007; Barker *et al.*, 2009; Owen *et al.*, 2012), with recently the suggestion of an initial lag phase in radiation damage lasting ~ 500 kGy at high dose rates ($\geq 1\,\mathrm{MGy\,s^{-1}}$) that was demonstrated to

be insensitive to crystal composition and incident beam energy[¶] (Owen *et al.*, 2014). Conversely, at lower dose rates ($0.05 - 300\,\text{kGy}\,\text{s}^{-1}$) a lack of dose rate effects at RT have been reported for a range of proteins (Leal *et al.*, 2013).

With the introduction of upcoming fourth generation synchrotrons (e.g. MAX IV in Lund, Sweden, and NSLS II at Brookhaven, USA) (Castelvecchi, 2015), X-ray fluxes of the order of 2×10^{13} ph/s will now be achievable at MX beamlines (Figure 1.4). The significant increase in dose rate at these facilities is predicted to be a limiting factor to MX, by incurring significant heating effects and by reducing the time to reach the experimental 30 MGy dose limit (Owen *et al.*, 2006) for cryocooled protein crystallography to just tens of µs (Garman and Weik, 2017b).

In the upper dose rate limit at X-ray free electron lasers (XFELs), in which crystal samples are irradiated by $\sim 10^{12}$ photons per single pulse in tens of femto seconds (Chapman *et al.*, 2011), it has been suggested that diffraction data can be collected prior to the development of any radiation-induced chemical or structural effects (Neutze *et al.*, 2000). However, both global (Lomb *et al.*, 2011) and site-specific (Nass *et al.*, 2015; Galli *et al.*, 2015) damage events have now also been reported in the case of unattenuated XFEL pulses of lengths over 50 fs.

1.1.3 Data reduction and radiation damage

Once a series of diffraction images has been collected from a crystal, the aim is to extract the intensity values, I_{hkl} for $h, k, l \in \mathbb{Z}$, for each Bragg peak present on each image, and convert these to a set of structure factor amplitudes $|\mathbf{F}_{hkl}|$, via the relationship:

$$|\mathbf{F}_{hkl}| = \left(\frac{KI_{hkl}}{Lp}\right)^{\frac{1}{2}} \tag{1.1.5}$$

where L is the Lorentz factor, p is the polarisation factor, and K is a scale constant. If both the structure factor amplitudes and phases $\phi(hkl)$ are known, the magnitude of the electron density

[¶]Owen *et al.* (2014) demonstrated that the $\sim$ 500 kGy lag phase in intensity decay held at RT for thaumatin crystals at a dose rate of $\geq 1.32\,\text{MGy}\,\text{s}^{-1}$ and a beam energy of 12.8 keV, *Bovine enterovirus* serotype 2 crystals at a dose rate of $\geq 4.8\,\text{MGy}\,\text{s}^{-1}$ and a beam energy of 12.8 keV and *Foot and mouth disease virus* crystals at a dose rate of $\geq 25.2\,\text{MGy}\,\text{s}^{-1}$ and a beam energy of 9.2 keV.

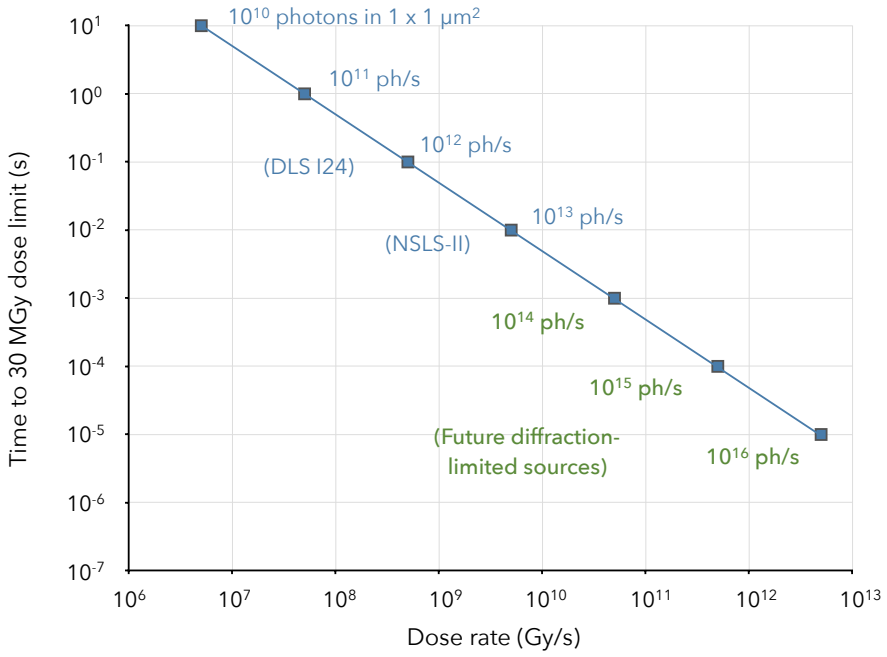


Figure 1.4: Approximate time for a typical protein crystal to reach the 30 MGy experimental dose limit (Owen *et al.*, 2006) for varying dose rates for an X-ray beam between 1×10^{10} photon/s and 1×10^{16} photon/s into a 1 µm $\times$ 1 µm area. Figure reproduced from the presentation of Elspeth Garman at the radiation damage workshop RD6, March 2010.

at each point (x, y, z) in the unit cell can be determined:

$$\rho(x, y, z) = \frac{1}{\mathrm{V}} \sum_h \sum_k \sum_l |\mathbf{F}_{hkl}| e^{-2\pi i(hx+ky+lz)+i\phi(hkl)} \tag{1.1.6}$$

where V is the unit cell volume (in Å^3). The integers h, k and l correspond to the reciprocal lattice indexing scheme for the crystal, with each hkl triplet corresponding to a single reciprocal lattice point. By a Fourier transform, the structure factor for each Bragg peak, $\mathbf{F}_{hkl}$, can itself be expressed as an integral of electron density ρ over the crystal unit cell:

$$\mathbf{F}_{hkl} = \iiint_{cell} \rho(x, y, z) e^{2\pi i(hx+ky+lz)} \, dx \, dy \, dz \tag{1.1.7}$$

The Fourier summation in Equation 1.1.6 is computed over all integer hkl indices with each contributing an individual term into the sum. However, due to inherent limitations in the diffracting power of the crystal, and the dimensions and sensitivity of the detector, only a subset of reflections are measurable during data collection ($\{hkl\}_{meas} \subset \mathbb{Z}^3$). Precise measurement of as many distinct hkl reflection intensities as possible (e.g. by rotating the crystal, by collecting highest

resolution data possible) is essential such that the truncated Fourier summation (calculated over $\{hkl\}_{meas}$) is still an appropriate estimator for $\rho(x, y, z)$.

Unfortunately, radiation damage leads to a reduced crystal lifetime in which to record diffraction data, and furthermore a reduction in the quality of intensity measurements that are obtained with accumulated dose.

1.1.3.1 Diffraction Image Processing

Various software packages such as *MOSFLM* (Leslie and Powell, 2007), *XDS* (Kabsch, 2010) and *DIALS* (Waterman *et al.*, 2016) are available to perform indexing and integration of diffraction images. Information on the unit cell dimensions and possible crystal lattice symmetry (within tolerances for cell edges and angles), as well as crystal orientation, can be extracted from a subset of diffraction images with knowledge of the crystal-detector distance, beam centre coordinates and incident beam wavelength. This information can then be used to index reflections on all the remaining images. For each reflection, the pixel intensity values in an assigned local region around the reflection are then integrated to obtain an intensity value I^i_{hkl} for that particular reflection on that image i. Following integration, the crystal point group (the rotational symmetry operators for the crystal) can then be determined using the program *POINTLESS* (Evans, 2006).

1.1.3.2 Scaling and Merging Intensities and Global Radiation Damage

Due to internal crystal symmetry, in a continuous sweep of diffraction images, the same hkl reflection can appear multiple times. Single intensity values for each hkl (representative of the dataset) are then derived through a process called merging. However, due to changes in crystal thickness and exposed volume, and deviations in beam intensity throughout data collection, the resulting intensities are not anticipated to be identical in practice (Evans, 2011).

It is also known that reflection intensities change as a function of increasing dose, with the first observations of intensity decay with irradiation time being made as early as 1962 on myoglobin crystals at RT (Blake and Phillips, 1962) and then for lamprey haemoglobin in 1973 (Hendrickson *et al.*, 1973). This is now known as global radiation damage, and has been widely reported at

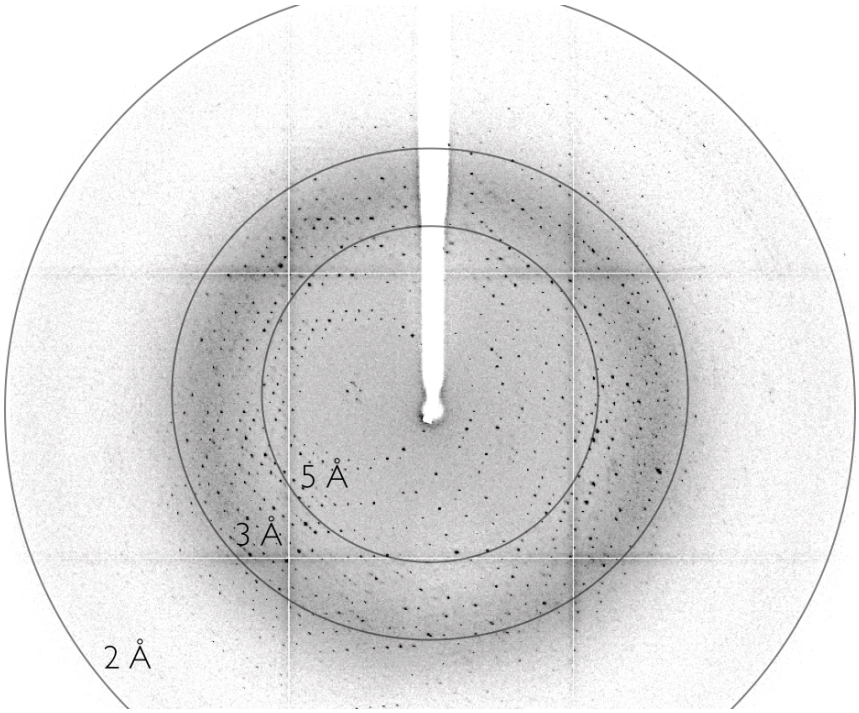
both RT and for cryocooled crystals. The intensity decay is resolution dependent, with the high resolution reflections decaying quickest. In general, the intensity of individual reflections may not monotonically decrease with dose, with intensities of particular reflections recorded to instead increase with dose (Blake and Phillips, 1962; Hendrickson *et al.*, 1973; Abrahams, 1973). However, the average of the measured intensities, I_D at a dose D, relative to those recorded at the start of data collection (I_1), denoted as I_D/I_1, has been widely observed to decay with dose (Owen *et al.*, 2006; Zeldin *et al.*, 2013a).

The diffracting life-dose of a crystal is typically quoted as the dose at which the average of the recorded intensity values has decayed to a fixed fraction of the original value (typically 50 $\%$ for the half-dose, $D_{1/2}$). At 100 K, it is predicted that an upper dose limit exists, as an upper bound for the accumulated dose required to reduce the diffractive power of any macromolecular crystal to below 50 $\%$. An original $D_{1/2}$ dose limit for MX of 20 MGy was proposed by Henderson, conjectured from comparison with 2-dimensional electron diffraction studies conducted at 77 K using 100 keV incident electrons (Henderson, 1990). However, more recently, a value of $D_{1/2} =$ 43 MGy was experimentally derived by comparison of apoferritin with iron-rich holoferritin crystals, with holoferritin chosen since it represents an upper limit for the X-ray absorption cross-section of any known protein, with 2 amino acids per iron atom (Owen *et al.*, 2006). A resolution-dependent dose limit, $D_{1/2}(s) = s \times 10$ MGy, for resolution s Å has also been established (Howells *et al.*, 2009). Overall, these dose limits are all consistent to the same order of magnitude.

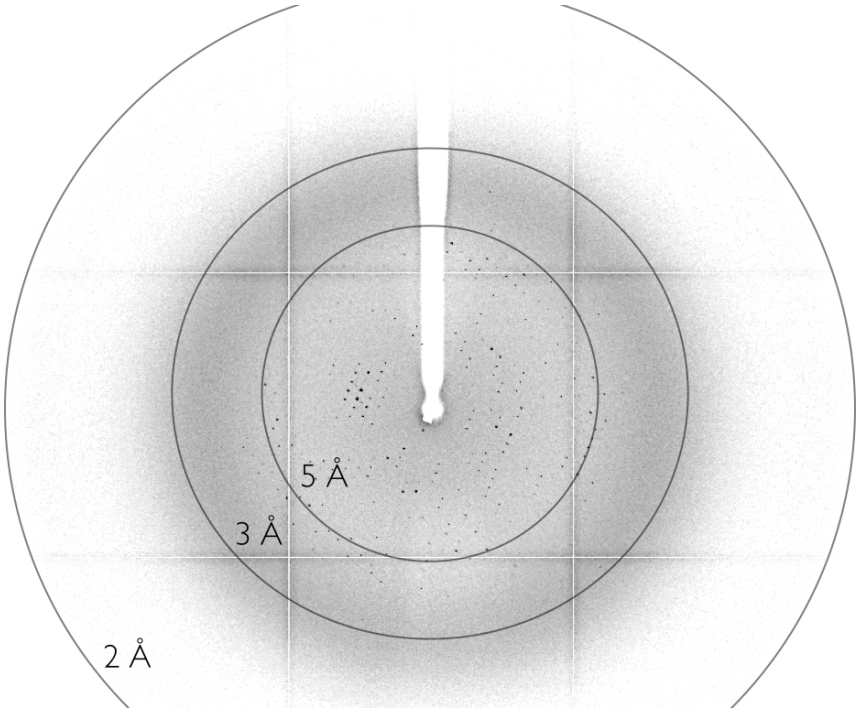
An intensity scaling model must be applied to account for all the above inconsistencies, caused by both experimental fluctuations and radiation damage, to put all intensities on an internally consistent scale prior to merging. Scaling and merging can be achieved by programs such as *AIMLESS* (Evans and Murshudov, 2013) and *XSCALE* (Kabsch, 2010), with the aim of minimising the intensity difference between symmetry-related reflections, and to output an average intensity value I_{hkl} and measurement of error σ_{hkl} for each Miller index.

Several models for intensity decay due to radiation damage have been proposed (Blake and Phillips, 1962; Sygusch and Allaire, 1988; Holton, 2009), however current scaling programs adapt a form of the decay model:

$$I_{hkl}(D) = I(h_d) \times K(D) \times \exp(-B(D){h_d}^2/2) \tag{1.1.8}$$



(a)



(b)

Figure 1.5: Diffraction images obtained from a crystal of Trp RNA-binding attenuation protein (TRAP) in complex with RNA at (a) low dose (11 kGy) and (b) high dose (50 MGy). The full data collection experiment will be reported later in Chapter 4. It was reported that the crystal originally diffracted to 1.98 Å. The crystal was in an identical orientation relative to the incident X-ray beam for both images, such that an identical region of reciprocal space had been sampled. Diffraction images were obtained on beamline ID14-4 at ESRF using an incident wavelength of 0.94 Å and an ADSC Q315R mosaic CCD detector at a distance of 304.5 mm from the crystal throughout the data collection. The angular resolution was fixed at $\Delta\phi = 1^{\circ}$ with an exposure time of 1 s per image frame.

where $I_{hkl}(D)$ is the expected intensity of a reflection I_{hkl} at dose $D > 0$, $I(h_d)$ is the expected intensity of a reflection at a reciprocal distance of h_d Å^{-1} from the diffraction pattern centre in the absence of any radiation damage, and $K(D)$ and $B(D)$ are dose-dependent scale and *B*-factors. A study by Leal *et al.* (2013) indicated the suitability of this model for 15 different protein crystals at RT down to a 60 % reduction in mean intensity. However, it should be noted that the analytical forms of $B(D)$ and $K(D)$ implemented in the Leal *et al.* (2013) model:

$$K(D) = C \exp(-\gamma^2 D^2) \tag{1.1.9}$$

and

$$B(D) = B_0 + D\beta \tag{1.1.10}$$

where C, γ, B_0 and β are contants to be experimentally determined, are not assumed in current scaling programs.

Following scaling, the program *CTRUNCATE* (Evans, 2011) can then be used to convert merged intensity values to structure factor amplitudes $|\mathbf{F}_{hkl}|$, with the use of Bayesian statistics to rectify intensity values that are close to zero (or below) as a result of background noise subtraction in the previous image integration step (French and Wilson, 1978).

1.1.3.3 Metrics to Assess Global Radiation Damage

The intensity statistics provided through the data reduction pipeline indicate the internal consistency of a sweep of diffraction images (Evans, 2011), and multiple metrics are accessible to monitor progressive radiation damage with dose. The $I_D/\sigma(I_D)$ metric has been deemed an unsuitable metric, since whereas the summed mean intensity I_D of a complete dataset collected at dose D MGy is anticipated to decrease with increasing dose, the standard deviation in intensity values $\sigma(I)$ is expected to increase, and consequently $I_D/\sigma(I_D)$ reduces at a more rapid rate than the diffracting power of the crystal. The I_D/I_1 metric[‖], where I_1 is the summed mean intensity of the first dataset collected on a crystal, has been proposed as an alternative, more suitable metric for intensity decay (Owen *et al.*, 2006).

[‖] I_D/I_1 is sometimes denoted as I_n/I_1 in the literature, depending on whether the dose $D > 0$ or dataset number $n > 1$ has been used to identify the intensities belonging to the higher dose dataset.

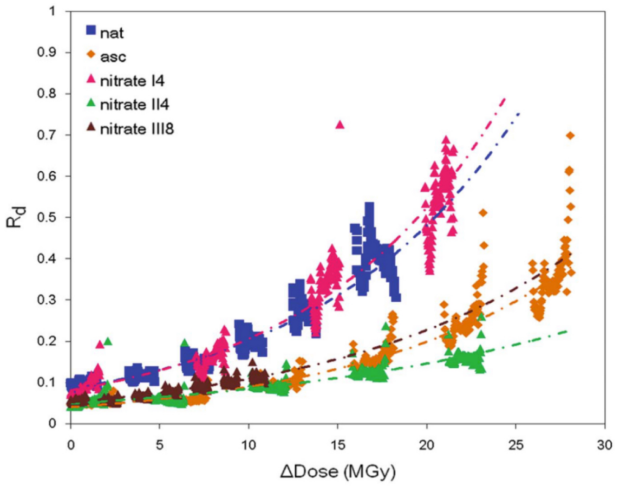
The relative *B*-factor, $B_{rel}(D)$, has also been proposed as a robust metric for radiation damage at 100 K (Kmetko *et al.*, 2006). $B_{rel}(D)$ is the change in the absolute isotropic *B*-factor obtained from a Wilson plot at dose D MGy and the first dataset, $B_{rel}(D) = B^{D}_{Wilson} - B^{1}_{Wilson}$. Absolute B_{Wilson} values cannot be measured directly from the Wilson plot, and in practice $B_{rel}(D)$ is derived explicitly by the scaling procedure implemented by the program *SCALEIT* to scale a higher reflection set to a low dose reference set (Howell and Smith, 1992). The *coefficient of sensitivity* $s_{AD} = \Delta B_{rel}/\Delta D 8\pi^2$ has additionally been defined as a measure for the dose-sensitivity of a crystal, and is proportional to the gradient of a plot of B_{rel} against dose.

The decay R-factor, $R_d(\Delta D)$, calculated between symmetry-related reflections recorded on diffraction images at a difference in dose ΔD, is another proposed metric to indicate the radiation sensitivity of protein crystals (Diederichs, 2006):

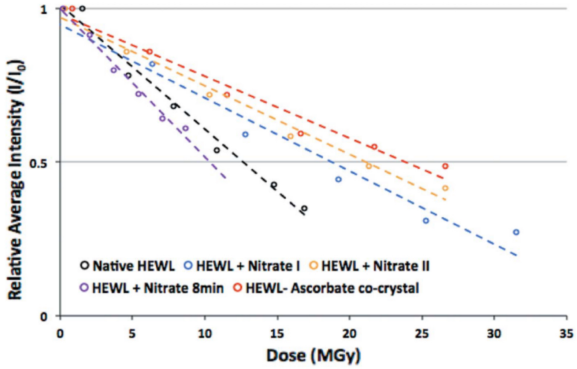
$$R_d = \frac{\sum\limits_{hkl}\sum\limits_{|i-j|=d}|X^i_{hkl} - Y^j_{hkl}|}{\sum\limits_{hkl}\sum\limits_{|i-j|=d}(X^i_{hkl} + Y^j_{hkl})/2} \tag{1.1.11}$$

where X^i_{hkl} and Y^j_{hkl} are the intensities of reflections with Miller index hkl with reflection centroids located on image numbers i and j respectively. The presence of a positive gradient in a plot of R_d versus ΔD is thus an indicator for progressive global radiation damage with dose. It has been observed that the R_d metric displays high variability between crystals of the same chemical composition (Nowak *et al.*, 2009; De la Mora *et al.*, 2011) (Figure 1.6 a).

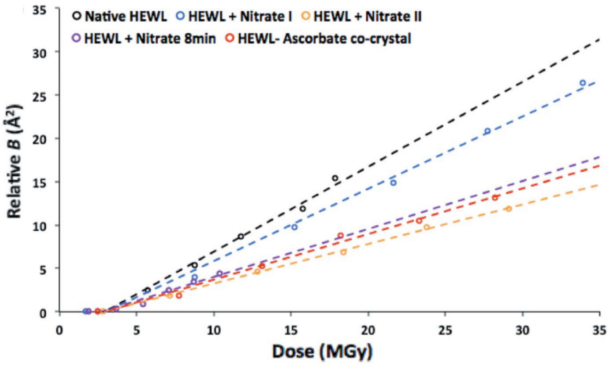
The I_D/I_1 and $B_{rel}(D)$ metrics have been most widely utilised to measure protein crystal radiation sensitivity, however they do not always yield consistent results. The change in I_D/I_1 with diffraction-weighted dose (DWD) has been shown to be invariant of incident beam size for insulin crystals at 100 K, whereas for the $B_{rel}(D)$ metric, such independence on beam size was less clear (Zeldin *et al.*, 2013a). Additionally, the reported efficacy of small molecule scavengers has varied depending on whether the I_D/I_1 or $B_{rel}(D)$ metric is used to assess radiation sensitivity (Figure 1.6 b-c) (Allan *et al.*, 2013; Kmetko *et al.*, 2011). It has also been observed that the I_D/I_1, $B_{rel}(D)$ and $R_d(\Delta D)$ metrics may yield different results when computed through different software packages (De la Mora *et al.*, 2011; Krojer and von Delft, 2011).



(a)



(b)



(c)

Figure 1.6: Variation in global damage metrics with increasing dose for 100 K lysozyme data originally reported in De la Mora *et al.,* (2011). In (a), the across-crystal variation in $R_d(\Delta D)$ metric for each scavenger condition indicates the non-reproducibility of the metric across crystals of lysozyme (De la Mora *et al.*, 2011). (b) Relative intensity I_D/I_1 and (c) $B_{rel}(D)$ metrics have been retrieved from Allan *et al.* (2013) upon reanalysis of the original De la Mora *et al.* (2011) data, indicating that the two metrics predict a different relative efficacy for each scavenger condition.

1.1.4 Phasing and radiation damage

In order to calculate the electron density equation 1.1.6, the phases $\phi(hkl)$ must be known in addition to the structure factor amplitudes $|\mathbf{F}_{hkl}|$. However the phase information is lost in the diffraction pattern derived from the native protein crystal, and additional techniques must be employed to extract the phases. This is known as the *phase problem* (Taylor, 2010). As will be discussed below, radiation damage is a significant hindrance to phase determination.

1.1.4.1 Molecular replacement

Molecular replacement is a widely exploited technique for cases in which there already exists a previously solved model that is similar in structure to the target protein. Typically, over 25 % sequence identity to the target protein is required, in addition to a C_α backbone rmsd of < 2 Å between the molecular replacement search model and target protein (Taylor, 2010). In a 2013 study, it was estimated that over 60 % of structures deposited in the PDB to that date had been solved by molecular replacement; however this number is exponentially increasing with time (Scapin, 2013).

Molecular replacement traditionally involves calculation of a series of inter-atomic vector maps called Patterson maps for both the structural homologue and target crystal. Crucially, a Patterson map can be calculated for an unknown structure from only diffraction intensities and no phase information. First a rotation search is performed to match the orientation of the homologue with that of the target protein in the unit cell. This is performed by determining the rotation matrix that best rotates the Patterson map for the homologue structure to match the subset of peaks nearest to the origin of the Patterson map derived from the crystal (which correspond only to intra-molecular atomic vectors). A translation search is then performed to determine the position of the target protein in the unit cell relative to its cell origin whilst accounting for any space group symmetry operations. In practice this is achieved by matching the Patterson map for the homologue (now calculated including inter-molecular vectors between symmetrically related copies) and the Patterson map derived from the crystal.

The program *PHASER* (McCoy *et al.*, 2007) instead utilises a maximum-likelihood approach to

locate the most probable orientation and position of a molecule, as a more robust method that explicitly accounts for both experimental and model errors, which Patterson methods omit (Evans and McCoy, 2008).

To the author's knowledge there have been no explicit studies assessing the effect of radiation damage on the success of molecular replacement. However, it has long been known that full-body rotation of a molecule in the unit cell can occur with dose (Ravelli and McSweeney, 2000), and this would thus directly affect the Patterson map rotation function described above at different doses.

1.1.4.2 Experimental phasing

In cases where no suitable homologous model exists, experimental methods must be deployed to derive phases. Most typically, knowledge of the location a heavy atom substructure (either incorporated into a crystal or intrinsic to the native protein) can be used to determine the phases for the whole protein structure.

In isomorphous replacement techniques (either single isomorphous replacement, SIR, or multiple isomorphous replacement, MIR), the incorporation of heavy atoms in a crystal, whilst preserving isomorphism to the native protein crystal, can lead to measurable changes in the intensities of reflections (and consequently structure factor amplitudes) between the native, $|\mathbf{F}_P|$, and derivatized crystal, $|\mathbf{F}_{PH}|$ (Perutz, 1956). The approximation $|\mathbf{F}_H| \simeq |\mathbf{F}_{PH}| - |\mathbf{F}_P|$ is required to provide an initial estimation** of the heavy atom structure-factor amplitude $|\mathbf{F}_H|$, from which direct methods can then be employed to determine the location of each heavy atom, and consequently the heavy atom phases ϕ_H. The heavy metal incorporation that is required for SIR and MIR significantly increases the overall crystal X-ray absorption cross-section, leading to a faster accumulation of absorbed dose for the derivatized crystal at a particular incident beam wavelength (Pike *et al.*, 2016). It is thus difficult to determine whether diffraction intensity changes are due to derivatization or unwanted radiation damage (Owen and Sherrell, 2016).

**This approximate relationship holds since $\mathbf{F}_{PH} = \mathbf{F}_P + \mathbf{F}_H$, assuming that isomorphism is preserved between the native and heavy atom derivatized protein crystal. The approximation is most easily visualised by displaying the vectors $\mathbf{F}_P$, $\mathbf{F}_{PH}$ and $\mathbf{F}_H$ on an Argand diagram, and derived using the cosine rule $|\mathbf{F}_H|^2 = |\mathbf{F}_{PH}|^2 + |\mathbf{F}_P|^2 - 2|\mathbf{F}_{PH}| \cdot |\mathbf{F}_P| \cos(\phi_P - \phi_{PH})$, whilst assuming that $\phi_P \approx \phi_{PH}$ holds.

In anomalous dispersion techniques, the crystal is irradiated at an incident beam energy close to the absorption edge of a particular heavy atom constituent within the protein crystal, at which Friedel's law $|\mathbf{F}_{hkl}| = |\mathbf{F}_{-h-k-l}|$ breaks down, due to differences in the absorption (f'') component of the atomic scattering factor ($f(\theta,\lambda) = f_0(\theta) + f'(\lambda) + if''(\lambda)$) between Bijvoet pairs[††]. These anomalous differences, $\Delta F^{\pm} = |\mathbf{F}_{PH}(+)| - |\mathbf{F}_{PH}(-)|$, can be exploited to determine the heavy atom contribution to the scattering. In multiple anomalous diffraction/dispersion, MAD, diffraction data are collected at multiple wavelengths, typically above, below and at the absorption peak for an anomalous scatterer, in order to maximise the dispersive (f') and absorption (f'') effects that give rise to isomorphous and anomalous differences.

Anomalous differences are typically small, and radiation-induced crystalline non-isomorphism throughout the collection of multiple MAD datasets can dominate anomalous signal. It has been demonstrated that a 0.5 % change in each dimension of a $100 \times 100 \times 100$ Å^3 unit cell (or alternatively, a 0.1 Å shift in the position of the macromolecule in the unit cell) could result in an average intensity change greater than 15 % at 3 Å resolution (Crick and Magdoff, 1956). It has since been simulated for the FcγRIII protein (with unit cell dimensions $a = 67$ Å, $b = 86$ Å, $c = 36$ Å) that a 0.5 % expansion in each unit cell dimension, and a 0.15 Å translation in the molecule could induce a median intensity change of 30 % (Owen and Sherrell, 2016).

Single-wavelength anomalous dispersion, SAD, only requires data collection at one single wavelength, but with the compromise that there is a resulting phase ambiguity (Taylor, 2010). Selenomethionine (Se-Met) incorporation into proteins has been widely exploited to achieve anomalous signal from selenium (Se-SAD), and the majority of third generation synchrotron MX beamlines are tuned to close to the Se K-shell edge (12.66 keV, 0.98 Å). Obtaining anomalous signal from sulphur in native protein (S-SAD) (or phosphorus in nucleic acids) removes the need to incorporate Se-Met or heavy metals into a target protein system, and is becoming more achievable at low-energy beamlines (for example I23 at Diamond Light Source (Wagner *et al.*, 2016)) that can reach the sulphur K-shell edge (2.47 keV, 5.02 Å) (Doutch *et al.*, 2012). Determination of the optimal wavelength for S-SAD must maximise the anomalous signal close to the S K-shell edge, whilst accounting for the increased radiation damage at lower incident beam energies due to the increased photoelectric absorption cross-sections of elements (Nave, 1995). The theoret-

[††] The dispersive component (f') causes only isomorphous differences not anomalous differences between Bijvoet pairs.

ical optimal S-SAD wavelength has been predicted to be dependent on the crystal size, and has been reported to exceed $\lambda = 2\,\text{Å}$ for crystals under $200 \times 200 \times 200\,\mu\text{m}^3$ (Wagner *et al.*, 2016). However, in practice the optimal wavelength has appeared to differ depending on the criteria used to assess data quality (Mueller-Dieckmann *et al.*, 2005). Furthermore, dose-dependent intensity changes between diffraction images can still dominate the anomalous difference $\Delta F^{\pm}$ between Bijvoet pairs of reflections in a single SAD dataset (which is typically only between $1-5\,\%$ (Terwilliger *et al.*, 2016)), unless precautions are taken to ensure that Bijvoet pairs are recorded on the same diffraction images at equivalent doses (de Sanctis *et al.*, 2016).

1.1.5 Model building and refinement

Once a set of structure factor amplitudes $|\mathbf{F}_{hkl}|$ and corresponding phases $\phi(hkl)$ have been derived, an electron density map $\rho(xyz)$ can be calculated using the electron density equation 1.1.6, to quantify the magnitude of electron density at each point (x, y, z) in the crystal unit cell. An atomic representation of the macromolecule can then be built into this electron density map.

The atomic model information is contained in a `.pdb` coordinate file, which provides the x, y and z Cartesian positions of atoms as well as their atomic displacement factors (B-factor) and atomic occupancies. The B-factor (Å^2) provides a measure of the uncertainty of position of the nucleus of an atom, and the occupancy of an atom quantifies the fraction of unit cells in the crystal for which the atom is present at that position. For protein crystals, typically the majority of the protein atoms have occupancy of 1, however bound ligand cofactors frequently have much lower values.

Most typically, atomic positions are built with the aid of likelihood-weighted $2mF_{obs} - DF_{calc}$ maps, calculated explicitly at position (x, y, z) as the Fourier summation:

$$\rho_{2mF_o-DF_c}(x,y,z) = \frac{1}{V}\sum_{hkl}\left[2m|\mathbf{F}_{obs}| - D|\mathbf{F}_{calc}|\right]e^{i\phi_{calc}} \cdot e^{-2\pi i\cdot[hx+ky+lz]} \tag{1.1.12}$$

where V is the volume of the unit cell and ϕ_{calc} is the calculated phase set derived from a known coordinate model. The figure of merit weighting $m \in [0,1]$ and multiplier $D \in [0,1]$ downweight the experimentally-derived and calculated structure factor amplitudes, $|\mathbf{F}_{obs}|$ and

$|\mathbf{F}_{\text{calc}}|$, respectively, in order to account for the rms error in the true phase value (Blow and Crick, 1959). This map formulation has been found to minimise model bias, in cases where the ϕ_{calc} phase set can bias the electron density towards the model instead of the experimental data, and introduce false artefacts in the electron density map (Read, 1986). In particular, these weightings maximise the heights of features from parts of the structure that have not yet been built. Correct placement of atoms can be assisted using a similarly defined $\text{m}F_{\text{obs}} - \text{D}F_{\text{calc}}$ difference map, which can directly indicate regions where too much or too little electron density has been built into the model.

Highly optimised refinement programs now exist to maximise the agreement of the atomic model with the experimentally observed data (such as structure factor amplitudes $|\mathbf{F}_{\text{obs}}|$ and experimental phase information) (Tronrud, 2004). The most modern refinement programs (for example *REFMAC* (Murshudov *et al.*, 2011), *phenix.refine* (Adams *et al.*, 2010) and *BUSTER-TNT* (Blanc *et al.*, 2004)) utilise maximum-likelihood based approaches to minimise a target function f_{total} of the form:

$$f_{\text{total}} = f_{\text{geometry}} + w \cdot f_{\text{xray}} \tag{1.1.13}$$

where the f_{geometry} and f_{xray} components are influenced by the geometry of the macromolecule and the experimental diffraction observations respectively. Minimisation is achieved by optimising the parameters of the model (atomic x, y, and z positions, *B*-factor, occupancy), whilst accounting for stereochemical restraints such as bond angles, bond lengths and torsional angles that have been widely documented from high-resolution small molecule studies (Allen, 2002) and extended to protein structure refinement (Engh and Huber, 1991). The weighting parameter w assigns the relative contribution of each component, enabling prior knowledge of ideal protein geometry to be upweighted in cases where only low resolution intensity information is available (Murshudov *et al.*, 2011).

Multiple rounds of refinement are performed, coupled with manual inspection of model to density map fit (with the aid of visualisation software such as *COOT* (Emsley *et al.*, 2010)) to manually update the positions of ordered solvent molecules and correct the orientations of side-chains that the automated refinement has struggled to assign to electron density. The validity of the built model relative to the experimentally-determined data can be assessed through the use of the

R-factor metric:

$$R_{\text{work}} = \frac{\sum_{hkl} \left| |\mathbf{F}_{\text{obs}}^{hkl}| - |\mathbf{F}_{\text{calc}}^{hkl}| \right|}{\sum_{hkl} |\mathbf{F}_{\text{obs}}^{hkl}|} \tag{1.1.14}$$

Where the above summations are over the *working set* of reflections that have been used in the refinement of the structural model. The R_{free} metric is used to monitor overfitting of data during refinement, using an identical formula to that of R_{work}, however calculated on a small *test set* of reflections (typically 5 %) that are kept aside and not included in refinement. For protein crystallography, the R_{work} and R_{free} metrics are typically around 20 % and much higher than those in small molecule crystallography (Holton *et al.*, 2014). As of early February 2017, $R_{\text{free}} \approx 0.23$ was the modal value over all refined structures deposited in the PDB to that date.

Model validation programs such as *MOLPROBITY* (Chen *et al.*, 2010) and *PROCHECK* (Laskowski *et al.*, 1993) are available, providing metrics such as Ramachandran score (Ramachandran and Venkatachalam, 1968), rms deviation from ideal bond length and angle, and steric clash predictions to assess the final stereochemical properties of a built model. Additionally, prior to model deposition, the PDB offers a rigorous validation service to complement those performed by the depositor (Read *et al.*, 2011).

Radiation-induced changes to localised regions of protein structures have been widely reported to occur in protein crystals at 100 K, leading to artifacts in electron density maps and disrupting the validity of resulting refined models. A detailed review of such *specific damage* events will be presented in the next section.

1.2 Specific damage in protein crystallography

1.2.1 Specific damage observations in proteins

The first predictions that X-ray irradiation might induce localised changes to the chemical configuration of crystalline proteins were made by Blake & Phillips (1962) during an experiment conducted on sperm-whale myoglobin crystals at RT, however the authors did not yet have access to a refined structure or amino acid sequence to support these claims. The first published evidence for structural damage in protein crystals was presented much later in 1988, for

des-pentapeptide insulin crystals irradiated at a second generation synchrotron at RT (Helliwell, 1988). In that study, by collecting multiple increasing dose datasets from a single crystal, difference Fourier maps $(F_{\text{damaged}} - F_{\text{undamaged}})e^{i\alpha_{\text{calc}}}$ revealed the opening of the disulphide bridges between successive dose states. That study also suggested that comparable structural changes had been observed in other systems, namely the oxidation of disulphide bonds and the opening of aromatic rings in ribonuclease, however that work was never published (S.K. Burley, G. Petsko & D. Ringe, unpublished).

At modern third generation synchrotrons, the first systematic studies on protein crystals at 100 K were published in 2000. Site specific damage was observed for *Torpedo californica* acetylcholinesterase (*Tc*AChE), hen egg white lysozyme (HEWL) (Weik *et al.*, 2000), winged bean chymotrypsin inhibitor (WCI) (Ravelli and McSweeney, 2000) and myrosinase (Burmeister, 2000) by collecting series of increasing dose complete diffraction datasets from single crystals. Analysis of $2mF_{\text{obs,n}} - DF_{\text{calc}}$, $mF_{\text{obs,n}} - DF_{\text{calc}}$, $3F_{\text{obs,n}} - 2F_{\text{calc}}$ and $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps (where $F_{\text{obs},n}$ corresponds to the set of structure factors corresponding to dataset $n \geq 1$ in the dose series), all indicated localised electron density loss at disulphide bridges, the carboxylates of glutamate and aspartate residues, and the methionine sulphur (Weik *et al.*, 2000; Ravelli and McSweeney, 2000), as well as for tyrosine aromatic -OH groups in myrosinase (Burmeister, 2000). Examples of such damage events in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps have been illustrated for the case of *Tc*AChE in Figure 1.7. Observations from such maps were also supported with *B*-factor (Weik *et al.*, 2000) and occupancy analysis (Burmeister, 2000). In addition, for HEWL, $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps illustrated the clear reorientation of cysteine sulphurs following disulphide bridge cleavage (Weik *et al.*, 2000). The reproducible preferential damage to disulphides, carboxylates and methionine residues has since been reinforced by specific damage observations on further non-metalloprotein systems at 100 K (e.g. Leiros *et al.* 2001; Nukaga *et al.* 2003; Borek *et al.* 2007; Juers and Weik 2011; Matsui *et al.* 2002; Fioravanti *et al.* 2007).

Radiation-induced reduction of metal centres has also been widely reported for metalloproteins at 100 K, at lower doses than disulphide bridge cleavage. X-ray absorption spectroscopy (XAS) has long been utilised to detect X-ray induced metal photoreduction (Penner-Hahn *et al.*, 1989), and can lead to notable changes in electron density at metal-ligands as binding site coordination changes with dose (Schlichting *et al.*, 2000). The oxygen-evolving Mn_4Ca active site of

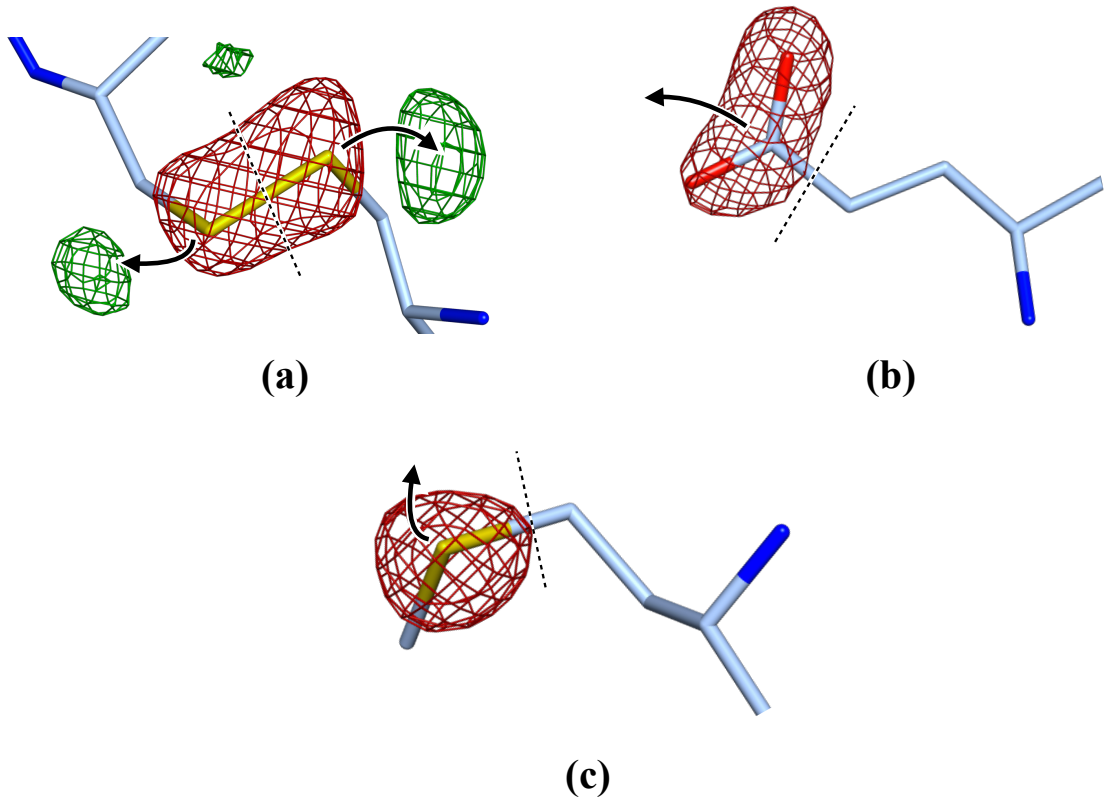


Figure 1.7: (a) Disulphide bond cleavage, (b) Glu decarboxylation, and (c) Met sulphur disordering within *Torpedo californica* acetylcholinesterase (*Tc*AChE) at 100 K (Weik *et al.*, 2000). $F_{\text{obs},5} - F_{\text{obs},1}$ Fourier difference maps calculated dataset 1 and 5 collected on the same crystal are shown, corresponding to PDB accession codes *1qid* and *1qih* respectively. All maps are contoured at $\pm 4\sigma$. The doses for dataset 1 and 5 were originally reported to be of the order of 10 MGy and 50 MGy respectively. Negative difference density (red) indicates disordering of the atomic positions with accumulated dose. In (a) the positive difference density (green) indicates the reorientation of the left sulphur atom to occupy a new well-defined conformation at the higher dose dataset. Dashed lines indicate radiation-induced bond cleavage and arrows indicate the reorientation/dislocation of subsequent products.

photosystem II has been systematically monitored through XAS, with the reduced Mn(II) state present at doses < 1 MGy (Yano *et al.*, 2005). Photoreduction of the Fe_2S_2 active site of putidaredoxin has been reported at doses as low as 0.35 MGy at 110 K (Corbett *et al.*, 2007). It is thus predicted that the functionally-relevant oxidation states of many redox metalloenzymes may be incorrect when solved by standard crystallographic techniques (Berglund *et al.*, 2002; Carugo and Djinovic Carugo, 2005).

X-ray irradiation inevitably triggers light-induced structural changes to photosensitive proteins (Baxter *et al.*, 2004). The true intermediate states of the light-driven proton pump bacteriorhodopsin have been difficult to characterise, due to the photoreactivity of the retinal Schiff base (Matsui *et al.*, 2002). For phosphoserine aminotransferase, functionally-relevant strain at the pyridoxal-5'-phosphate active site has been observed to decrease under X-ray irradiation (Dubnovitsky *et al.*, 2005). In a recent study, loss of electron conjugation on the chromophore of the Green Fluorescent Protein (GFP) homologue mNeonGreen has been detected as low as 0.1 MGy and it has thus been highly recommended that a radiation damage study follow structural elucidation of any fluorescent protein (Clavel *et al.*, 2016).

In selenomethionine-derivatized protein crystals, the Se atom sites are highly prone to electron density disordering in $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps (Leiros *et al.*, 2006). X-ray absorbance near-edge spectrum (XANES) measurements have been used to probe the breakage of the selenomethionine C_γ – Se covalent bond, indicating a minimum half-decaying dose of 5 MGy for crystals of the GCN4-N16A-p1 peptide at 100 K (Holton, 2007). Such specific damage significantly hinders SAD and MAD phasing techniques that utilise Se anomalous signal, since it must be assumed for them that the Se contribution to the structure factor vector sum remains constant with dose (Holton, 2009). It has however been demonstrated that SAD phasing at the Se peak anomalous wavelength can allow phases to be derived in cases where Se site-specific damage proceeds too rapidly for successful MAD data collection (Rice *et al.*, 2000).

1.2.2 Specific damage mechanisms

It has long been predicted that photoelectrons and the large number of low-energy secondary electrons (Auger electrons, recoil electrons) produced following initial ionisation by the incident

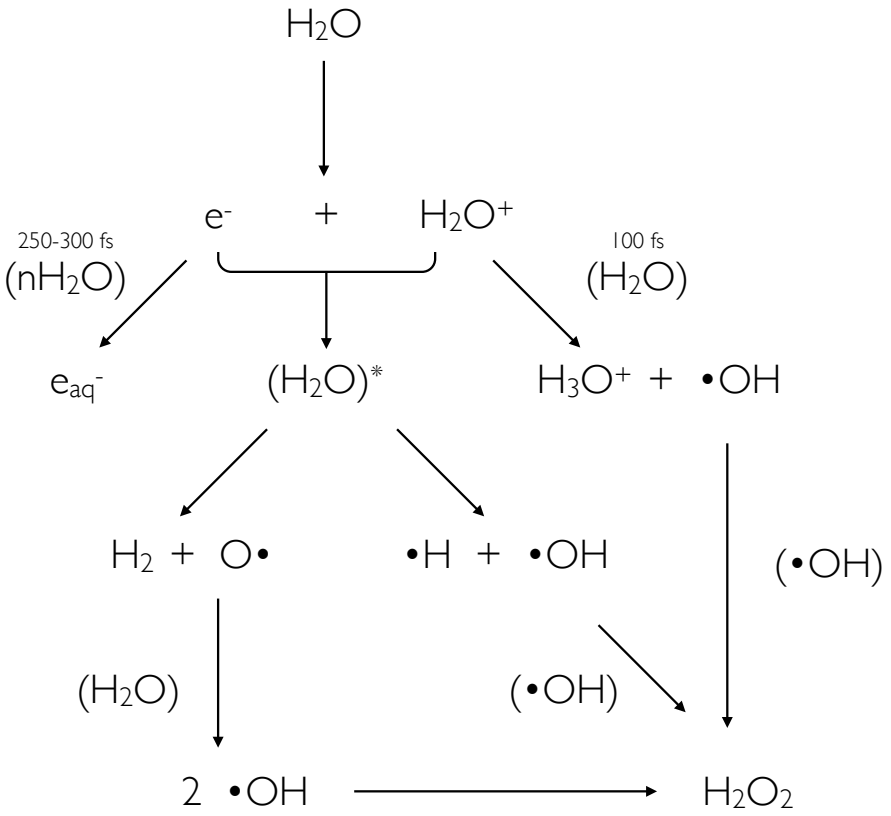


Figure 1.8: The decomposition scheme for H_2O following initial absorption of energy by the H_2O molecule. Figure has been reproduced from Hiroki *et al.* (2002) in the context of γ-radiolysis of water. Ionisation of water produces an electron and a water cation, which is rapidly followed by either hydration of the electron (producing e_{aq}) and proton transfer by the water cation, or dissociative recombination of the electron-cation pair. Both pathways lead to the formation of OH radicals, and the eventual formation of hydrogen peroxide through combination reactions of hydrogen peroxide with other reaction intermediates.

X-ray beam, must deposit their kinetic energy along tracks throughout the crystal, leading to the production of free radicals, namely hydroxyl radicals, electrons and hydrogen atoms in significant yields[‡‡], through radiolysis in bulk solvent channels on the picosecond time scale (Nave, 1995; Garrison, 1987). The available pathways for the radiolysis of water are shown in Figure 1.8.

It is predicted that the reactive hydroxyl radical is immobile below 110 K (Owen *et al.*, 2012), and the main damaging species for MX experiments at 100 K is anticipated to be low energy electrons (LEEs) with energies of only several eV (O'Neill *et al.*, 2002). Electrons have been shown to be mobile at 77 K through electron spin resonance (ESR) spectroscopy studies (Jones *et al.*, 1987), and have been postulated to rapidly quantum tunnel along protein backbone (and stacked base

[‡‡]The yields of radicals formed are $G(\mathrm{OH}) \simeq 2.80$, $G(\mathrm{e_{aq}}) \simeq 2.65$ and $G(\mathrm{H}) \simeq 0.60$, where $G(X)$ indicates the number of chemical species X formed per 100 eV of energy absorbed by the total sample. These quoted G values have been experimentally derived for ^{60}Co γ-irradiation of 0.4 M H_2SO_4 aqueous solutions at RT and neutral pH (Ferradini and Jay-Gerin, 2000).

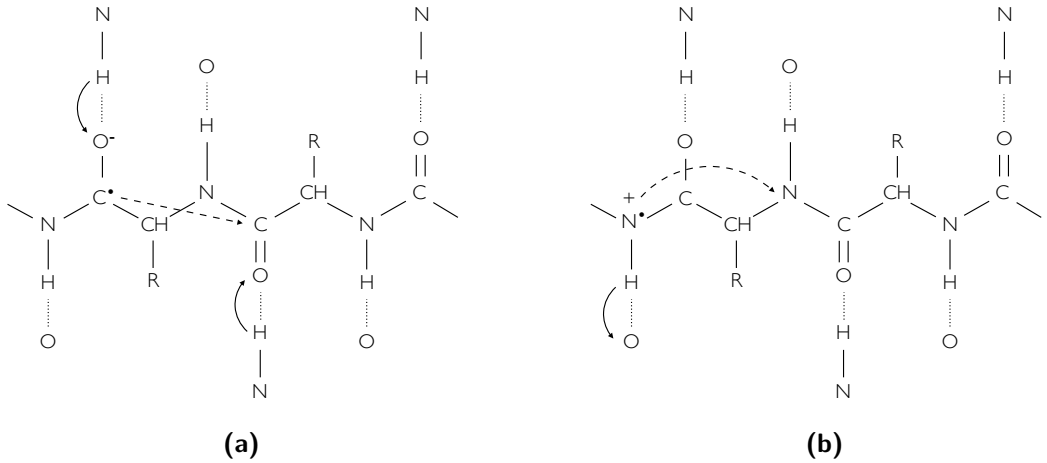


Figure 1.9: Proposed models for rapid electron and hole transfer along protein amino acid backbones between adjacent amido groups. Figure has been reproduced from Symons (1997) and was originally proposed as the transport mechanism in irradiated proteins and polypeptides in frozen aqueous solution (at 77 K). In (a), electron migration (dashed arrow) is suggested to be more rapid than the shown proton transfer (black arrows), and thus the electron can avoid trapping. In (b), hole transfer (dashed arrow) is slower than amide deprotonation (black arrow), and thus the hole is more rapidly trapped.

π-systems for nucleic acids) to relocate from the initial sites of ionisation (Figure 1.9) (Symons, 1997). Electron loss centres (holes) also are believed to be mobile at 100 K, but more easily trapped than electrons, predominantly by deprotonation of the amide backbone and carboxyl groups in proteins, and to form $G^{\bullet+}$ centres in nucleic acids (Jones *et al.*, 1987). Electrons can be trapped at electron-affinic sites in proteins, such as carbonyl groups and disulphide bridges, and to form $T^{\bullet-}$ and $C^{\bullet-}$ centres in nucleic acids (Symons, 1997).

For MX data collection at RT, reactive hydroxyl radicals are presumably mobile however, and are able to attack solvent exposed regions of protein. Hydroxyl radicals can abstract hydrogen atoms from C−H and N−H bonds, leading to the formation of carbon and nitrogen centred radicals (Southworth-Davies *et al.*, 2007). Hydroxyl radicals can additionally add to unsaturated bonds, for example in aromatic ring side-chains (O'Neill *et al.*, 2002). The yield of hydroxyl radicals is approximately equal to that of hydrated electrons. Hydrated electrons can interact with oxygen to produce superoxide ($O_2^{\bullet-}$) and hydroperoxide ($HO_2^{\bullet}$) radicals that can then also directly attack the protein (Hart, 1964). Despite the high yield of reacting radicals, it has been suggested that formation of these might be outrun through the use of high-brilliance X-ray beams and fast readout modern detectors at modern synchrotron beamlines (Owen *et al.*, 2012).

Radiation chemistry literature has been consulted to provide damage mechanisms for the reported site-specific radiation damage in MX experiments at 100 K. In aqueous solution the disulphide

bond is susceptible to oxidation (by hydroxyl radicals for example):

$$(\mathrm{p})\mathrm{-CH_2SSCH_2}^{-}(\mathrm{p}) + \mathrm{OH}^{\bullet} \longrightarrow (\mathrm{p})\mathrm{-CH_2SS}^{\bullet+}\,\mathrm{CH_2}^{-}(\mathrm{p}) + \mathrm{OH}^{-} \tag{1.2.1}$$

and reduction by solvated electrons:

$$(\mathrm{p})\mathrm{-CH_2SSCH_2}^{-}(\mathrm{p}) + \mathrm{e}^{-} \longrightarrow (\mathrm{p})\mathrm{-CH_2SS}^{\bullet-}\,\mathrm{CH_2}^{-}(\mathrm{p}) \tag{1.2.2}$$

$$(\mathrm{p})\mathrm{-CH_2SS}^{\bullet-}\,\mathrm{CH_2}^{-}(\mathrm{p}) \longrightarrow (\mathrm{p})\mathrm{-CH_2S}^{-} + (\mathrm{p})\mathrm{-CH_2S}^{\bullet} \tag{1.2.3}$$

Microspectrophotometry measurements on irradiated lysozyme crystals have revealed changes in the optical spectra as a function of dose to suggest reduction by electrons (De la Mora *et al.*, 2011). The peak at 500 nm, corresponding to the trapping of radiation-induced secondary hydrated electrons, was observed to diminish with dose to be replaced by a 400 nm peak that was attributed to the formation of the disulphide radical anion. This is in good accord with measured elongated bond lengths of disulphides in irradiated acetylcholinesterase crystals, which are consistent with the disulphide radical anion (Weik *et al.*, 2002). Electron paramagnetic resonance and online UV-visible absorption microspectrophotometry have since been used to predict that disulphide radicalisation rapidly saturates below 1 MGy (Sutton *et al.*, 2013). Since disulphide bond cleavage is not typically reported at such low doses, that study proposed that each disulphide may require multiple consecutive ionisations to initiate cleavage.

Oxidative decarboxylation of amino acids is also well documented in aqueous radiation chemistry literature (Sevilla *et al.*, 1979). Electron loss in a zwitterionic amino acid can lead to $\mathrm{C}_{\alpha}\mathrm{-CO_2}$ bond lengthening at the C-terminus and eventual loss of $\mathrm{CO_2}$ (Wisniowski *et al.*, 2002). In peptide derivatives the electron loss centre is preferentially located instead at the side-chain carboxyl group of Glu & Asp residues (Garrison, 1987):

$$(\mathrm{p})\mathrm{-CH_2CH_2COO}^{-} \rightleftharpoons (\mathrm{p})\mathrm{-CH_2CH_2COO}^{\bullet} + \mathrm{e}^{-} \tag{1.2.4}$$

$$(\mathrm{p})\mathrm{-CH_2CH_2COO}^{\bullet} \longrightarrow (\mathrm{p})\mathrm{-CH_2CH_2}^{\bullet} + \mathrm{CO_2} \tag{1.2.5}$$

In protein crystals, it is assumed that the above mechanism is mediated by positive hole migration to the carboxyl group, followed by rapid proton transfer to trap the hole at the carboxyl group.

For irradiated solutions of methionine amino acids and peptides, degradation of methionine has been predicted to occur by both direct scission of the $-SCH_3$ bond and also by oxidation to produce predominantly methionine sulfoxide (Shimazu *et al.*, 1964). Initial oxidation products are then believed to enter irreversible pathways to ultimately lead to formation of α-amino-n-butyric acid as a major reaction product (Kopoldová *et al.*, 1967). Cleavage of the CH_3SH group is thus well documented.

Tyrosine has also been reported to be radiation-susceptible in MX crystallography, and to undergo cleavage of the C−O bond of the aromatic ring (Burmeister, 2000). However, from a large body of radiation chemistry studies on both aqueous and solid state peptide samples, there appears to be no energetically favourably mechanism for C−O bond rupture in MX, and the tyrosyl ($TyrO^{\bullet}$) radical appears to be a robust radical species (Houée-Levin *et al.*, 2015). In Chapter 6 of the current thesis, the original reports of Tyr −OH cleavage in MX will be critically reanalysed, and the susceptibility of tyrosine will be revised.

1.2.3 Factors influencing specific damage events

1.2.3.1 Local protein environment effects

There is high interest in determining how local protein environment influences site-specific damage. Such insight would enable prediction of radiation-susceptible loci in protein structures for previously solved structures, and even prior to structure solution if knowledge about structural homologues is available.

It is unclear whether solvent accessibility plays a role in residue damage susceptibility. For disulphide bonds, it has been reported in several studies that the solvent exposed residues are particularly prone to breakage (Weik *et al.*, 2000; Meents *et al.*, 2010; Petrova *et al.*, 2010). Other studies have suggested that no such correlation exists (Homer *et al.*, 2011; Sutton *et al.*, 2013) since buried disulphides can also undergo cleavage (Ravelli and McSweeney, 2000). It has been reported for thermolysin that the formation of new secondary cysteine sulphur conformations correlates with solvent accessibility (Petrova *et al.*, 2010). For Glu & Asp residues, a lack of correlation between solvent accessible area and decarboxylation has been reported using $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$

difference map peak heights (Ravelli and McSweeney, 2000; Fioravanti *et al.*, 2007). Conversely, a systematic study on the protein data bank (PDB) using the B_{Damage} metric, designed to deconvolute atomic *B*-factor from atomic packing density, revealed a positive correlation between B_{Damage} and solvent accessibility for Asp, Cys, Glu, Met and Tyr residues (Gerstel *et al.*, 2015).

It has reported for the acidic residue-rich malate dehydrogenase structure that Glu and Asp residues interacting with the side-chains of serine, threonine and tyrosine are more readily decarboxylated than those interacting with lysine and arginine side-chains (Fioravanti *et al.*, 2007). Conversely, glutamate residues partaking in intermolecular Glu-Lys salt bridge interactions have been reported in a separate study to be most susceptible to electron density loss in $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ maps (Murray *et al.*, 2005). It has been suggested that carboxylate-Tyr hydrogen bonding interactions may be more labile under irradiation than their non-bonded counterparts (Burmeister, 2000; Leiros *et al.*, 2006). The type of disulphide bond (hook, spiral, staple, as categorised by the five dihedral angles that make up the bond) has been indicated to play a role in damage susceptibility, with staple-type disulphides most susceptible (Gerstel *et al.*, 2015). Staple-type disulphides have been reported to comprise of allosteric bonds (permitting control of protein function through conformational change by breaking open or forming), but not the majority of structural or catalytic disulphides (Schmidt *et al.*, 2006).

Active site Glu and Asp residues and disulphides appear notably susceptible to damage across a number of protein case studies (Fioravanti *et al.*, 2007; Roberts *et al.*, 2005). Conformational changes to active site histidine (Baxter *et al.*, 2004), serine (Leiros *et al.*, 2006) and lysine residues (Dubnovitsky *et al.*, 2005; Matsui *et al.*, 2002) in active sites have also been reported in $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps. It is suggested that functional flexibility and strain in active sites might be responsible for their high radiation-sensitivity (Fioravanti *et al.*, 2007). Photoreduction of metals occupying active sites may trigger nearby disulphide cleavage (Adam *et al.*, 2004). Conversely, Zn binding to an aspartate carboxyl group has been suggested to offer significant protection against radiation-induced decarboxylation (Burmeister, 2000).

There has been sustained interest in using small molecule additives to prevent site-specific damage (Allan *et al.*, 2013). The presence of ascorbate ions has been shown to mitigate disulphide bond breakage in HEWL at 100 K (Murray and Garman, 2002). Ordered nitrate ions have also been reported to act as an effective electron scavenger at 100 K, to be preferentially reduced over

adjacent disulphide bonds (Borek *et al.*, 2007; De la Mora *et al.*, 2011). This appears to be a short-range effect, since for nitrate-soaked crystals in which no ordered nitrate sites could be identified, no change in disulphide bond susceptibility was reported at 100 K (Kmetko *et al.*, 2011). The Kmetko *et al.* (2011) study did, however, report protection to the most radiation-sensitive disulphide in HEWL in nitrate-soaked crystals at RT, even in the absence of any ordered nitrate sites.

Despite suggestions that Glu and Asp decarboxylation susceptibility may be related to their protonation state, no correlation has been detected between $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map peak heights and theoretically calculated pK_{a} values of Glu/Asp carboxyl groups in a systematic study on malate dehydrogenase (Fioravanti *et al.*, 2007). However, a lower pH, particularly in the presence of ascorbic or nitric acid, has been shown to increase C_{γ} – Se bond lifetime in SeMet-derivatized protein (Holton, 2007). In that study, in order to explain the distribution of site-specific damage, it was hypothesised that mobile electrons might migrate along electric field lines surrounding the protein, and were thus diverted from certain SeMet sites.

1.2.3.2 The role of beam energy and dose rate at 100 K

It has been suggested that experimental conditions may also affect site-specific damage rates. The rate of electron density loss per unit dose associated with disulphide bond cleavage has been reported to be greater using a 14 keV incident X-ray beam compared to 9 keV at 100 K (Homer *et al.*, 2011). However, this was contradicted in another study on lysozyme, where no clear dependence between disulphide bond cleavage and beam energy was determined over a range of nine increasing incident beam energies between 6.5 and 33.0 keV (Shimizu *et al.*, 2007). At 100 K (and an incident beam energy of 12.6 keV), lower rates of dose deposition in crystalline samples have been suggested to lead to a small reduction in electron density loss at disulphide bridges in bovine trypsin (at 7.9 kGy s^{-1} versus 0.19 MGy s^{-1}) and SeMet in a selenomethionine-substituted maltooligosyltrehalose trehalohydrolase (at 48 kGy s^{-1} versus 0.48 MGy s^{-1}) (Leiros *et al.*, 2006).

1.2.3.3 Temperature effects

A reduction in data collection temperature below 100 K has been reported to reduce site-specific damage rates. At 50 K, disulphide bond cleavage in insulin has been shown to be reduced by an approximate factor of 4 relative to at 100 K, however a further reduction of damage at 5 K was not observed (Meents *et al.*, 2010). A two-fold increase in the dose-tolerance of disulphide bridges was detected for elastase crystals at 15 K relative to 100 K (Petrova *et al.*, 2010). Photoreduction of the Fe_2S_2 active site in putidaredoxin crystals has been reported to be eliminated by reducing the data collection temperature from 110 K to 40 K (Corbett *et al.*, 2007). Photoreduction of the Cu-site in plastocyanin was reported to reduce by a factor of 2 between 190 K and 4 K, but for frozen solutions as opposed to crystals (Penner-Hahn *et al.*, 1989).

Radiation-induced conformational changes to the catalytic triad at the active site of acetylcholinesterase have been detected at 155 K but not at 100 K, in addition to significantly increased damage to disulphides and methionine residues (Weik *et al.*, 2001). A large *B*-factor increase with dose has been reported for many additional residue types (Lys, Gln, Pro, Thr, Asn) in thermolysin at 160 K relative to 100 K (Juers and Weik, 2011). Additionally in that study, a correlation between atomic *B*-factor increase and solvent channel proximity was present at 160 K but not at 100 K. In agreement with this, seven solvent-exposed turns in thaumatin show increased radiation-induced flexibility above 180 K, but not at 155 K and below (as measured by *B*-factor increase with dose) (Warkentin *et al.*, 2012).

For room temperature data collection, disulphide bonds appear particularly radiation-sensitive in HEWL, and a lower dose deposition rate (6.7 Gy s^{-1} versus 9.7 Gy s^{-1}) was observed to increase electron density loss around disulphides in $F_{obs,n} - F_{obs,1}$ maps above 1 MGy (Southworth-Davies *et al.*, 2007). That study provided little evidence for damage to other residue types (e.g. Glu & Asp decarboxylation), however it could not be ruled out that at RT the lack of damage observations in $F_{obs,n} - F_{obs,1}$ maps could be attributed to the rapidly degrading data quality due to global damage. In a recent study on insulin at RT, disulphide bond cleavage was not indicated in $mF_{obs,n} - DF_{calc,n}$ maps up to 0.57 MGy, and disulphide bond cleavage was suggested to not be a temperature dependent reaction (Roedig *et al.*, 2016).

1.2.4 Exploitation of specific radiation damage

Attempts have been made to exploit the changes in structure factor amplitudes owing to site-specific damage. Radiation-damage induced phasing (RIP) techniques have proved capable of deriving experimental phases, by treating the data as a single isomorphous replacement phasing problem, with the low-dose undamaged protein as a 'derivatized'-protein and the higher-dose damaged protein as a 'native'-protein (Ravelli *et al.*, 2003). RIP initially proved successful by exploiting disulphide bond cleavage in typsin and base debromination in a small oligonucleotide (Ravelli *et al.*, 2003), and its applicability has since been demonstrated on a range of protein systems (Nanao *et al.*, 2005). In a similar technique, RIPAS (radiation-damage induced phasing with anomalous scattering), radiation damage to iodinated tyrosine residues has been shown to break the inherent SAD phase ambiguity for the iodine anomalous scatterers (Zwart *et al.*, 2004). RIPAS phasing of a mercury-derivatized haloacid dehydrogenase protein enabled elucidation of a structure, where previous SAD and MAD phasing attempts had failed due to rapid disruption of the Hg−Cys bonds in the protein (Ramagopal *et al.*, 2005). For radiation-sensitive samples in which site-specific damage occurs rapidly within a single dataset, it has been suggested that phasing with unmerged data (where each reflection is assigned a *dose stamp*), and treating specific damage as a continuous dose-dependent process, can improve the efficacy of RIP and RIPAS (Schiltz and Bricogne, 2007).

Knowledge of radiation-induced changes has proved beneficial for elucidating transient dynamics in certain proteins. The structure of tryparedoxin with its redox-active disulphide in a reduced form was only achievable by photoreduction using an intense X-ray beam provided by a third generation synchrotron (Alphey *et al.*, 2003). X-ray induced photoreduction has also provided a tool to probe the expansion of the superoxide reductase active site upon Fe(III) to Fe(II) reduction (Adam *et al.*, 2004), and to capture the redox intermediates of horseradish peroxidase (Berglund *et al.*, 2002). More recently, X-ray irradiation has proved beneficial to induce and monitor with accumulated dose the progressive nitrite turnover in copper nitrite reductase (Horrell *et al.*, 2016). In addition, disulphide bond cleavage in thaumatin crystals has been used to validate the Hadamard transform approach to improve time-resolution and signal-to-noise ratios in time resolved X-ray crystallography (TRX) experiments (Yorke *et al.*, 2014).

1.3 The aims of this thesis

This thesis aims to present novel research on specific radiation damage to macromolecules in MX experiments. In Chapter 2, a computational approach, *RIDL* (radiation-induced density loss), has been developed to objectively detect, classify and quantify specific damage events directly from $F_{obs,n} - F_{obs,1}$ Fourier difference maps, and thus directly improve on previous standard approaches for assessing specific damage. Such an approach provides a per-atom analysis of electron density change with increasing dose, and thus a significant information increase is immediately achievable.

Using this objective approach, experiments have been conducted to systematically investigate site-specific damage events to a range of both crystalline nucleic acids and nucleoprotein complexes. Radiation damage to nucleic acids has previously been poorly characterised, however the work in Chapters 3, 4 and 5 now aims to address this issue.

Chapter 6 will address the radiation susceptibility of tyrosine in MX. Whereas Tyr $-$OH cleavage from the aromatic ring has been reported in MX, this is in contrast with a large body of radiation chemistry literature in which the C$-$O linkage is highly robust to radiation effects. Using the newly developed analytical tool, RIDL, in Chapter 2, the previous reports of MX radiation-induced bond cleavage in Tyr will be critically re-analysed.

Chapter 7 will report further methods to best inform experimenters at what point during a diffraction data collection that the onset of significant site-specific damage has occurred. A method has be designed to calculate half dose values for the decay of radiation-susceptible groups in proteins (Cys, Glu and Asp), and such half doses have be calculated for various available MX damage series. In addition, a protocol has be developed using *RIDL* to enable detection of the onset of site-specific damage events occurring within a single sweep of diffraction images, all of which would be required for structural solution.

CHAPTER 2

Development of tools to investigate site-specific radiation damage

2.1 Introduction

At doses of several MGy, before an observable decay in the diffraction intensities, radiation-induced changes to localised regions of protein, including changes to oxidation state, bond scission, and conformational changes, have been reported for a range of model protein crystals at 100 K (Burmeister, 2000; De la Mora *et al.*, 2011; Nanao *et al.*, 2005; Ravelli and McSweeney, 2000). When undetected, specific damage events will inherently lead to modelling inaccuracies; this is of particular importance since active site regions appear to be particularly radiation-sensitive, and failure to identify damage can lead to poor assignment of cofactor geometry, and ultimately the derivation of false enzymatic pathways (Dubnovitsky *et al.*, 2005; Matsui *et al.*, 2002; Clavel *et al.*, 2016).

Specific damage has been typically studied by collecting multiple complete, non-overlapping MX diffraction datasets collected from a single crystal (Figure 2.1), and observing how the electron density evolves as a function of dose. $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps (where $n > 1$ is the index of a higher dose dataset) have proved to be an invaluable tool in determining the presence of site-specific damage in proteins, by permitting changes in electron density with dose to be directly visualised around individual atoms (Ravelli and McSweeney, 2000; Sutton *et al.*, 2013; Fioravanti *et al.*, 2007). In practice, $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps are calculated at Cartesian position (x, y, z) as:

$$\frac{1}{V}\sum_{hkl}\left[|\mathbf{F}_{\text{obs},n}| - |\mathbf{F}_{\text{obs,1}}|\right]\cdot\exp(i\phi_{\text{calc,1}})\cdot\exp(-2\pi i\cdot[hx+ky+lz]) \tag{2.1.1}$$

which is the Fourier transform of the difference in structure factor amplitudes between the initial and higher dose datasets $n > 1$, $|\mathbf{F}_{\text{obs,1}}|$ and $|\mathbf{F}_{\text{obs},n}|$ respectively, and typically using the calculated phases derived from the refined initial dataset coordinate model, $\phi_{\text{calc,1}}$. Here, V is the volume of the unit cell. $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps are a spatial average of the behaviour of electron density over all unit cells in a crystal, and as such these maps are only sensitive enough to changes that occur reproducibly across a non-negligible fraction of unit cells.

Difference map analysis has been a significant contributing factor in the establishment of a reproducible ordering of such specific damage events for non-metalloprotein crystals at 100 K (Weik

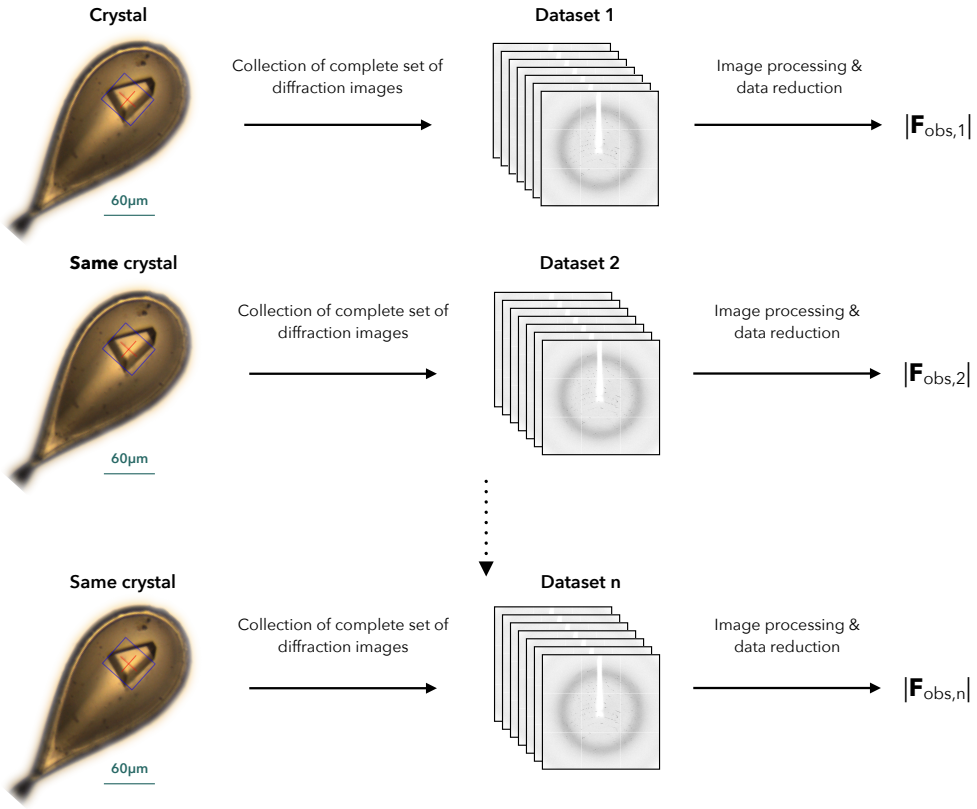


Figure 2.1: For collection of a typical MX damage series from a single protein crystal, $n > 1$ complete datasets of diffraction images are collected from the same crystal. Each diffraction image set is typically processed individually, to produce a set of corresponding structure factor amplitudes $|\mathbf{F}_{\text{obs,n}}|$ following data reduction. Each dataset of images must be non-overlapping, such that a diffraction image cannot be shared by adjacent datasets. For each dataset, the same region of the crystal can be exposed (by maintaining the same start and stop goniometer rotation angles, $\psi_{\text{start},n}$ and $\psi_{\text{end},n}$, for each dataset), or alternatively, in the scenario in which the crystal is homogeneously irradiated in a beam, the crystal can be continuously rotated such that $\psi_{\text{end},n-1} = \psi_{\text{start},n}$. Ideally, the beam transmission and energy, exposure time per image and $\Delta\psi$ angular resolution of the goniometer should not be varied throughout.

et al., 2000): (a) disulphide bond elongation and cleavage, (b) glutamate and aspartate decarboxylation, and (c) methionine methylthiol group loss. However, it is important to note that other protein moieties have been observed to undergo radiation-induced changes at 100 K (e.g. tyrosine: Burmeister (2000), selenomethionine: Leiros *et al.* (2006), lysine: Juers and Weik (2011), histidine: Fioravanti *et al.* (2007)), and consequently the entire coordinate model must be visually inspected for indications of damage, and subjective judgement made as to its significance. For larger, more complex structures, however, this process quickly becomes intractable.

Whereas cryotemperature data collection remains the dominant successful strategy for structure solution, an interest in data collection above 100 K (Owen *et al.*, 2012; Keedy *et al.*, 2015) has re-emerged in recent years, due to the potential loss of functionally-relevant protein conformational states as a result of the cryocooling process (Fraser *et al.*, 2011). This has been facilitated by advances in detector technology, particularly regarding reductions in their readout time (Owen *et al.*, 2014). For data collection above 100 K, specific damage has been far less characterised across proteins*, partly due to the faster deterioration of diffraction data quality with dose, which then severely hinders the derivation of electron density that is of sufficient quality to probe local radiation-induced changes between multiple dose states. Regardless of this limitiation, difference maps have still enabled the location of disulphide cleavage events in RT data collection (Southworth-Davies *et al.*, 2007; Kmetko *et al.*, 2011).

A challenge remains in how to quantify manifestations of damage in $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps. In practice, the difference map is stored as a finite grid of difference density values (which will be referred to as voxels) expanding over a pre-determined volume (typically the unit cell or asymmetric unit). The mean density value calculated over all map voxels is $\sim 0\,\mathrm{e/\AA^3}$, with a corresponding standard deviation denoted as σ. The σ-height of a difference map peak in close proximity to a protein residue (the number of standard deviations, σ, that the peak value attains above/below the map mean) has been most typically quoted as a metric for damage to that residue (e.g. Borek *et al.* (2007)). In particular, σ-heights have been used in attempts to assess correlations between damage to residues and their local environment (solvent accessibility, pKa) (Fioravanti *et al.*, 2007).

*Studies have indicated the presence of disulphide bond cleavage in proteins at RT for HEWL at 1.08 MGy (Southworth-Davies *et al.*, 2007) and at 0.12 MGy (Kmetko *et al.*, 2011), however for another study on Porcine insulin at 0.57 MGy, disulphide cleavage was not reported (Roedig *et al.*, 2016).

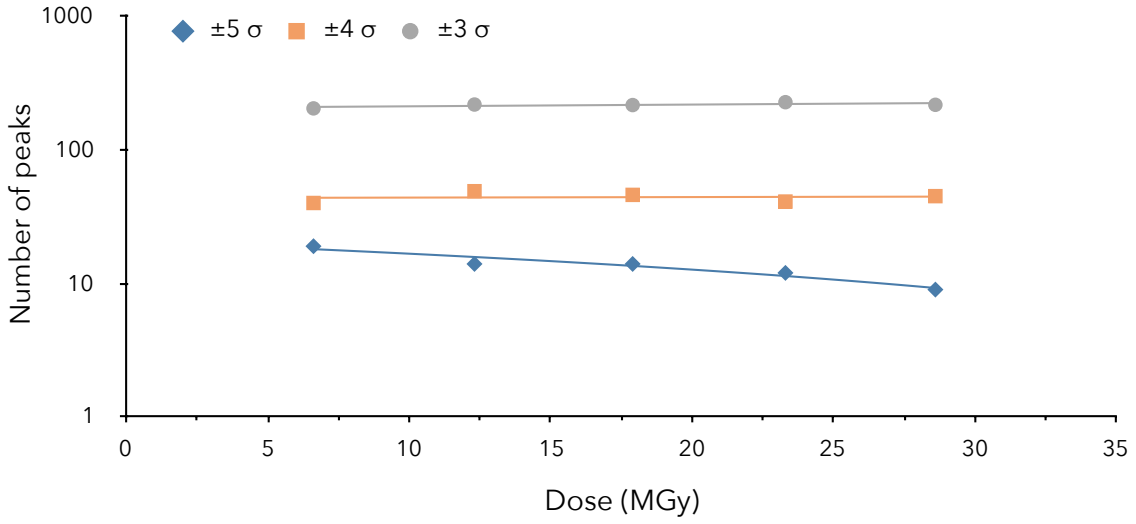


Figure 2.2: The variation in the number of detected peaks (local minima and maxima) detected in $F_{obs,n} - F_{obs,1}$ difference maps above $\pm 5\sigma$, $\pm 4\sigma$ and $\pm 3\sigma$ map thresholds, as a function of dose (calculated using *RADDOSE v2*) for the De la Mora *et al.* (2011) lysozyme damage series (deposited in the PDB with accession codes *2ybh*, *2ybi*, *2ybj*, *2ybl*, *2ybm* and *2ybn*). Coordinate models and structure factor amplitudes were retrieved from the *PDB_REDO* online server (Joosten *et al.*, 2014) prior to re-refinement and model rebuilding. Difference peaks were detected using the '*Difference map peaks*' tool available in *COOT* (Emsley *et al.*, 2010).

However, unfortunately σ-levels of difference map peaks cannot be rigorously compared between different MX radiation damage studies or even for different dose datasets collected from a single crystal, since the value of σ is dataset-specific[†]. To exemplify this, in the case of a lysozyme crystal damage series deposited in the PDB consisting of 6 increasing dose datasets between 2.3 and 28.6 MGy (De la Mora *et al.*, 2011), the number of $F_{obs,n} - F_{obs,1}$ map peaks detected above fixed 5 and 4 σ-thresholds (per-dataset) monotonically diminish and oscillate, respectively, as n increases (Figure 2.2). Setting a constant σ-threshold (here either ± 4 or $\pm 5\sigma$) to screen for damage sites risks neglecting informative difference peaks at higher doses.

2.2 Outline of chapter

In short, whereas $F_{obs,n} - F_{obs,1}$ maps provide a direct visualisation of site specific damage events in proteins, and have been fundamental in characterising the radiation chemistry at play, manual inspection is a highly subjective and time consuming process. This is compounded by the fact that maps are inherently noisy due to global damage effects with increasing dose (a problem that

[†]This is even for datasets that are scaled to a single reference dataset (usually the first dataset) due to decreasing $I/\sigma(I)$.

is particularly detrimental to room temperature studies). Furthermore, it is non-trivial to quantify damage using these maps, since the map information most easily accessible to the standard user (the σ-height of difference map peaks) is dataset-specific, and moreover provides no information of electron density change for atoms that are not adjacent to well defined difference peaks.

The aim of the work presented in this chapter is to improve upon these limitations of using difference maps, and to present the development of new tools that have been designed to investigate site specific radiation damage.

In Section 2.4, the first tool (*PEAKS*) is described, which primarily aims to assist the process of manual map inspection by providing an efficient protocol to systematically inspect difference map peaks for indications of true damage. The subjective bias of map inspection remains, however *PEAKS* is designed to improve the tractability of manual map inspection, when such an approach is desired.

In Section 2.5, the second tool (*RIDL*, radiation induced density loss) is presented. It was developed subsequently to *PEAKS* and aims to remove the subjective element of manual map inspection entirely, thus providing an objective high throughput and automated analysis of difference density around each atom in a structure. It should be noted that further extensions of the *RIDL* method have been implemented to achieve additional tasks and are detailed in other chapters in this thesis. These modifications will not be discussed in the current chapter.

A detailed breakdown of the functionality of each tool will be demonstrated for an MX damage series on a glycoside hydrolase protein, the structure of which was recently solved by MX. This protein was chosen due to it being of larger size than model proteins that have previously been characterised in the literature. It is thus presented as an example for which manual map inspection is deemed to be intractable. The crystallisation and data collection information for this series is presented in Section 2.3. The further application of the tools to other macromolecular systems can be found in later chapters.

2.3 Glycoside hydrolase family 7 data collection

The glycoside hydrolase family 7 (GH7) cellulase from *Daphnia pulex* (*Dp*Cel7B) was chosen for the purpose of the current study primarily due to its large size: 96.8 kDa per asymmetric unit, containing 2 monomers each of 445 amino acids. As such, this protein represents a case where the characterisation of specific damage through manual inspection of difference maps at multiple doses would have been challenging. Each asymmetric unit contains a large test set of traditionally radiation-sensitive residues (Cys: 36, Glu: 28, Asp: 74, Met: 22) which is ideal for verifying the development of damage detection tools. Moreover, the structure of the selected *Dp*Cel7B enzyme was only recently solved by crystallographic means at 100 K (J. McGeehan, unpublished), making the structure a novel candidate for specific damage analysis.

Cellobiohydrolase proteins such as GH7s hydrolyse cellulose chains to release cellobiose as their major product (Kern *et al.*, 2013). GH7 enzymes are endogenous in fungi, bacteria and arthopods, that assist in the digestion of lignocellulose to glucose, along with endoglucanases and β-glucosidases (Cragg *et al.*, 2015). GH7s are thus of high interest to studies concerned with improving the efficiency of lignocellulosic biomass deconstruction for biofuel production. The main focus of this chapter will be the development of novel damage detection tools. However in addition, the current investigation could provide a novel insight into the *Dp*Cel7B structure, particularly involving the radiation-sensitivities of interactions at the cellulose-degrading active site.

2.3.1 Crystallisation

As another major benefit of selecting the *Dp*Cel7B system, well-diffracting crystals (~ 2 Å) were already available and could be readily supplied by a collaborator, Prof. John McGeehan, University of Portsmouth, UK. The original *Dp*Cel7B protein had been supplied by the National Renewable Energy Laboratory (NREL), USA, having been overexpressed in a *Trichoderma reesei* system. The *Dp*Cel7B crystals were grown using a hanging-drop crystallization protocol with 2.75 mg ml^{-1} *Dp*Cel7B protein mixed with 0.1 M monosodium citrate and 0.9 M ammonium sulphate, pH 4. Crystals grew with approximately cuboid-dimensions; the crystal from which diffraction data were collected had approximate dimensions of $75 \times 75 \times 10\,\mu m^3$.

*Dp*Cel7B crystals were retrieved at the University of Portsmouth by the current author, where they were then looped and soaked for 2 min in buffer solution diluted with 30 % (v/v) glycerol (previously identified to be a suitable cryoprotectant (J. McGeehan, private communications)) immediately prior to storage in liquid nitrogen.

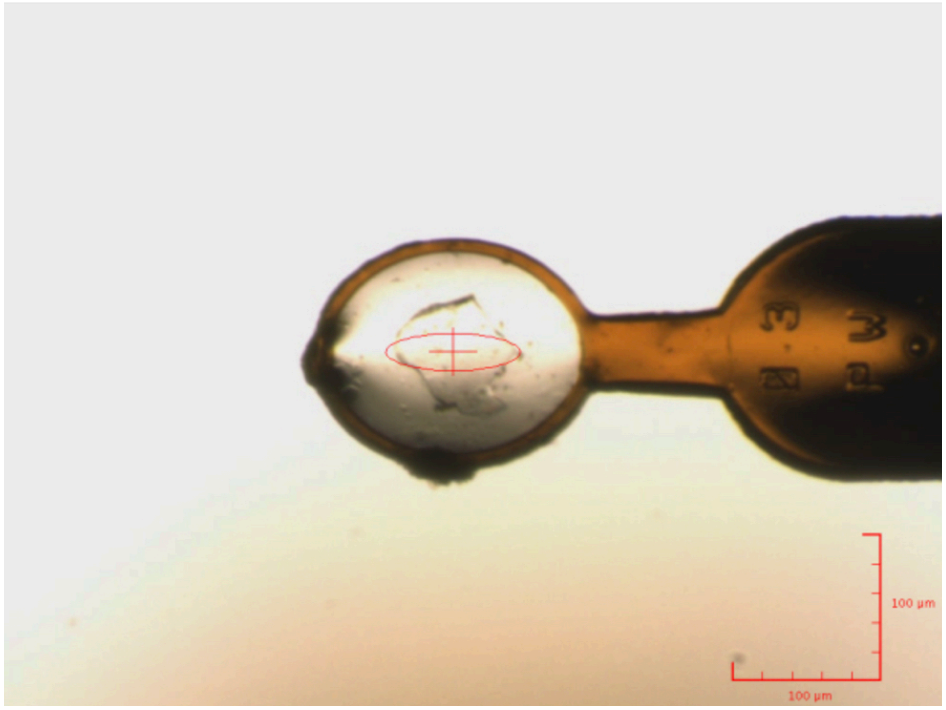
All subsequent diffraction data collection and analysis was performed by the current author.

2.3.2 X-ray data collection

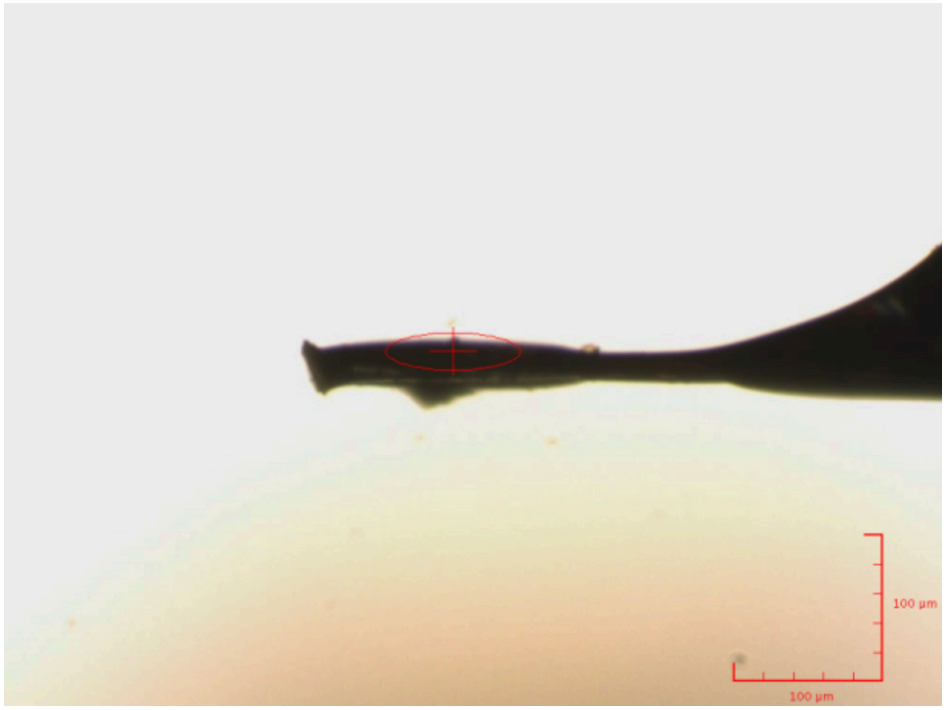
Diffraction data were collected from a single *Dp*Cel7B crystal at 100 K on beamline I02 at the Diamond Light Source, using an incident wavelength of 0.980 Å (12.7 keV) and a Pilatus 6M detector, positioned at a distance of 244.9 mm from the crystal throughout data collection. The beam passed through a 200 µm aperture and was slitted to 60 × 120 µm (vertical × horizontal), with an approximate Gaussian profile (vertical × horizontal full-width half-maximum (FWHM) of 17.7 × 104.8 µm). The flux at the sample position was determined using a pre-calibrated 500 µm-thick silicon PIN photodiode (Owen *et al.*, 2009) to be $\sim 1.77 \times 10^{12}$ ph/s at 100 % beam transmission. The crystal was initially orientated with the small crystal dimension ($z = 10$ µm) set parallel to the incident beam direction. A 2000° continuous sweep of data was collected, consisting of 9999 frames (with an angular resolution of $\Delta\phi = 0.2°$) of 0.04 s exposure time each. The beam transmission was held at 25 % throughout.

2.3.3 Diffraction data processing

The single of 2000° sweep of diffraction data was divided into 11 disjoint 180° wedges (which are sufficient for completeness due to the C121 space group), each consisting of 900 images. Each wedge of data was integrated using *DIALS* (Waterman *et al.*, 2016), and then scaled and merged in *AIMLESS* (Evans and Murshudov, 2013). To obtain an initial set of phases for the first dataset, molecular replacement was performed in *PHASER* (McCoy *et al.*, 2007), using an identical *Dp*Cel7B structure already deposited in the PDB (accession code: *4xnn*, resolution 1.9 Å, (McGeehan, unpublished)) as a search model. It is noted that molecular replacement was not a necessary step for the current data processing since an identical *Dp*Cel7B coordinate model (with 100 % sequence identity was present); a suitable procedure would have been to perform rigid-body



(a)



(b)

Figure 2.3: A crystal of *Dp*Cel7B enzyme in a nylon loop on beamline I02 at the Diamond Light Source. (a) Initial orientation of crystal with the smallest crystal dimension ($z = 10\,\mu m$) parallel to the incident beam direction, and (b) the orientation of the crystal after 18 s, corresponding to a 90° rotation of the loop in the beam path.

refinement (in e.g. *REFMAC5* (Murshudov *et al.*, 2011)) to refine the *4xnn* coordinate model directly against the $|\mathbf{F}_{\text{obs}}|$ structure factor amplitudes for the current *Dp*CelI7B crystal.

The resulting initial coordinate model was then refined using *REFMAC5*, first with 10 cycles of rigid body refinement, followed by repeated rounds of 10 cycles of restrained and TLS refinement, coupled with manual inspection in *COOT* (Emsley *et al.*, 2010). For TLS refinement, 10 groups were specified (chain A: residues 21-92, 93-284, 285-437, 438-464, and chain B: 21-134, 135-191, 192-235, 236-401, 402-437, 438-464) as determined using the *phenix.find_tls_groups* tool as part of the *PHENIX* crystallographic suite (Adams *et al.*, 2010). Local non-crystallographic symmetry (NCS) restraints were also included to account for the NCS between chains A and B in the asymmetric unit.

$2\text{m}F_{\text{obs,n}} - \text{D}F_{\text{calc}}$ density consistent with ordered glycerol molecules were found at the binding pocket of each monomer, in addition to at a crystal contact interface with the adjacent asymmetric unit. In each monomer, it proved challenging to fit the C_α-backbone of the solvent exposed loop, amino acids $121 - 125$. In the previous PDB *4xnn* reference model, the $2\text{m}F_{\text{obs,n}} - \text{D}F_{\text{calc}}$ map quality at these loop regions was also unconvincing, and did not provide a suitable guide as to the correct placement of the loop. Consequently, these residues were omitted from the current refined model. Final protein geometry was assessed with *Molprobity4* (Chen *et al.*, 2010).

Data reduction and refinement statistics are reported in Tables 2.2 and 2.3.

2.3.3.1 Discussion on refinement of higher dose datasets

In the subsequent radiation damage analysis that will be carried out in this chapter, the calculation of $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference map coefficients has used phases, $\phi_{\text{calc,1}}$, derived above for the initial dataset refined coordinate model, such that the experimentally-determined structure factor amplitudes, $|\mathbf{F}_{\text{obs,n}}|$, are the only variables with increasing dose. As such, there was no requirement to obtain more accurate phase sets, $\{\phi_{\text{calc},n}\}_{n>1}$, using coordinate-restrained refinement of models for higher dose datasets. Furthermore, full restrained refinement of high dose datasets appears to be unreliable, due to the diminished quality of the experimentally-collected diffraction data at high doses.

Table 2.1: Refinement summary statistics for each higher dose dataset of the *Dp*Cel7B protein test series, in which ten cycles of restrained refinement have been performed for each dataset in *REFMAC*, using the refined coordinate model corresponding to the first dataset, coupled with the structure factor amplitudes, $|\mathbf{F}_{\text{obs,n}}|$, for each higher dose dataset. With increasing dose, the geometry weighting automatically selected by *REFMAC* (the *weight matrix* parameter) is observed to decrease, such that tighter geometry restraints are selected to account for the diminishing quality of the diffraction data. The RMSD bond lengths and bond angle values are observed to fluctuate with increasing dose, but do not vary smoothly with dose.

Dataset	Weight matrix	RMSD bond length (Å)	RMSD bond angle (°)	R_{work}	R_{free}
2	0.36	0.019	1.81	0.16	0.20
3	0.33	0.018	1.79	0.16	0.21
4	0.35	0.019	1.85	0.16	0.21
5	0.28	0.018	1.80	0.17	0.21
6	0.27	0.018	1.83	0.17	0.22
7	0.23	0.018	1.84	0.18	0.22
8	0.22	0.018	1.81	0.18	0.22
9	0.17	0.017	1.81	0.19	0.24
10	0.19	0.018	1.82	0.19	0.23
11	0.14	0.017	1.79	0.20	0.25

Indeed, as shown in Table 2.1, unless care account is taken of the decreasing data quality, by default *REFMAC* appears to increasingly up-weight ideal geometry restraints at higher doses. Reported RMSD bond length and angle values following restrained refinement are also much higher than reported for dataset 1 in Table 2.3, and indeed significantly higher than the anticipated RMSD values over a survey of structures at 2.0 Å as reported by *POLYGON* (Urzhumtseva *et al.*, 2009) as part of the *PHENIX* crystallographic suite (Adams *et al.*, 2010). This indicated that restrained refinement at higher doses had introduced substantial error into the model. Consequently, no coordinate restrained refinement was performed for each of the higher dose datasets.

It is noted in the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map calculations reported in Section 2.4.1 and 2.5.4 that each map has been generated over the unit cell dimensions as specified by the higher dose dataset $n > 1$, which do differ depending on n, since the crystal unit cell dimensions generally increase as a function of dose (Ravelli and McSweeney, 2000). However, it will be demonstrated later in Section 2.5.9.2 (refer to Figure 2.32) that the analysis of radiation damage is insensitive to whether the map is generated over the cell dimensions for each later dataset or fixed as the initial dataset throughout. In the current protocol, to compensate for slight re-definition of the unit cell parameters with increasing dose, the atomic positions within the model needed to be

updated at each higher dose dataset to match the $F_{obs,n} - F_{obs,1}$ map dimensions. To achieve this, for the model from each later dataset, 10 cycles of rigid body refinement were performed in *REFMAC*, using the refined coordinate model derived from the initial dataset coupled with the merged structure factor amplitudes, $|\mathbf{F}_{obs,n}|$, from the later dataset. This is a standard protocol for refining multi-dataset protein damage series (Southworth-Davies *et al.*, 2007). It is noted that if no rigid-body refinement was performed, the initial dose coordinate model would not align exactly with each $F_{obs,n} - F_{obs,1}$ difference map.

An increased thermal motion of atoms with deposited dose has widely been reported in previous studies (Weik *et al.*, 2000). In the current protocol, the rigid body refinement above was followed by 3 cycles of individual atom isotropic *B*-factor refinement in *phenix.refine* (Adams *et al.*, 2010). This allowed the isotropic *B*-factors to also be tracked with increasing dose as additional radiation damage indicator. In contrast to the first dataset, no TLS refinement was performed for each higher dose dataset, and as such only an isotropic description of atomic displacement assigned for each higher dose dataset. The *PEAKS* pipeline presented in Section 2.4 did not depend on *B*-factor values, however it is noted that the per-atom *B*-factor values will play a direct role in the per-atom radiation damage metrics derived by the *RIDL* pipeline in Section 2.5. In the current analysis, it has been argued that allowing the *B*-factors to increase with dose accounts for the increasing uncertainty in the location of each atom with increasing dose. However, in Section 2.5.9.3 it will be demonstrated that the decision to include isotropic or TLS *B*-factor refinement at each dose did not significantly affect the results of *RIDL* in the current analysis.

The above refinement strategy has been considered the default strategy for *Dp*Cel7B protein damage series. It must be noted, however, that in subsequent sections of this chapter (particularly Section 2.5.9 in which the *RIDL* pipeline is tested in detail), variations of the refinement protocol above will be explicitly stated where used.

Data reduction and refinement statistics are also reported in Tables 2.2 for each higher dose dataset. Coordinates and structure factor amplitudes for the full damage series have been deposited in the PDB with accession codes: *5mcc*, *5mcd*, *5mce*, *5mcf*, *5mch*, *5mci*, *5mcj*, *5mck*, *5mcl*, *5mcm* and *5mcn*.

As can be seen in Table 2.2, the diffraction data quality deteriorated with increasing dose, with

most notably an increase in R_{merge} and decreases in $CC_{1/2}$ and I/sig(I) with dose. In additional, the quality of the refinement statistics also diminished with dose with an increase in both R_{work} and R_{free}. Interestingly, the mean isotropic *B*-factor per dataset did not monotonically increase, but appeared to oscillate with increasing dose between odd and even dataset numbers. This is suggested to be a consequence of the continuous rotation of the crystal throughout data collection: since each dataset comprised of 180° of images, an identical crystal volume has been irradiated for across all even (or all odd) datasets.

2.3.4 Dose calculations

The diffraction weighted dose (DWD) (Zeldin *et al.*, 2013a) was calculated using *RADDOSE-3D* (Zeldin *et al.*, 2013b) for each dataset in the damage series. The crystal was modelled as a rectangular cuboid with dimensions $75 \times 75 \times 10\,\mu m^3$. Information on the crystal composition (the unit cell dimensions, number of amino acids per monomer, number of monomers per unit cell, and crystal solvent content) was determined from the *4xnn* coordinate model, which has an identical amino acid sequence to the current protein. The contribution of heavy elements within the crystallisation buffer (0.1 M Na and 0.9 M S) was included to calculate the crystal absorption and attenuation coefficients. Lighter elements (e.g. N, O, C and H) in the buffer were not explicitly included since these have a negligible relative contribution to the overall crystal absorption coefficient as the atomic absorption coefficients increase in approximate proportion to the fourth power of the atomic number (Murray *et al.*, 2005).

A Gaussian-shaped beam was modelled with parameters as provided in Section 2.3.2. As a result of the similarity in data collection strategy across datasets, for each dataset n = 1,2,... the resulting DWD value was calculated to be $1.11 + 2.16 \times (n-1)$ MGy. By the end of the final dataset ($n = 11$) the crystal had absorbed an accumulated DWD of 22.7 MGy.

2.4 Facilitated manual inspection of difference maps

Manual inspection of electron density remains an inevitable component of specific damage detection, however the process is severely limited by the subjectivity of the inspector. The decision of a

Table 2.2: Data processing and refinement statistics for the *Dp*Cel7B protein damage series. Values in parentheses are for the highest-resolution shells (fixed at 2.05 – 2.00 Å throughout). R_{free} has been calculated using a small (5 %) test set of randomly selected reflections, with the same test set maintained between datasets. Unit cell dimensions of $\alpha = \gamma = 90^{\circ}$ are conserved throughout. The mean isotropic *B*-factor has been reported for all refined atoms in each coordinate model. PDB accession codes are provided for the deposited series. A full assessment of model geometry has been provided in Table 2.3.

Dataset	1	2	3	4	5	6	7	8	9	10	11
PDB codes	5mcc	5mcd	5mce	5mcf	5mch	5mci	5mcj	5mck	5mcl	5mcm	5mcn
Dose (MGy)	1.11	3.27	5.43	7.59	9.75	11.9	14.1	16.2	18.4	20.6	22.7
Resolution range (Å)	164.1	82.1	82.3	82.4	82.6	82.6	82.8	82.6	83.0	82.7	83.1
	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00
Cell dimensions											
a (Å)	127.0	127.0	127.2	127.2	127.4	127.3	127.5	127.3	127.6	127.5	127.8
b (Å)	46.6	46.6	46.7	46.7	46.7	46.7	46.7	46.7	46.8	46.8	46.8
c (Å)	172.8	172.9	173.4	173.5	173.9	173.8	174.4	173.9	174.7	174.1	174.9
β (°)	108.3	108.3	108.3	108.2	108.3	108.2	108.2	108.2	108.2	108.2	108.2
No. Observations	214907	215330	215898	215805	215589	215268	215641	215058	216104	215297	216335
	(15647)	(15653)	(15609)	(15623)	(14888)	(14873)	(14775)	(14794)	(14796)	(14768)	(14487)
Unique reflections	65500	65572	65895	65997	66183	66172	66457	66250	66726	66424	66960
	(4626)	(4632)	(4619)	(4662)	(4433)	(4451)	(4423)	(4440)	(4444)	(4450)	(4455)
R_{merge}	0.06	0.06	0.06	0.07	0.08	0.10	0.11	0.16	0.20	0.25	0.41
$CC_{1/2}$ (%)	99.7	99.7	99.5	99.6	99.4	99.5	99.3	98.6	98.1	96.6	90.3
	(95.4)	(94.4)	(89.4)	(83.1)	(64.5)	(52.8)	(28.9)	(23.1)	(10.3)	(6.8)	(2.7)
I/sig(I)	12.2	12.2	10.9	10.0	8.3	7.7	6.4	6.4	5.2	5.6	4.3
	(4.3)	(4.1)	(3.1)	(2.8)	(1.6)	(1.7)	(1.0)	(1.4)	(0.8)	(1.2)	(0.8)
Multiplicity	3.3	3.3	3.3	3.3	3.3	3.3	3.2	3.2	3.2	3.2	3.2
	(3.4)	(3.4)	(3.4)	(3.4)	(3.4)	(3.3)	(3.3)	(3.3)	(3.3)	(3.3)	(3.3)
Wilson *B*-factor (Å^2)	23.8	24.9	27.1	25.9	32.8	27.4	34.7	24.9	30.0	24.7	24.8
Refinement											
R_{work} (%)	15.1	18.0	18.7	19.3	20.6	20.8	22.4	21.9	24.4	23.2	26.2
R_{free} (%)	17.3	20.2	20.8	21.3	22.5	22.9	24.4	24.1	26.5	25.22	28.4
Mean isotropic *B*-factor (Å^2)	30.0	27.4	30.0	30.9	37.8	32.9	40.2	31.1	35.3	30.2	30.6

Table 2.3: Assessment of final macromolecular geometry for the refined coordinate model corresponding to the first dataset, as reported by *Molprobity4* (Chen *et al.*, 2010). It is noted that one group of local NCS restraints had been used between chain A and chain B in the asymmetric unit. Since for the coordinate models corresponding to the higher dose datasets, only rigid-body refinement was performed on the refined initial dataset coordinate model coupled with the higher dataset merged structure factor amplitudes, the reported geometry statistics were conserved for the higher dose datasets.

Number of:	
Non-H atoms	7298
Protein atoms	6717
Water molecules	575
Other atoms	6
Bond parameters:	
RMSD bond length (Å)	0.007
RMSD bond angle (°)	1.258
Ramachandran analysis:	
Favoured (%)	98.6
Outliers (%)	0.0
Allowed (%)	1.4
Rotamer outliers (%)	0.8
All atom clash-score	0.69

correct map threshold at which to distinguish between damage signal and map noise is non-trivial, and moreover no standardised criteria exist to access whether density can be attributed to a true site-specific damage event. Furthermore, inspection is a laborious, time-consuming process, and the ability of the investigator to maintain a consistent assessment of the map is questionable.

The program *PEAKS* was developed in *Python 2.7* to permit higher throughput manual inspection of difference maps, and is purposely aimed to address the above issues. It depends on the openly available CCP4 crystallographic suite (Winn *et al.*, 2011) for specific tasks, and exists as a series of custom scripts, that must be run sequentially by the user from a command line environment. The input to *PEAKS* consists of a series of increasing dose `.mtz` merged structure factor files (output from *CTRUNCATE* following intensity data reduction) as well as a refined `.pdb` coordinate model (corresponding to the lowest dose dataset). The architecture of the *PEAKS* pipeline is presented in Figure 2.4. The functionality of each step will be discussed below and demonstrated for the *Dp*Cel7B protein case study.

2.4.1 Generation of Fourier difference maps

$F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps between the each high dose dataset n and the initial dataset are generated using a standard protocol (Southworth-Davies *et al.*, 2007). For each higher dose dataset n, *CAD* is used to combine the necessary `.mtz` columns for dataset n ($|\mathbf{F}_{\text{obs,n}}|$ and $\sigma\mathbf{F}_{\text{obs,n}}$, a measure of the error in the recorded $|\mathbf{F}_{\text{obs,n}}|$ value) with those for the initial dataset. The $\sigma\mathbf{F}_{\text{obs}}$ terms are required by the program *FFT* used below to identify any missing reflections and to set the corresponding $F_{\text{obs,n}} - F_{\text{obs,1}}$ map coefficients to zero. No $\sigma\mathbf{F}_{\text{obs}}$-defined cutoffs are applied to reject any measured reflections from the following calculation of map coefficients.

By default, using *SCALEIT*, an anisotropic *B*-factor scaling function is then applied to the higher dose dataset's structure factor amplitudes $|\mathbf{F}_{\text{obs,n}}|$ to scale them to $|\mathbf{F}_{\text{obs,1}}|$. Consequently all higher dose datasets are scaled to the same initial dataset. The anisotropic scaling model implemented in *SCALEIT* was chosen for this, however the effect of different scaling models on resulting difference maps has been investigated in detail in Section 2.5.4.

Following scaling, the Fourier difference map is generated over the asymmetric unit using *FFT* (Ten Eyck, 1973) in the `.map` file format, using default grid sampling axes and coarseness.

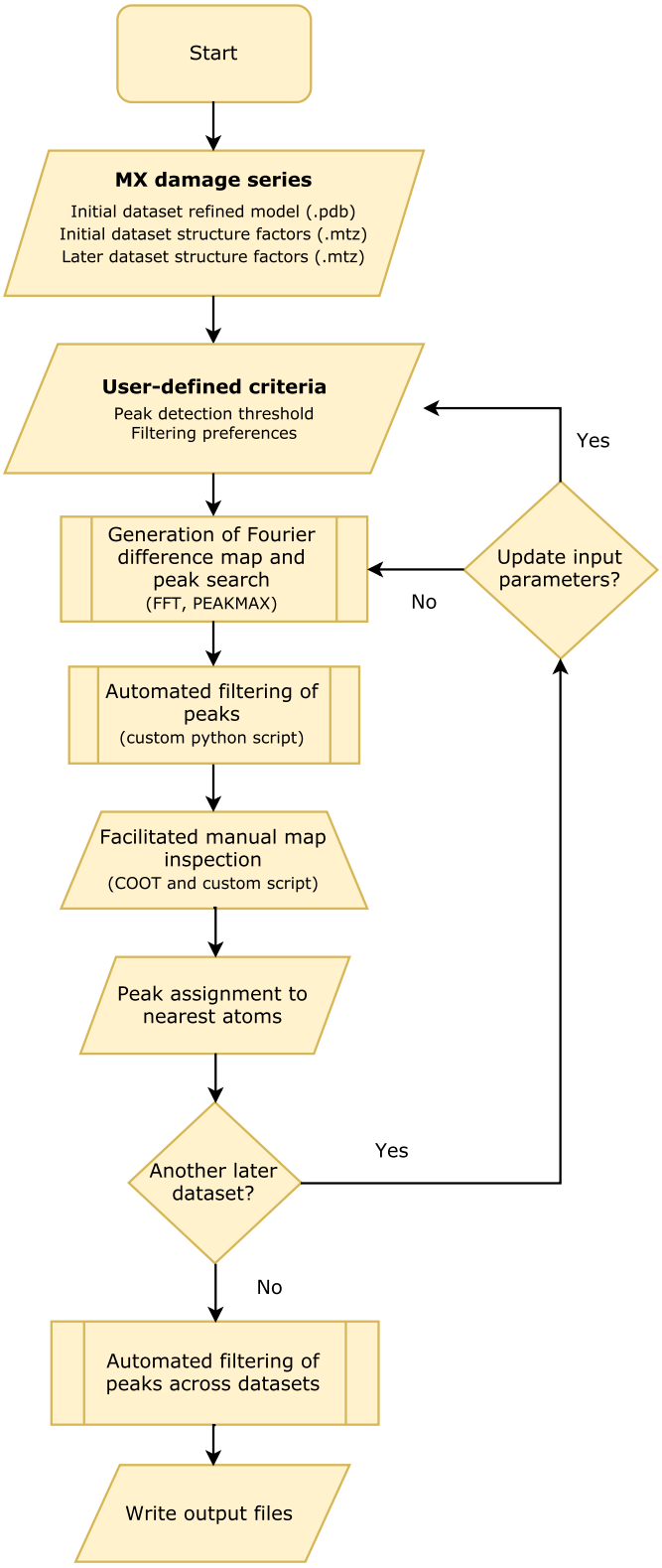


Figure 2.4: Flow chart illustrating the structure of the *PEAKS* pipeline to assist manual inspection of $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps to detect site-specific damage to atoms in a macromolecular coordinate model.

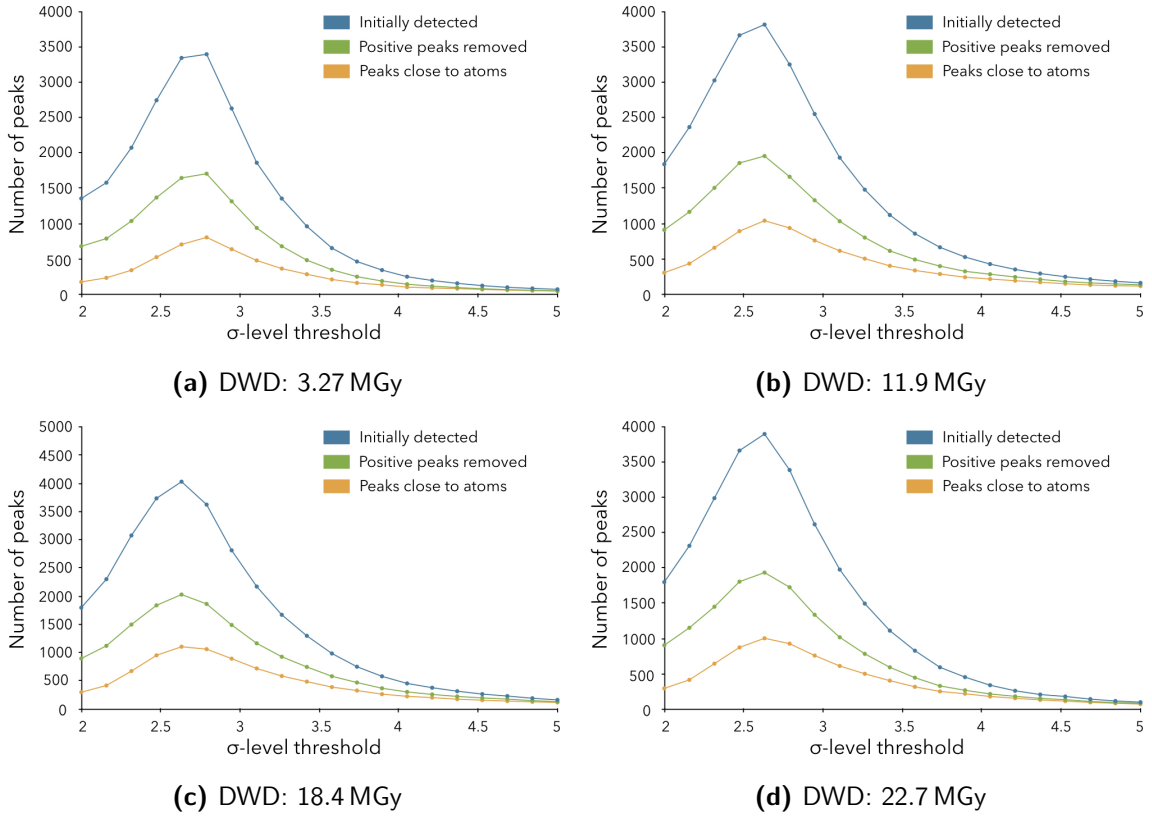
The calculated phases, $\phi_{\text{calc},1}$, corresponding to the refined structure for the initial dataset are used. Finally, *MAPMASK* is used to extend the map to ensure that it overlaps with the atomic coordinates as specified in the supplied coordinate model. The map extension in *MAPMASK* included a border of 3 Å around the edge of the model to ensure that the Van der Waals radii of all atoms typically found in proteins and nucleic acids were covered (see Table 2.4).

2.4.2 Difference peak detection

Difference map local maxima and minima (peaks) over the asymmetric unit are located using the CCP4 program *PEAKMAX*. A user-defined cut-off threshold is required by *PEAKMAX* (in units of σ), above which both positive and negative peaks will be detected. Figure 2.5 illustrates the variation in detected peaks for different σ thresholds for the *Dp*Cel7B test system. Unfortunately, the determination of a valid threshold is non-trivial, since a high threshold may neglect low-lying damage events, whereas a low threshold will inevitably lead to the inclusion of many unwanted noise peaks. To mitigate this issue, the default threshold was set as low as possible for the current step, and criteria detailed in Sections 2.4.3 and 2.4.4 were implemented to then reject noise peaks.

Unfortunately there is an issue with setting the σ threshold very low, since in the current work it has been observed that *PEAKMAX* becomes unstable and is unable to detect peaks when a threshold below $\sim 2.6\sigma$ is set (Figure 2.5). It is suggested that at low σ thresholds, *PEAKMAX* can no longer distinguish local map maxima/minima since an increasing number of neighbouring map voxels will also exhibit values above the low σ threshold. Whereas it could be argued that by increasing the coarseness of the map, this effect should be suppressed since less voxels are sampled in the vicinity of a local maximum/minimum, this may also lead to the failure to detect some maximum/minimum. Consequently this instability is non-trivial to address.

To avoid this instability at low σ, a default threshold of $\pm 3\sigma$ has been selected for the current analysis, and it will be validated in Section 2.4.5 that this threshold is sufficient to locate all damage sites in the *Dp*Cel7B protein test system. By the same reasoning, it was chosen to fix an upper detection limit of 8000 peaks as a default (this is the maximum permitted in *PEAKMAX*), to minimise the omission of low-lying peaks representing damage. These *PEAKMAX* parameters can be modified by the user if necessary for their specific investigation.



(a) DWD: 3.27 MGy **(b)** DWD: 11.9 MGy

(c) DWD: 18.4 MGy **(d)** DWD: 22.7 MGy

Figure 2.5: For the test case of the *Dp*Cel7B GH7 family protein at various doses, the lower density cut-off (in units of σ) for peak detection in *PEAKMAX* has been varied (blue). Following automated peak filtering (see Section 2.4.3), the number of included peaks following removal of positive peaks (green) and subsequent restriction to peaks proximal to the protein (orange) are also shown.

The output to *PEAKMAX* is a coordinate file `peaks.pdb` containing the x, y and z coordinates of peaks and their corresponding height (in units of e/Å^3) specified in the occupancy column.

2.4.3 Filtering of peaks

Several automatic filtering steps are now performed to reject unsuitable peaks. In a protein crystal, the protein itself will only occupy a proportion of the asymmetric unit, with the remainder being occupied by solvent (typically protein crystals have solvent fractions of between $27 - 65\,\%$ (Matthews, 1968)). Consequently, peaks detected by *PEAKMAX* may not be associated with the refined protein component.

The CCP4 program *WATPEAK* was initially tested in order to remove any peaks further than 5 Å from the protein atoms[‡], however this was observed to result in the unexpected loss of significant

[‡]As discussed in Section 2.4.1, the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map extension performed by the program *MAPMASK* included a border of 3 Å around the edge of the model. As such, for atoms on the surface of the model, the program *WATPEAK*

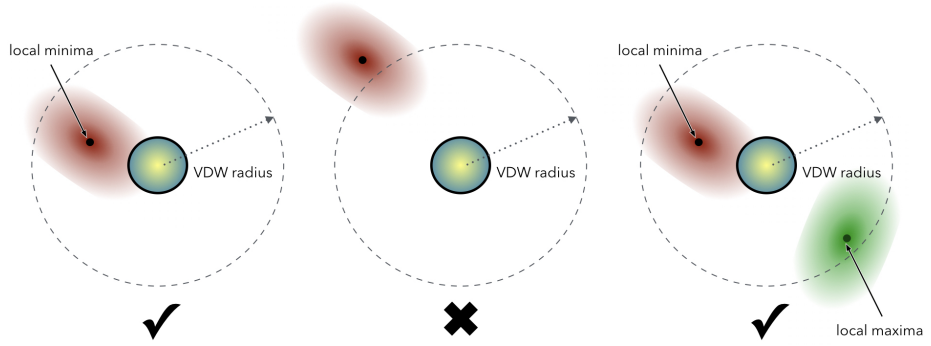


Figure 2.6: Difference map peaks are accepted only when the coordinates of the local maxima lie within the Van der Waals radius of an atom (sphere) (left), otherwise they are rejected (middle). A potential radiation-induced conformational change can be detected on the same principle if both a negative and positive peak lie within the VDW search radius of an atom (right). A current limitation of *PEAKS* is that the conformational change will only be detected if the positive peak lies inside the search radius, however this may not be true.

Table 2.4: Examples of Van der Waals radii (in Å) for atoms typically found in proteins and nucleic acids. Values have been retrieved from Housecroft and Sharpe (2008).

Element	H	C	N	O	P	S
Radius (Å)	1.2	1.7	1.55	1.52	1.8	1.8

peaks detected by *PEAKMAX* (for example at the Cys-162 sulphur in *Dp*Cel7B). Furthermore, *WATPEAK* is sub-optimal since it only permits a single cut off distance for all atoms in a structure. Consequently, a custom script was written to remove difference peaks that exceed the Van der Waals (VDW) radius of each individual atom (Figure 2.6 and Table 2.4), and thus allow a relevant peak *search radius* to be assigned for each atom. The suitability of the VDW radius threshold will be discussed in Section 2.4.5. Alternatively, the user can specify a singular distance threshold threshold for all atoms. An option has also been provided to allow the user to restrict the selection to include only peaks adjacent to a subset of residue types, or to only side-chain atoms.

In addition, in order to suit the specific objective of the user, options have been provided to restrict the search to (a) negative peaks (sites of electron density loss only), or (b) positive peaks (sites electron density gain only). It is also possible to detect potential radiation-induced conformational changes. This is achieved by locating peaks for which another peak of the opposite sign (negative/positive) has also been detected within the VDW search radius of an atom (Figure 2.6). It is

was ineffective at removing any peaks (within 3 Å of the model) located in the solvent region surrounding the model.

Table 2.5: The number of difference map peaks included by the *PEAKS* program at each step of filtering and manual inspection. The lower peak detection threshold has been set at the $\pm 3\sigma$ default. This example is for the *Dp*Cel7B protein damage series, with the criteria that only negative peaks are of interest, and that peaks closest to water molecules should be ignored. The results of manual inspection (facilitated using *PEAKS*, see Section 2.4.4), following the automatic filtering of peaks, has also been reported.

Dataset	2	3	4	5	6	7	8	9	10	11
Detection										
PEAKMAX output	6069	5363	5070	5856	5159	5812	5125	5704	5744	5221
Filtering										
Proximal to atoms of interest	1158	1239	1293	1625	1319	1612	1280	1461	1452	1251
Negative peaks only	612	662	692	903	755	970	739	874	809	724
Peaks retained (%)	10.1	12.3	13.6	15.4	14.6	16.7	14.4	15.3	14.1	13.9
Manual inspection										
After inspection	114	170	206	258	261	350	257	315	228	254
Peaks retained (%)	1.88	3.17	4.06	4.41	5.06	6.02	5.01	5.52	3.97	4.86

noted that with this implementation, *PEAKS* currently cannot distinguish larger conformational changes, in which the positive peak lies outside of the VDW search radius of an atom. However, optionally the search radius around the atom for the positive peak can be made larger to account for more significant conformational shifts.

For the *Dp*Cel7B protein test system, Table 2.5 and Figure 2.7 indicate the number of included peaks at each stage of the *PEAKS* pipeline. Compared to those initially detected by *PEAKMAX*, only a small percent ($< 17\,\%$) of peaks must be manually inspected at all doses. This reduction would be even more beneficial if the above filtering restricted to only a subset of residue types (data not shown).

2.4.4 Facilitated manual inspection

Manual inspection of the difference map is performed using *COOT* (Emsley *et al.*, 2010) to render a 3D representation of the refined initial dataset coordinate model, with the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map overlaid. *PEAKS* automatically launches *COOT* at runtime, ready for manual inspection with the first peak centred in view. It is noted that in *COOT*, the built-in tool titled '*difference map*

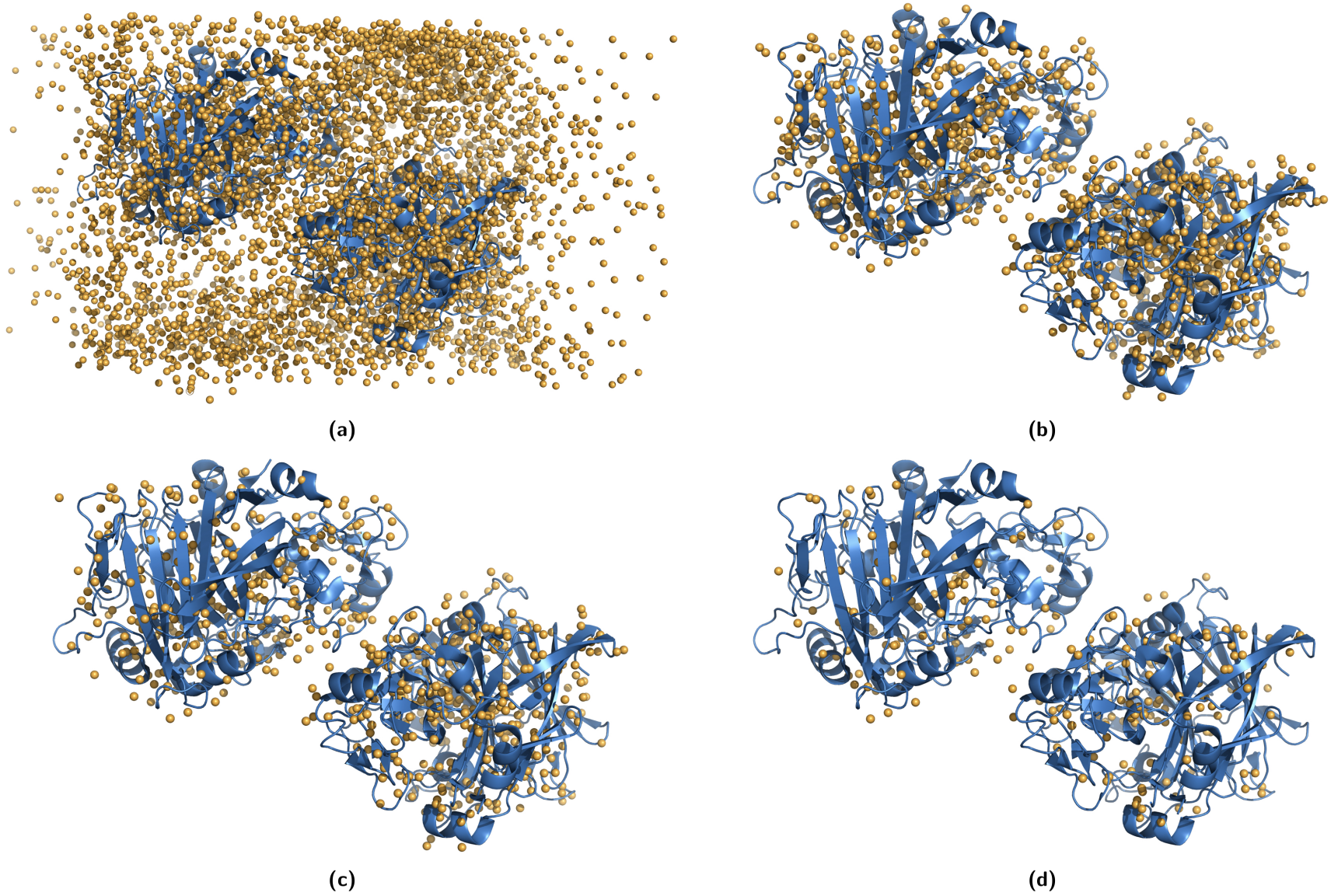


Figure 2.7: For the *Dp*Cel7B case study at 22.7 MGy, the distribution of difference map peaks included by the *PEAKS* program is shown after (a) initial detection using *PEAKMAX*, (b) restriction to peaks proximal to the non-water components of the asymmetric unit, (c) restriction to negative density loss peaks, and (d) manual inspection using *COOT*.

peaks' is available to isolate all difference map peaks in the close vicinity of the macromolecule. However, this tool is a suboptimal alternative to the current protocol, since it neglects the desired filtering criteria as defined in the above section, and furthermore it only reports difference peak heights in units of σ.

Currently, *PEAKS* requires the user to manually inspect each difference peak in turn in *COOT*, advancing with simply the keyboard *spacebar* command. In a separately generated graphical user interface (implemented in *PEAKS* using the Python *Tkinter* package), the user can efficiently assign whether a site corresponds to damage or not. Peaks are processed in order of decreasing peak height. For each selected peak, the peak height (in units of $e/Å^3$) is automatically assigned to the closest atom in the coordinate model, and this information is written to an output summary file for that dataset.

The output from *PEAKS* is subsequently reported in units of $e/Å^3$ as opposed to σ. As discussed previously in Section 2.1 (and demonstrated visually in Figure 2.2), reporting $F_{obs,n} - F_{obs,1}$ map peak heights in units of σ is sub-optimal, since σ is dataset-specific. Due to global radiation damage, the standard deviation, σ, of values in the $F_{obs,n} - F_{obs,1}$ map increases with dose, and the significance of map peaks (in units of σ) can diminish with dose. By instead reporting peak heights in raw units of $e/Å^3$, this issue is avoided and the output provides a clearer indication of the cumulative electron density loss at sites of specific radiation damage with dose.

Since a low peak detection threshold is used by default (see Section 2.4.2), as peak inspection proceeds, low-lying damage will eventually become indistinguishable from background map noise. If 20 successive peaks are rejected by the user, *PEAKS* will then request whether the user wishes to continue inspecting any remaining peaks. This criterion has been implemented to allow manual inspection to be halted once the user has inspected all true damage sites. It is further noted that future improvements to this implementation should include the full integration of this decision making process into a single plug-in window in *COOT*, to automatically advance to the next damage site upon labelling a site. The development of this was deemed beyond the scope of the initial implementation of this tool.

2.4.4.1 Manual inspection for the *Dp*Cel7B test case

For the *Dp*Cel7B protein test case, $F_{\text{obs,n}} - F_{\text{obs,1}}$ peaks were first filtered such that (a) only negative (density loss) peaks and (b) only peaks proximal to non-water atoms would be considered. Following application of these two conditions, Figure 2.8 illustrates the result of the manual inspection process. The following criteria were used to manually inspect peaks:

1. Peaks centred at radiation-sensitive atoms that can undergo cleavage events (disulphides, carboxylates) should be included.

2. Peaks proximal to other atom types should be included if it appears feasible that the residue may have undergone a conformational change. This assessment was based on where the peak is located (side or main chain) and whether side-chain movement would be sterically hindered. The presence of nearby positive peaks could help to support this decision, but were not required.

3. The presence of multiple peaks at a side-chain moiety supported the decision that a peak should be retained (for example, if two peaks overlap both the $O_{\epsilon 1}$ and $N_{\epsilon 2}$ atoms of Gln as opposed to just the $O_{\epsilon 1}$ atom).

4. Peaks that are also in close proximity to noisy peaks in the solvent region should be omitted.

Future development of *PEAKS* would aim to implement these criteria automatically as part of Section 2.4.3. However, whereas criterion 1 would be simple to implement, the remainder of the criteria are greatly aided by human insight and appear non-trivial to implement.

Figure 2.9a illustrates that as manual inspection progresses, the heights of manually inspected peaks (negative only) rapidly plateaued at the $\pm 3\sigma$ background level of the map. Difference peaks attained higher (more negative) peak heights as dose increased, indicating that the magnitude of damage at the detected damaged atoms increased with dose, as anticipated.

Table 2.5 and Figure 2.9b demonstrate that a higher dose does not necessarily result in the detection of a higher number of damage sites following manual inspection. It is suggested that either the number of available damage sites has saturated with increasing dose, or that the

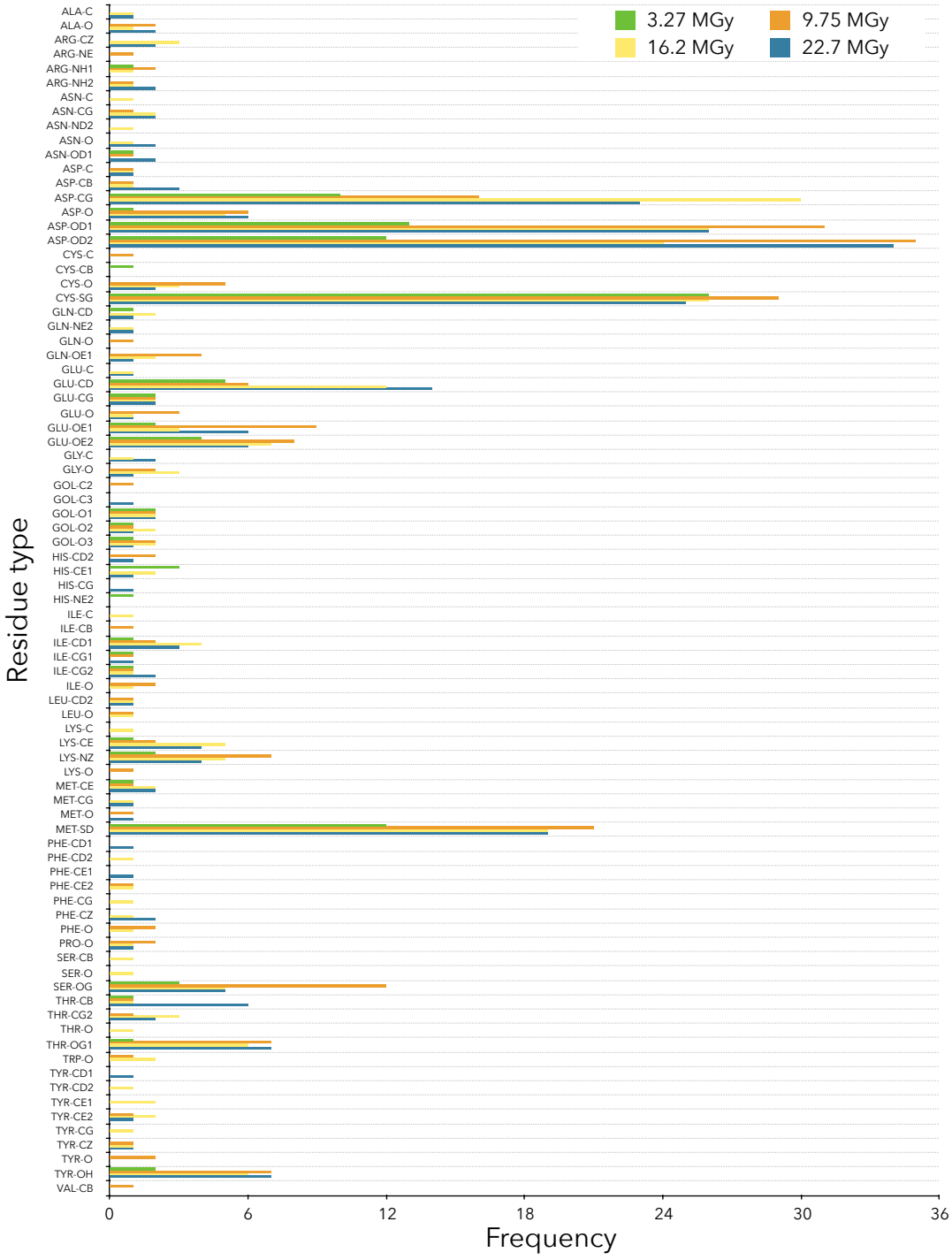
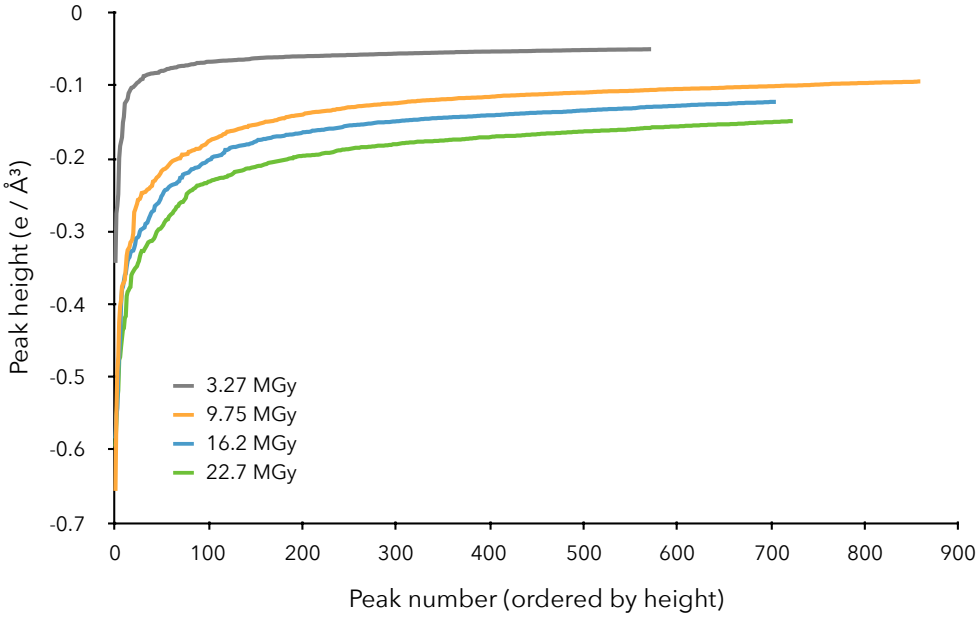
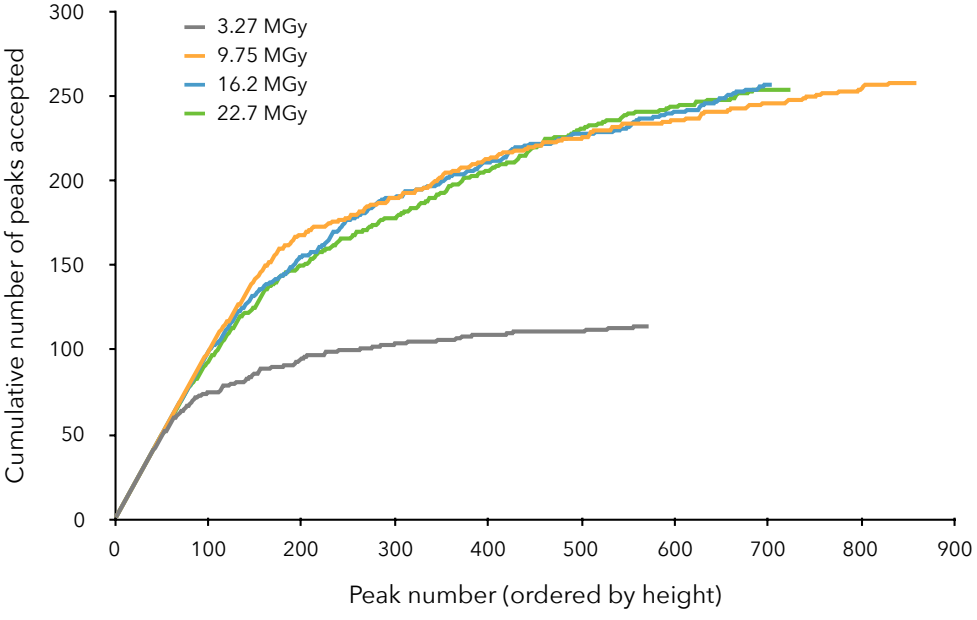


Figure 2.8: For the *Dp*Cel7B case study, the number of damage sites reported following manual $F_{\text{obs,n}} - F_{\text{obs,1}}$ map inspection is shown for the cases in which the higher dose dataset n corresponds to 3.27, 9.75, 16,2 and 22.7 MGy. Other doses show similar behaviour and are not shown. Atom types that are not assigned any peaks are omitted from the y-axis. As anticipated, a high frequency of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks are located proximal to the radiation-sensitive groups in Cys, Glu, Asp and Met. Other residues (Thr, Tyr, Ser and Lys) exhibited pronounced peaks at their side-chain atoms and were suspected to have undergone a conformational shift or become disordered. In addition to the protein, atoms of ordered glycerol molecules (labelled as GOL) have been included.



(a)



(b)

Figure 2.9: (a) The peak height (in e/Å^3) of each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peak that underwent manual inspection using *PEAKS*. (b) The cumulative number of individual atoms that were detected to be in close proximity to peaks that have been labelled as damage during manual $F_{\text{obs,n}} - F_{\text{obs,1}}$ map inspection. This is for the case of the *Dp*Cel7B GH7 family protein, using $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps with the higher dose dataset n corresponding to 3.27, 9.75, 16,2 and 22.7 MGy. Only negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks have been considered throughout. The initial peak detection threshold in *PEAKS* was $\pm 3\sigma$ for all four difference maps.

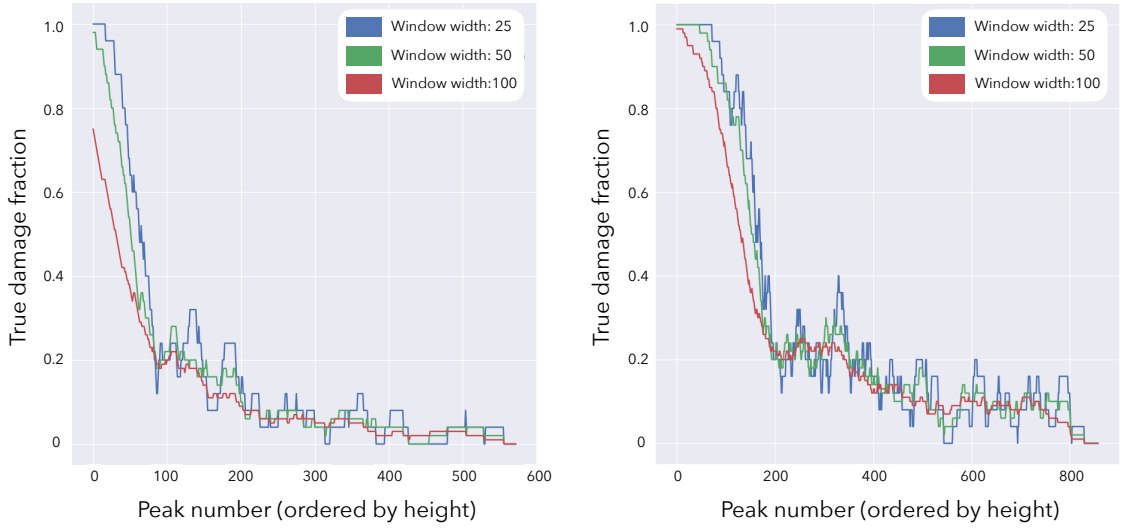
increased map noise at the higher doses investigated (due to deteriorating data quality at high doses), could mask the location of further damage sites in the $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps. In support of this, the $CC_{1/2}$ for the highest resolution bin (2.05 – 2.00 Å) was observed to reduce from 94.4 % to 2.7 % between the datasets at 3.27 MGy and 22.7 MGy (Table 2.2). It is also noted that the significant overall increase in peaks assigned to damage between 3.27 MGy and 9.75 MGy is due to the onset of many decarboxylation events between these dose values.

2.4.5 Assessment of manual inspection quality

Manual inspection should not be considered to be error free. Low-lying peaks (those at low map σ-heights) have been labelled as true damage (Figure 2.9b), however the veracity of this decision is uncertain due to the presence of comparable map noise. It is therefore essential to be able to retrospectively assess the quality of manual inspection.

A sliding window analysis (Figure 2.10) could help determine a suitable cut-off value beyond which map noise is dominant and the labelling of further sites should be deemed unreliable. At each dose for the *Dp*Cel7B case study, towards the end of manual inspection, less than 10 % of recently inspected peaks were classified as true damage. This observation appears to be independent of the sliding window size used to sample peaks. Consequently, the $\pm 3\sigma$ sigma detection threshold implemented in Section 2.4.2 appears sufficient, and a lower threshold would yield a negligible increase in assigned damage sites.

If a suitable manual inspection has been performed, it should also be anticipated that peaks positioned further from their nearest atoms have a greater probability of being rejected by the inspector. Figure 2.11 illustrates the distributions of distances for all manually inspected peaks in the *Dp*Cel7B case study. Since the range of distances for accepted peaks is visibly lower than for those that have been rejected, this indicates that the per-atom distance threshold implemented in Section 2.4.3 is suitable, and that a larger distance threshold would not lead to the inclusion of more peaks following manual inspection.



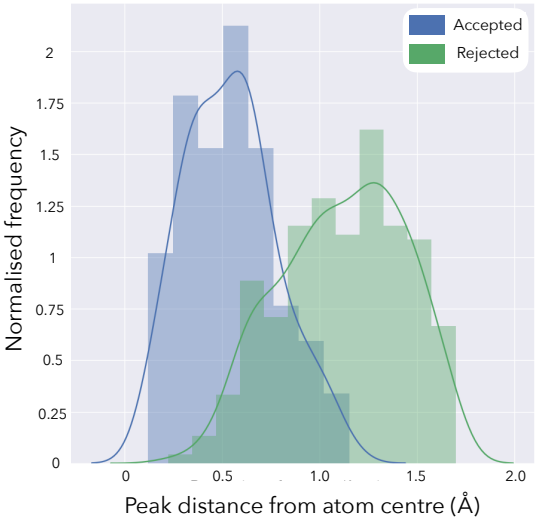
(a) DWD: 3.27 MGy

(b) DWD: 9.75 MGy

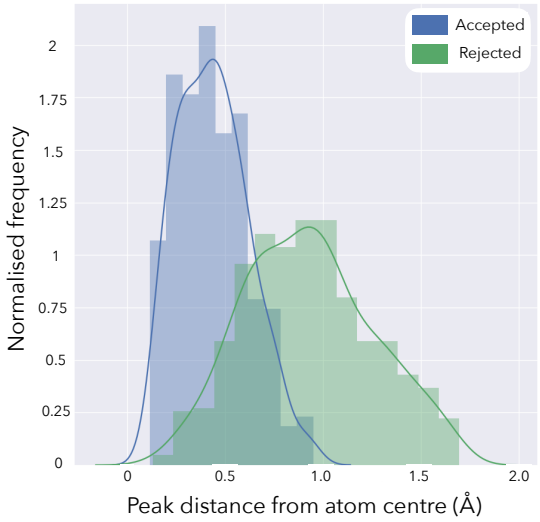
(c) DWD: 16.2 MGy

(d) DWD: 22.7 MGy

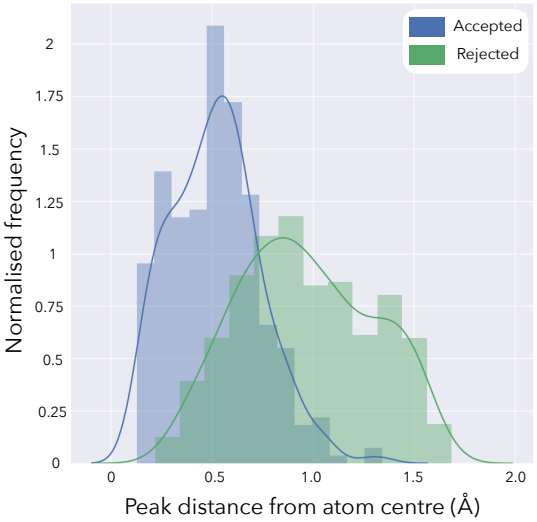
Figure 2.10: Sliding window analysis to determine when $F_{\text{obs,n}} - F_{\text{obs,1}}$ map noise becomes dominant during manual inspection of the *Dp*Cel7B test series. To perform a sliding window analysis of width w, peaks have been ordered by height (x-axis), and for each set of consecutive w ordered peaks, the fraction of peaks in the sliding window that are classified as damage in manual inspection is plotted. Graphically, for each individual instance of a sliding window, the resulting fraction has been plotted at x-axis value for the first peak number present in the current window. Three sliding window widths have been included of widths 25 (blue), 50 (green) and 100 (red) to demonstrate that the overall qualitative behaviour is insensitive to window size, however becomes smoother with increasing window size. For each plot, the DWD corresponding to the higher dose dataset n used to calculate the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map is reported. Other doses in the *Dp*Cel7B series exhibit similar distributions and are not shown.



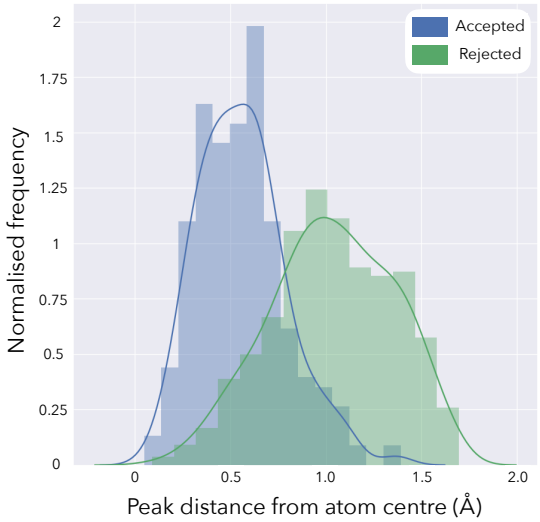
(a) DWD: 3.27 MGy



(b) DWD: 9.75 MGy



(c) DWD: 16.2 MGy



(d) DWD: 22.7 MGy

Figure 2.11: Distribution of distances from nearest atoms for peaks detected by *PEAKS* following assisted manual inspection for the *Dp*Cel7B protein test series. Peaks are classified as *accepted* if identified by the inspector as a true damage site, and *rejected* otherwise. For each plot, the DWD corresponding to the higher dose dataset n used to calculate the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map is reported. Other doses in the *Dp*Cel7B series exhibit similar distributions and are not shown. y-axis values have been normalised such that the overall summed area between each histogram and the x-axis is equal to 1. In addition, Gaussian kernel density estimations (KDE) have been overlaid over each histogram to provide a non-parametric estimate of the underlying distribution for each set of partitioned data.

2.4.6 Further filtering of peaks

Assuming that the previous steps have been performed for each higher dose dataset $n > 1$, the user can now opt to filter to include only *recurrent* damage sites: sites that persist once detected (if detected for dataset $m > 1$, they remain present for all $n > m$). Under the assumption that once a residue is damaged, it will remain damaged, this step is justifiable, and has been implemented to remove false noise peaks that the user may have included during inspection of a single map. However it is noted that if a user incorrectly omits a damage site during manual inspection, this could then lead to that damage site being omitted from all later datasets. Consequently this step is optional.

2.4.7 The output of *PEAKS*

For the *Dp*Cel7B case study, Figure 2.12 demonstrates the damage as a function of increasing dose for the top 10 recurrent damage sites in the protein. In this figure, the procedure in Section 2.4.6 has been applied to isolate only damage sites that persist once detected. As anticipated: (a) the majority of the top damage sites in *Dp*Cel7B correspond to the breaking of disulphide bridges, and (b) there is increasing electron density loss (measured by $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peak heights in units of e/Å^3) with increasing dose.

It was observed that despite the high conformational similarity between chain A and B (with a C_α backbone RMS deviation of 0.063 Å as calculated by the *align* command in *PyMol*), there were differences in $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peak heights as reported by *PEAKS* between NCS-related disulphide bonds at each dose. Figure 2.13 provides an illustration of how the $F_{\text{obs,2}} - F_{\text{obs,1}}$ map differed locally in the regions around NCS-related Cys residues. In particular, whereas A-Cys-162-S_γ was indicated to undergo a well-defined conformational change upon irradiation, no such conformational state was detectable for B-Cys-162-S_γ. As such, differences detected by *PEAKS* were concluded to be due to true differences between chain A and B, as opposed to due to map noise.

It is noted that in Figure 2.12, no attempt has been made to extrapolate data back to 0 MGy. Theoretically, since the data are derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps, the heights of all difference

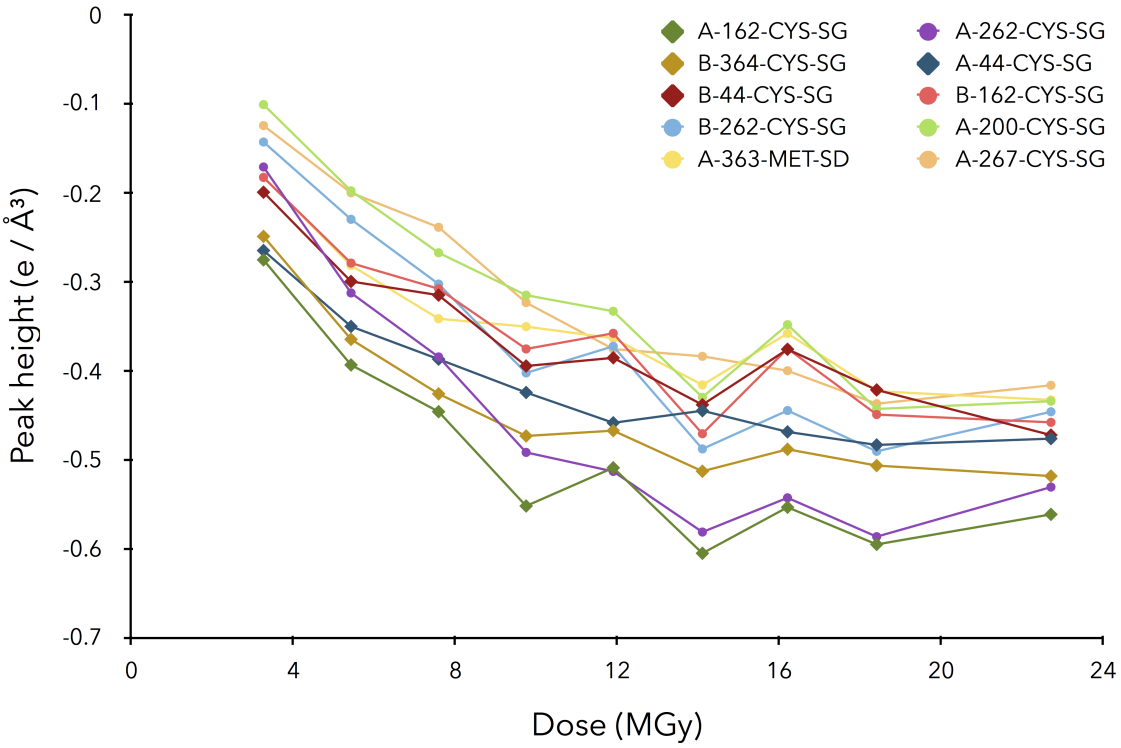


Figure 2.12: For the *Dp*Cel7B protein test series, the top 10 *recurrent* damage sites are plotted as a function of dose. A clear ordering of damage is visible, with A-Cys-162-S_γ and A-Cys-262-S_γ deemed as most radiation-sensitive.

map peaks must strictly equal $0\,\mathrm{e}/\mathrm{Å}^3$ when the x-axis DWD equals 1.11 MGy (corresponding to a $F_{\mathrm{obs},1} - F_{\mathrm{obs},1}$ map). However, without a current valid model to explain how the $F_{\mathrm{obs,n}} - F_{\mathrm{obs},1}$ map density ought to behaviour as a function of dose, any extrapolation is unjustified.

2.5 Automated detection of damage in difference maps

The program *RIDL* (Radiation Induced Density Loss) has been developed to automate the process of creating and interpreting $F_{\mathrm{obs,n}} - F_{\mathrm{obs},1}$ difference maps generated from an MX radiation damage series, and provide an objective tool for analysing site-specific radiation damage.

Instead of identifying difference map peaks in a map and then attributing these to specific atoms in a structure (as performed by *PEAKS*, see Section 2.4), the aim of *RIDL* is to sample the local region of the $F_{\mathrm{obs,n}} - F_{\mathrm{obs},1}$ map surrounding each refined atom in a macromolecular coordinate file, and define per-atom summary metrics to describe the radiation-induced electron density change

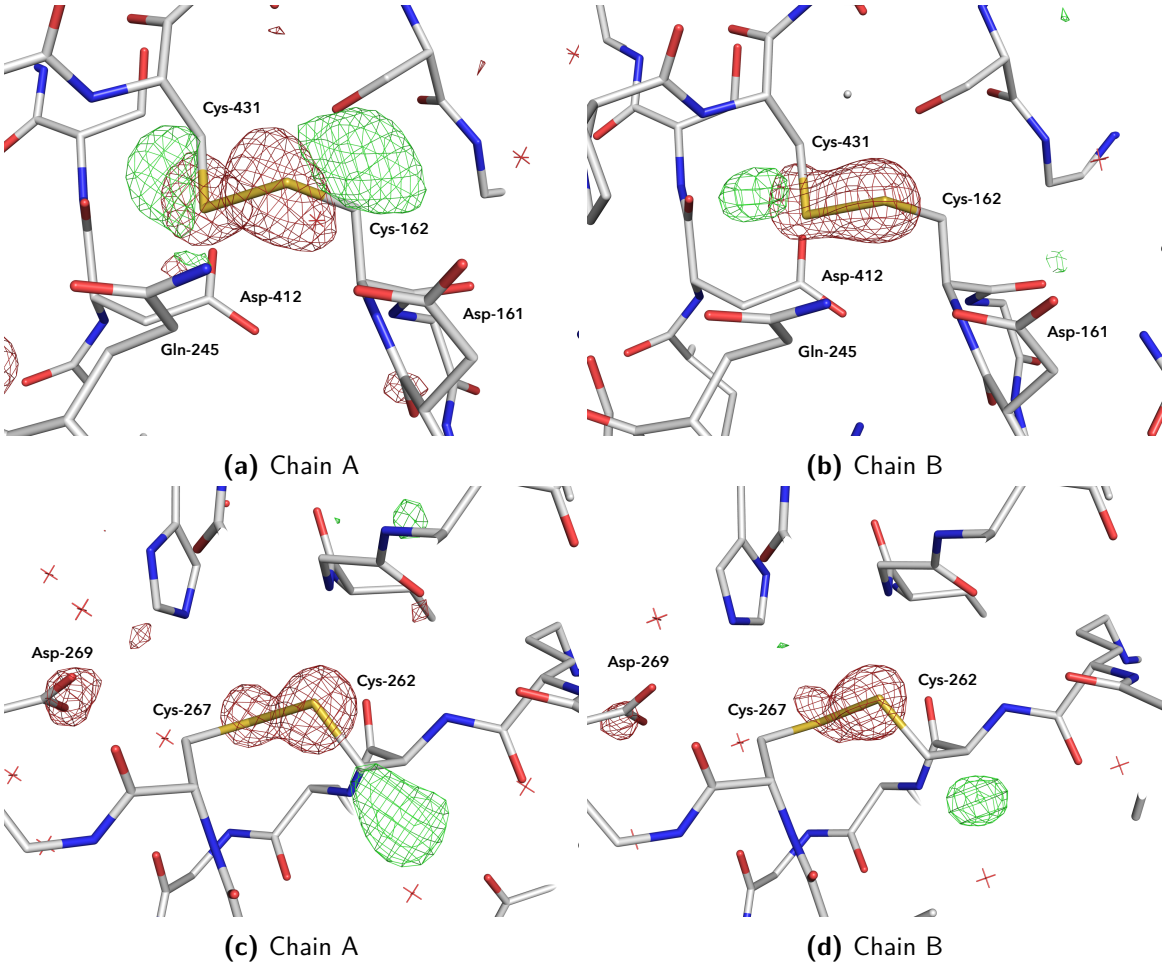


Figure 2.13: An illustration of $F_{\text{obs},2} - F_{\text{obs},1}$ difference map behaviour localised to example disulphide bonds in the NCS-related chains (chains A and B) of the *Dp*Cel7B protein case study. All maps are contoured at $\pm 4\sigma$ in green/red. Despite the NCS-related symmetry, differences in the radiation-induced changes to the disulphides were detected. Most notably, on comparison of Cys-162 in chain A and B, only for chain A did a well-defined reorientation of the sulphur atom appear with dose.

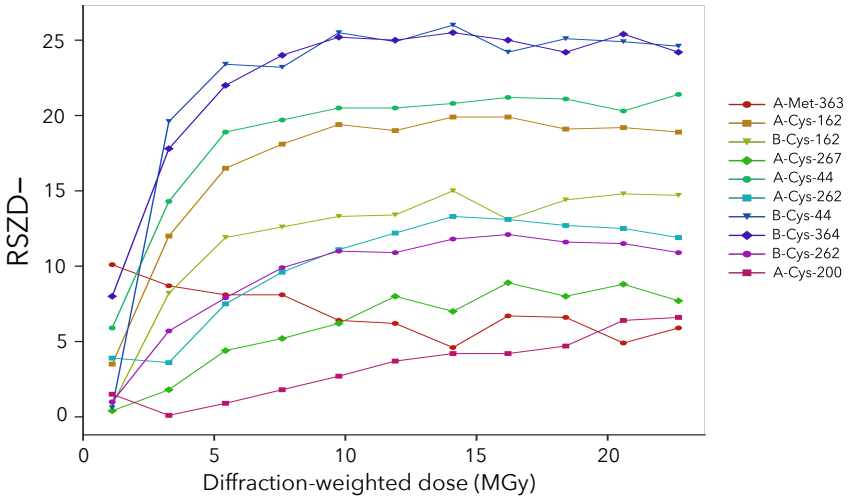
to individual atoms. Such an approach thus enables the relative measurement of radiation-induced density changes for every refined atom in a structure, as opposed to only atoms in close proximity to well-formed and user-interpreted Fourier difference map peaks above a predetermined σ-threshold. The subjective bias of the researcher is thus eliminated entirely.

2.5.1 The availability of other programs and motivation for *RIDL*

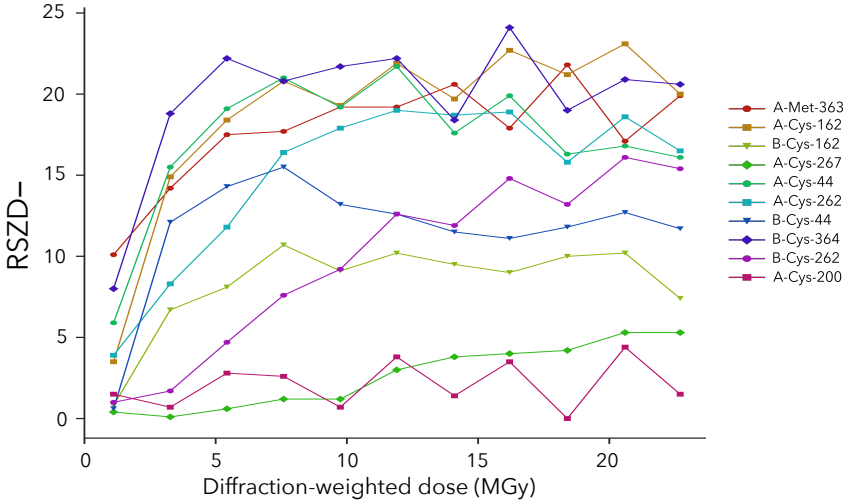
It is noted that other programs are available to assess the fit of individual residues to electron density (e.g. *EDSTATS* (Tickle, 2012) and *OVERLAPMAP* (Winn *et al.*, 2011)). Such programs typically report map correlation coefficients or Z-score-based measures to describe the agreement between the refined coordinate model and either $2mF_{obs} - DF_{calc}$ or $mF_{obs} - DF_{calc}$ maps (or using both). These programs have not been specifically designed for radiation damage analysis as a function of dose, but in principle could be reworked for this purpose.

However, it is unclear how the direct output from these programs could be used to isolate radiation damage effects. For example, the *real-space difference density Z-score* (RSZD) metric calculated by *EDSTATS* provides an assessment of the *accuracy* of the refined model, derived from the corresponding $mF_{obs} - DF_{calc}$ map. It is anticipated that the $mF_{obs} - DF_{calc}$ map will change with increasing dose, and indeed this has been reported in the earliest specific damage investigations at third generation synchrotrons (Weik *et al.*, 2000). Consequently, the RSZD should also vary with dose. Despite this, it is difficult to distinguish whether a poor RSZD score at high doses indicates radiation damage or merely incorrect model placement, due to worsening diffraction data quality as dose increases. This issue can be demonstrated for the *Dp*Cel7B protein test damage series, for which different refinement strategies at each dose led to inconsistent conclusions using *EDSTATS* (Figure 2.14 for the case of the RSZD$-$ metric, which is the RSZD metric calculated using only negative $mF_{obs} - DF_{calc}$ map points). Here, refinement of higher dose datasets has been employed as an efficient method to access updated $2mF_{obs} - DF_{calc}$ and $mF_{obs} - DF_{calc}$ map coefficients for each higher dose dataset.

Even if no refinement of higher dose datasets is performed, such that the first dataset coordinate model, and corresponding $\phi_{calc,1}$ phase and $|\mathbf{F}_{calc,1}|$ amplitude sets are instead used at all doses to calculate each $2mF_{obs} - DF_{calc}$ and $mF_{obs} - DF_{calc}$ map, such metrics based on model accuracy



(a)



(b)

Figure 2.14: For the test case of the *Dp*Cel7B protein damage series, the RSZD− metric has been calculated using *EDSTATS* for distinct refinement strategies. In (a), the higher dose coordinate models have been generated by performing 10 cycles of rigid body refinement in *REFMAC* (but no refinement of atomic *B*-factors), using the initial dataset refined coordinate model with the higher dose dataset structure factor amplitudes. In (b), the higher dose coordinates have been instead generated after 10 cycles of coordinate-restrained and isotropic *B*-factor refinement in *REFMAC*. The top 10 damaged residues, as detected by manual inspection of $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps using *PEAKS* (Figure 2.12) are included only, with the RSZD− metric calculated over all side chain atoms in each residue.
The change in the RSZD− metric for individual residues is inconsistent between the different refinement strategies, most evidently for A-Met-363, in which a small conformational shift of $\sim 0.3\,\text{Å}$ was observed between the Met-S_δ atoms following the rigid-body refinement and restrained refinement strategies. The results from *EDSTATS* do not agree with manual inspection of $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps as reported in Figure 2.12. Most notably, A-Cys-200 attains relatively low RSZD− scores in both (a) and (b) (RSZD− $\lesssim 5$), and in accordance with this, manual inspection of the corresponding mF_{obs} − DF_{calc} maps did not reveal any clear negative difference density at the site greater than -3σ. This is in contrast with the $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps generated in Section 2.4, where a clear negative difference map peak was detectable by manual inspection.

are highly sensitive to imperfections in the original model. The data in Figure 2.15 exemplify this for the case of the RSZD metric from *EDSTATS*. Overall, the change in *accuracy* of the model with dose is thus argued to be an unsuitable metric to track radiation damage.

The above conclusion also holds true for other output metrics from *EDSTATS* and *OVERLAPMAP*, which measure the agreement between the model and the electron density, e.g., the real-space R-factor (RSR) (Jones *et al.*, 1991):

$$RSR = \frac{\sum |\rho_{\text{obs}} - \rho_{\text{calc}}|}{\sum |\rho_{\text{obs}} + \rho_{\text{calc}}|} \tag{2.5.1}$$

where the above summation is over all density map grid points in a given radius centred on each atom, and the real-space correlation coefficient (RSCC):

$$RSCC = \frac{\text{cov}(\rho_{\text{obs}}, \rho_{\text{calc}})}{\sqrt{\text{var}(\rho_{\text{obs}}) \cdot \text{var}(\rho_{\text{calc}})}} \tag{2.5.2}$$

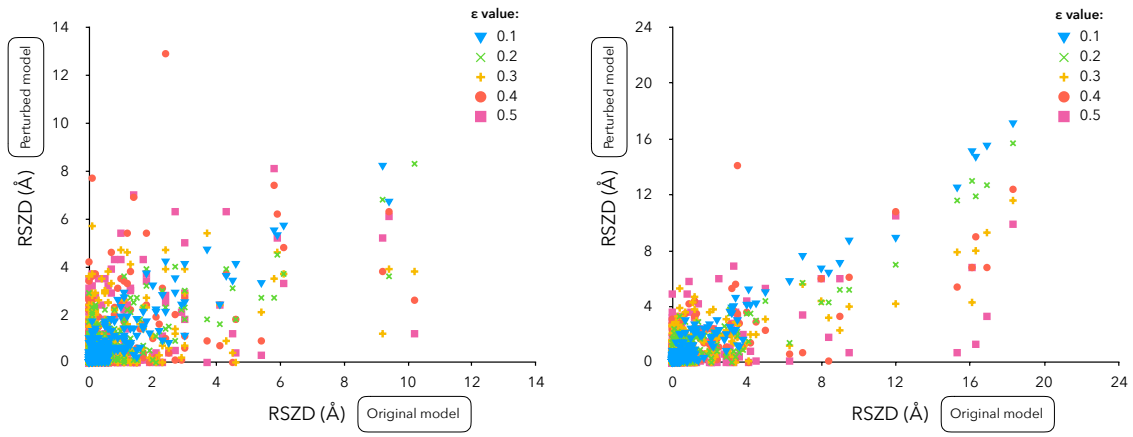
where $\text{cov}(x, y)$ and $\text{var}(x)$ are the covariance and variance, respectively, over the density map grid points surrounding each atom. Clearly errors in the model directly affect ρ_{calc}, and thus also affect both RSR and $RSCC$.

In contrast, a significant benefit of instead using $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps to investigate damage lies in the fact that these maps are less highly dependent on the local accuracy of the initial model around each atom[§]. To this end, the new program *RIDL* has been written to allow the calculation of new metrics that have been purposely designed to assess radiation damage, using the set of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map values around each atom.

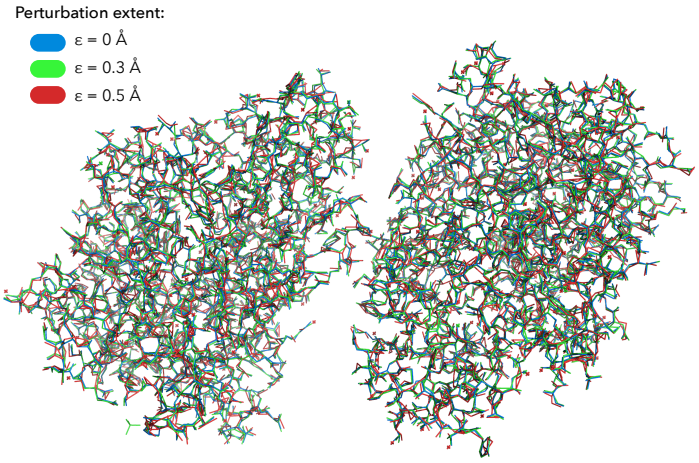
2.5.2 Construction of the *RIDL* pipeline

RIDL has been written in the *Python* object-oriented programming language, to allow cross-platform compatibility and so that it can be run directly from the command line together with a pre-defined `.txt` input file. At runtime, *RIDL* utilises a range of widely used CCP4 suite

[§]However, it is evident that model errors will affect the set of calculated phases that are used when calculating a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. Thus the dependence on the model cannot be entirely eliminated. It will be demonstrated later in Section 2.5.9.5 that metrics derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps are relatively robust to such phase error up to an overall mean error of approximately 40°.



(a) 3.27 MGy **(b)** 22.7 MGy



(c) 22.7 MGy

Figure 2.15: For the test case of the *Dp*Cel7B protein damage series, the sensitivity of the RSZD metric (calculated using *EDSTATS*) to inaccuracies in the refined coordinate model has been presented. In contrast to Figure 2.14, no subsequent model refinement has been performed at each later dose: (a) 3.27 MGy and (b) 22.7 MGy; instead, all $2mF_{obs} - DF_{calc}$ and $mF_{obs} - DF_{calc}$ maps at higher doses have been calculated using the $\phi_{calc,1}$ phase and $|\mathbf{F}_{calc,1}|$ amplitude sets corresponding to the initial dose coordinate model. As such, the coordinate model is constant with respect to dose. Each x-axis displays the RSZD metric values calculated over all refined residues in the original *Dp*Cel7B coordinate model.
Each y-axis displays the RSZD values computed for a series of *Dp*Cel7B models into which coordinate errors have been introduced. This was achieved by perturbing the Cartesian x, y and z positions for each individual atom by a uniformly distributed random amount $U[-\epsilon, \epsilon]$, between $-\epsilon$ and ϵ (in units of Å). In (c) the effect of this scheme on atomic positions has been displayed for $\epsilon = 0$ Å (the original model), $\epsilon = 0.3$ Å and $\epsilon = 0.5$ Å. The $\phi_{calc,1}$ phase and $|\mathbf{F}_{calc,1}|$ information was extracted from each modified coordinate model using the program *SFALL*. σ_A-weighted map coefficients were then calculated at each dose using *SIGMAA* in order to avoid needing to run any additional rounds of refinement to extract updated maps. Since *EDSTATS* requires that the unit cell dimensions for each input map and the coordinate model be identical, *SFTOOLS* was used to fix the unit cell dimensions specified in all generated `.mtz`-format files at each higher dose to those of the first dataset. *FFT* was then used to calculate maps over the extent of the unit cell dimensions for the first dataset.
The perturbation scheme presented in this figure is identical to that described later in Section 2.5.9.5 to test the sensitivity of the *RIDL* pipeline to phase error when calculating $F_{obs,n} - F_{obs,1}$ Fourier difference maps. In that section, it is demonstrated that for the case of the *Dp*Cel7B test case, *RIDL* is robust to an overall mean phase error of $\sim 40^\circ$, corresponding here to $\epsilon = 0.5$ Å. In contrast, in the current plots, the RSZD metric is sensitive to much smaller perturbations ($\epsilon = 0.2$ Å corresponding to a mean phase error of just 22.5°).

programs (Winn *et al.*, 2011) to generate $F_{obs,n} - F_{obs,1}$ maps and to determine a local region of interest around each refined atom within the structure (the *search radius*), for each dataset $n > 1$. *RIDL* then proceeds to calculate per-atom summary metric values for each higher dose dataset ($n > 1$) sequentially. The logical flow for the complete pipeline is provided in Figure 2.16. The following sections will next discuss the construction of this pipeline; the relevant section to each component of Figure 2.16 has been labelled on the figure.

The functionality of each step will be demonstrated for the *Dp*Cel7B protein MX damage series introduced in Section 2.3.

2.5.3 Data preparation for the *RIDL* pipeline

RIDL can be used to analyse MX damage dataset series collected from either single crystals or merged multi-crystal datasets. Here an MX damage series is defined as a series of `.pdb`-format coordinate models (e.g. output from refinement programs *REFMAC* (Murshudov *et al.*, 2011) or *phenix.refine* (Adams *et al.*, 2010)) and `.mtz`-format merged and scaled structure factor files (e.g. output from *CTRUNCATE*), corresponding to disjoint wedges of data collected on the same crystal at increasing doses. The information for a full MX damage series consisting of a low-dose dataset and either single or multiple higher dose datasets must be specified within a single plain-text input file to be parsed by *RIDL*. An example input file is provided in Figure 2.17. This input file format is efficient, since it allows multiple higher dose datasets to be specified and analysed in a single run of *RIDL*.

As will be discussed in Section 2.5.4, at each higher dose, a $F_{obs,n} - F_{obs,1}$ map is calculated using the $|\mathbf{F}_{obs,1}|$ and $|\mathbf{F}_{obs,n}|$ columns specified by the `mtzlabels1` and `mtzlabels2` labels, and with phases retrieved from the `mtz3` file (which may or may not differ from the `mtz2` file). It is important to note that in the current implementation of *RIDL*, by default each $F_{obs,n} - F_{obs,1}$ map will be calculated over the unit cell dimensions as specified by each file `mtz2`. As discussed in Section 2.5.5, for each higher dose `mtz2` file, the $F_{obs,n} - F_{obs,1}$ map will be overlaid over the corresponding `pdb2` coordinate model in *RIDL*, such that the atomic positions specified in that coordinate model file will define the local region of space in the $F_{obs,n} - F_{obs,1}$ map that will be assigned to each atom. As such it is a strict requirement that each `mtz2` and `pdb2` file pair

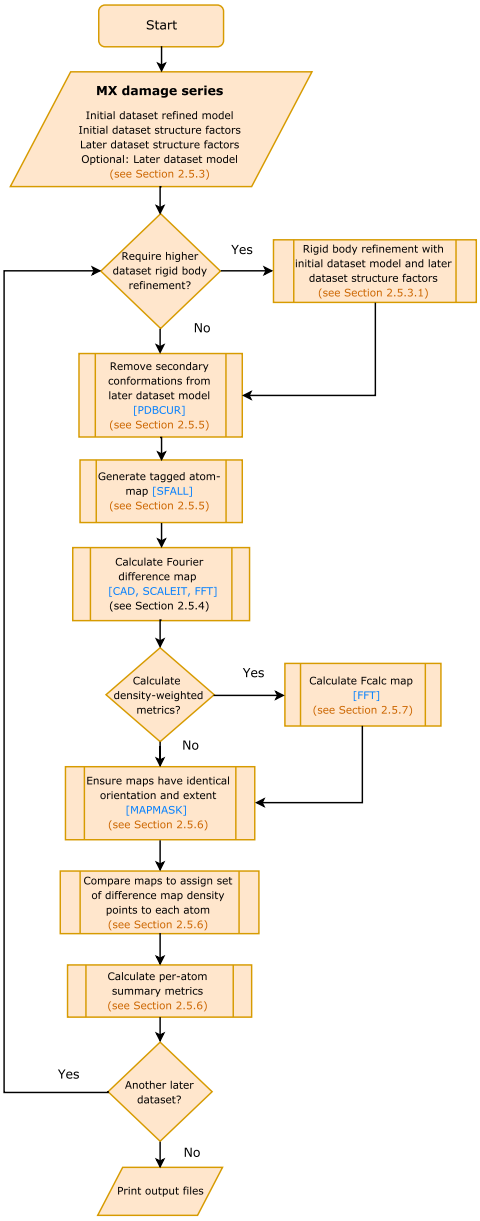


Figure 2.16: Flow chart illustrating the structure of the *RIDL* pipeline. CCP4 programs used by *RIDL* have been coloured in blue. The sections corresponding to each flowchart step are in brown. Please refer to the corresponding sections for full details of each step.

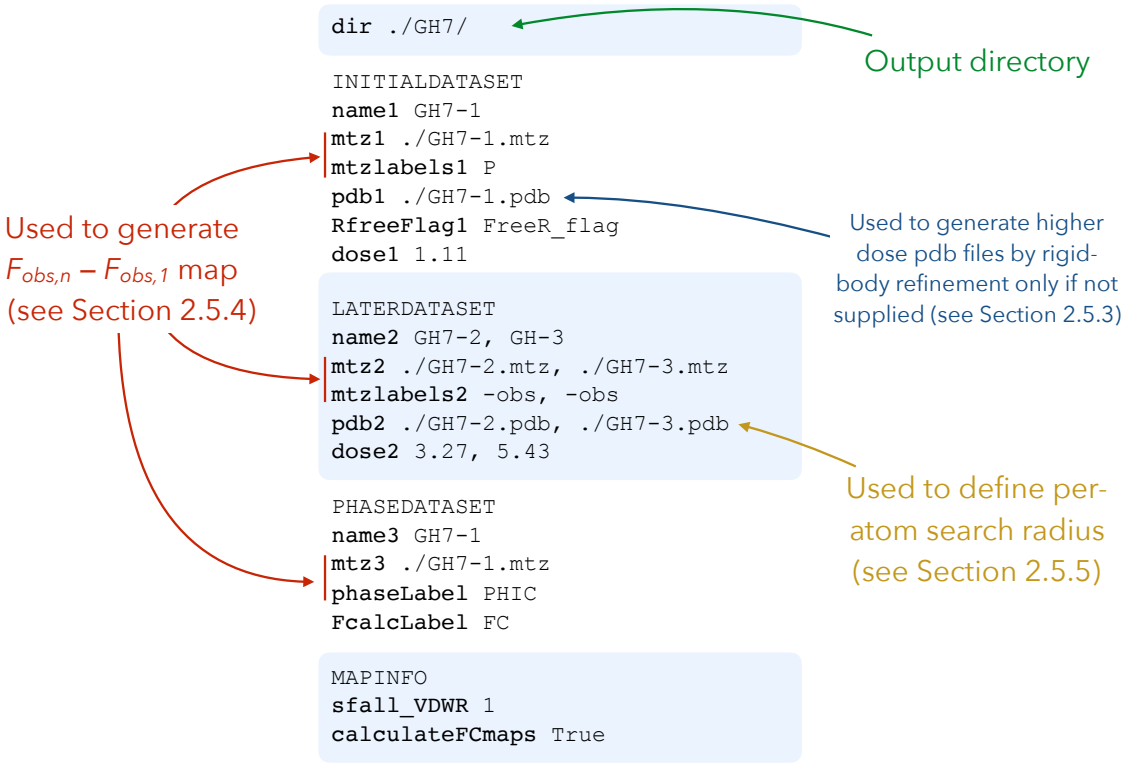


Figure 2.17: Example `.txt` input file for the *RIDL* pipeline. This is for the example of the *Dp*Cel7B case study, in which $F_{obs,2} - F_{obs,1}$ and $F_{obs,3} - F_{obs,1}$ maps are calculated, corresponding to the 3.27 MGy and 5.43 MGy higher dose datasets. Further higher dose datasets can be included by appending to the comma-separated lists in the `LATERDATASET` section. At each higher dose, a $F_{obs,n} - F_{obs,1}$ difference is calculated using the $|\mathbf{F}_{obs,1}|$ and $|\mathbf{F}_{obs,n}|$ columns specified by the `mtzlabels1` and `mtzlabels2` labels, and with phases retrieved from the `mtz3` file, which is fixed as the phase set corresponding to dataset 1 in this example. The α parameter used by *SFALL* (see Section 2.5.5) is set by the input `sfall_VDWR` to the default value of 1 used by *RIDL*. Since the `calculateFCmaps` parameter is set to `True`, *RIDL* will additionally calculate density-weighted summary metric values with $|\mathbf{F}_{calc}|$ amplitudes specified by the `FcalcLabel` label (see Section 2.5.6).

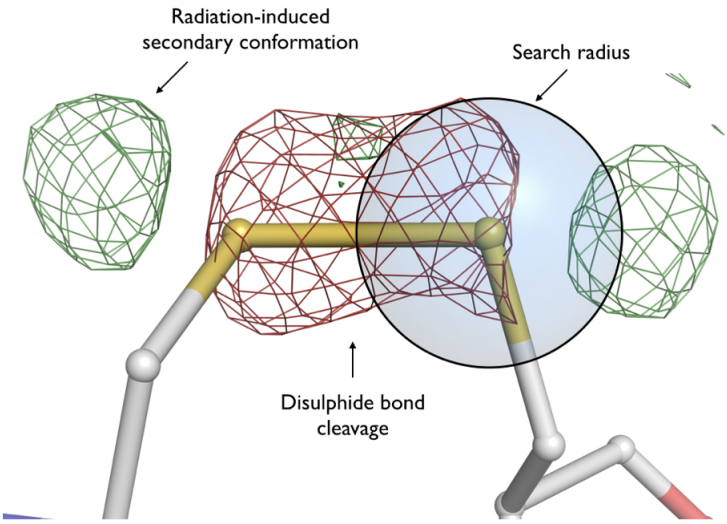


Figure 2.18: For the *Dp*Cel7B case study, the $F_{obs,2} - F_{obs,1}$ difference map indicates a disulphide bond cleavage event between Cys-44 (left) and Cys-38 (right) in chain A of the protein. The difference map is contoured here at $\pm 4\sigma$ in green/red. The search radius for the Cys-38-S_γ atom, as defined by the corresponding atom-tagged map output from *SFALL*, is also shown. Positive difference density (in green) indicates the presence of a stable radiation-induced secondary conformation for both Cys-S_γ atoms in dataset 2. These secondary conformations are not built into the coordinate model for dataset 2, such that the Cys-S_γ search radius samples a conserved density region per dataset over which summary metrics will be calculated.
It is noted that in general, depending on the extent of conformational change of each cysteine following disulphide bridge cleavage, the positive $F_{obs,n} - F_{obs,1}$ map peak may not be situated within the search radius of each sulphur atom (as defined by Section 2.5.5). In those cases, the conformational change will remain undetected by *RIDL*. However, the original site of electron density loss can always be monitored.

must be defined over identical unit cell dimensions.

2.5.3.1 General considerations for generation of higher dose coordinate models

Since *RIDL* aims to extract electron density change information from $F_{obs,n} - F_{obs,1}$ difference maps near atomic positions at increasing dose, if protein side-chains are permitted to take new rotameric positions at different doses (e.g. reorientation of cysteine sulphurs following disulphide bond cleavage, Figure 2.18), the region of space sampled is different at each dose, and not truly measure the effect of dose on electron density change at the atom's original position on refinement of the first structure. It is thus evident that the coordinate model (`pdb2`) should not deviate significantly between different higher dose datasets $n > 1$.

It is therefore suggested that for each higher dose dataset, only rigid-body refinement should be performed to generate a suitable higher dose coordinate model (using the higher dose $|\mathbf{F}_{obs,n}|$

structure factor amplitudes coupled with the initial dataset coordinate model). This refinement will account for whole-molecule rotation/translation within the unit cell induced by irradiation, as well as unit cell expansion, as first documented by Ravelli and McSweeney (2000) for crystals held at 100 K. However, this will not permit relative movement of atoms. It is noted that in the current implementation of *RIDL*, at least one round of rigid body refinement must be performed to ensure that each `mtz2` and `pdb2` file pair are defined over identical unit cell dimensions, as required.

The proposed rigid body refinement will also ensure that the atom numbering format is conserved for successive `.pdb`-format coordinate models, and the behaviour of an individual atom can be more-easily tracked across datasets. If the higher dose coordinate models supplied by the user do not satisfy the criterion of a conserved atom numbering between datasets, *RIDL* can call *REFMAC* directly to perform this rigid-body refinement step at run-time.

The requirement for rigid body refinement has been enforced throughout the current thesis. If alternatively all $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps had been generated over the unit cell dimensions for the first dataset, such a rigid-body refinement step could have been omitted. Future work will aim to implement this alteration. However, as will later be demonstrated in Figure 2.32 in Section 2.5.9.2, the output to *RIDL* is insensitive to both the use of rigid body refinement and whether unit cell dimensions corresponding to dataset 1 or n are used to calculate the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map.

2.5.3.2 Default input parameters for *Dp*Cel7B test case

In the current demonstration of *RIDL* for the *Dp*Cel7B test case, unless otherwise specified, to satisfy the above criteria, higher dose coordinate models have been generated following 10 cycles of rigid-body refinement coupled with isotropic *B*-factor refinement in *REFMAC*, to refine the initial dataset coordinate model against the higher dose $|\mathbf{F}_{\text{obs,n}}|$ structure factor amplitudes. Please refer back to Section 2.3.3.1 for the full details of this refinement strategy.

In the current example, it is noted that the per-atom search radii as defined by Equation 2.5.4 in Section 2.5.5 will depend on the refined isotropic *B*-factors at each dose. Isotropic *B*-factor refinement has been included to account for the increased uncertainty in the location of each atom with increasing dose, with the effect of broadening the per-atom map search radius with

dose (see Section 2.5.5). However, the decision to include *B*-factor refinement has been shown to not significantly affect the results of *RIDL* (see Figure 2.33 later in Section 2.5.9.3).

In fact, in Section 2.5.9 below, it has been demonstrated in Figure 2.33 that the *Dp*Cel7B test case is in fact highly robust to the effects of full restrained coordinate refinement at high doses. However in general, rigid-body refinement is still a safer option.

2.5.4 $F_{obs,n} - F_{obs,1}$ difference map generation

The protocol for generating $F_{obs,n} - F_{obs,1}$ maps is similar to that detailed in Section 2.4.1 for the *PEAKS* pipeline. For each higher-dose dataset of an MX damage series $n > 1$, *CAD* combines the observed structure factor amplitudes $|\mathbf{F}_{obs}|$ and corresponding standard deviation $\sigma\mathbf{F}_{obs}$ columns from the merged `.mtz`-format input files for the initial (`mtz1`) and higher-dose datasets (`mtz2`). *CAD* maintains a record of the unit cell dimensions from both input `.mtz`-format files, and these may differ.

The high-dose $|\mathbf{F}_{obs,n}|$ values are then scaled to the low dose $|\mathbf{F}_{obs,1}|$ reference set with *SCALEIT*, performed with the default anisotropic temperature factor scaling function of the form:

$$C \cdot e^{-(h^2B_{11}+k^2B_{22}+l^2B_{33}+2hkB_{12}+2hlB_{13}+2klB_{23})} \tag{2.5.3}$$

being applied to a higher dose $|\mathbf{F}_{obs,n}|$ value for Miller index hkl, where C and B_{ij} $(i, j \in \{1, 2, 3\})$ are the scale and the anisotropic temperature factor values that must determined through a least squares fitting approach to all the structure factors in *SCALEIT*. This scaling step is performed sequentially for all higher-dose datasets of the damage series, such that all datasets are scaled to an identical reference set. Figure 2.19 indicates that the calculation of the summary metrics by *RIDL* (see Section 2.5.6 below and Quick reference box 1 for the details of metric construction) can be influenced by the scaling protocol declared by *SCALEIT*, and consequently the specified scaling function should remain fixed between independent runs of *RIDL*. In Figure 2.19a, anisotropic scaling differs significantly from the other tested scaling protocols (isotropic or single scale factor scaling); manual inspection of the $F_{obs,2} - F_{obs,1}$ map revealed that only in the map resulting from anisotropic scaling could clear site-specific damage events be distinguished. At higher doses,

however, there is less difference between scaling regimes (Figure 2.19d).

Quick reference box 1: The D_{loss} metric

This reference box is designed to provide the necessary information regarding summary metrics calculated by *RIDL* that is required for the current section. For a detailed discussion on each summary metric, refer to Section 2.5.6.

- The D_{loss} metric is a per-atom summary metric derived from a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map by the program *RIDL*. The full definition will be provided below in Section 2.5.6. It is defined as the most negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map value attained in the localised region of space around an atom (as specified by the *SFALL* α parameter), with units of electrons/Å^3.
- The notation $D_{\text{loss}}^{\text{C}_\alpha}$ corresponds to the C$_\alpha$-normalised D_{loss} metric, such that the the D_{loss} value for each atom is normalised to the mean D_{loss} value attained by the set of protein backbone C$_\alpha$ atoms. In Section 2.5.8 below, two normalisation schemes, ${}^1D_{\text{loss}}^{\text{C}_\alpha}$ (given by Equation 2.5.10), and ${}^2D_{\text{loss}}^{\text{C}_\alpha}$ (given by Equation 2.5.14) will be explicitly presented and tested.

FFT (Ten Eyck, 1973) is then used to generate a $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference map covering the whole crystal unit cell dimensions corresponding to the higher dose dataset n. This is a consequence of the order in which the $|\mathbf{F}_{\text{obs}}|$ columns are supplied to *FFT*: the $|\mathbf{F}_{\text{obs,n}}|$ column must be supplied first to generate a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map as opposed to a $F_{\text{obs,1}} - F_{\text{obs,n}}$ map.

A model phase set ϕ_{calc} must be supplied by the user through the *RIDL* input file (as retrieved from file `mtz3` in Figure 2.17). It is recommended to use an identical fixed phase set for all datasets in the damage series, such that the experimentally recorded $|\mathbf{F}_{\text{obs}}|$ remain the only variable between increasing dose datasets. However, the *RIDL* input file can specify differing phase sets per dataset if required by the user. For the case of the *Dp*Cel7B example, the phase set $\phi_{\text{calc,1}}$, corresponding to the initial coordinate model have been used throughout unless otherwise specified.

It is also important to note that crucially each generated $F_{\text{obs,n}} - F_{\text{obs,1}}$ map must have identical grid dimensions to the corresponding atom-tagged map for the same higher dose dataset, as

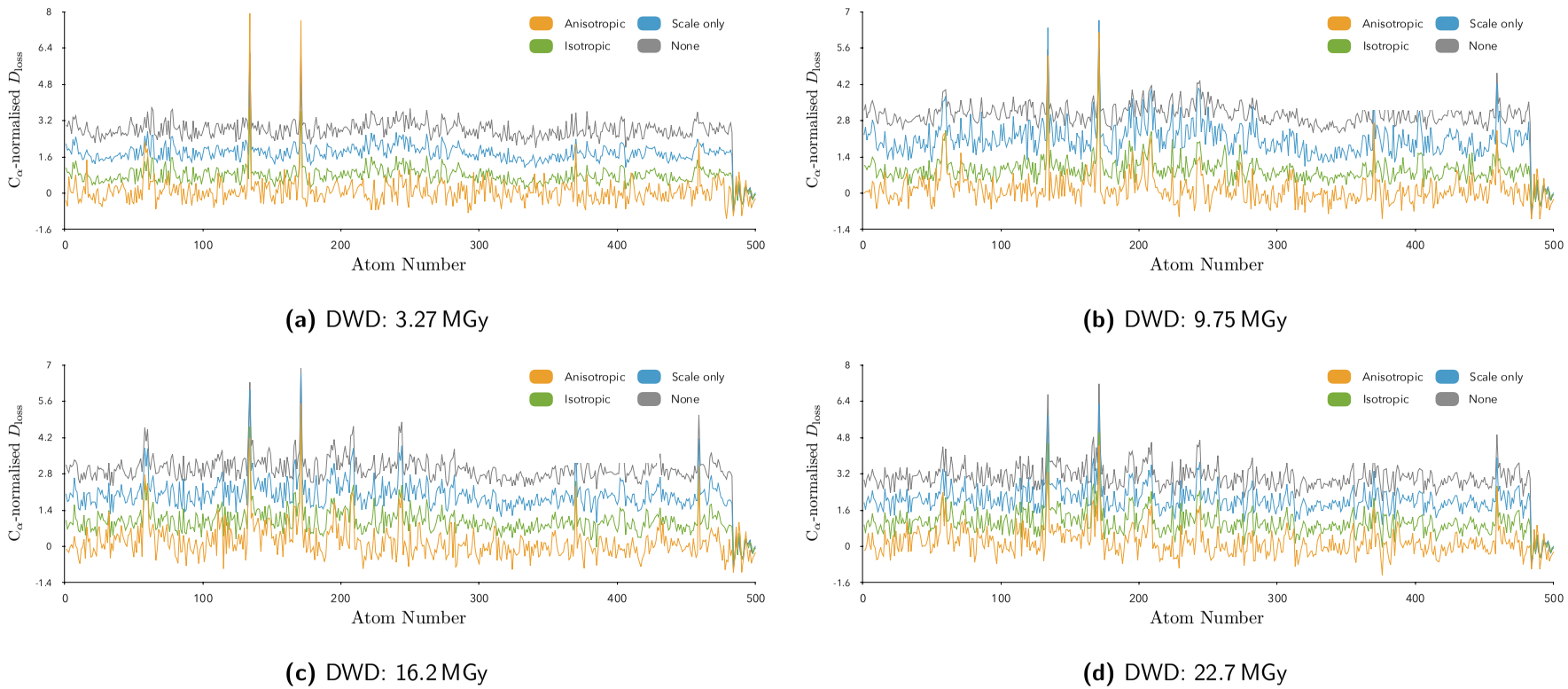


Figure 2.19: For the case of the *Dp*Cel7B GH7 family protein, the C_α-normalised D_{loss}(atom) metric has been calculated using *RIDL* for varying scaling protocol using *SCALEIT*. Metric values have been derived from $F_{obs,n} - F_{obs,1}$ maps with higher dose dataset n corresponding to 3.27, 9.75, 16,2 and 22.7 MGy. The tested scaling functions are: an anisotropic *B*-factor scaling, an isotropic *B*-factor scaling, an overall scale, and no scaling. To aid visibility, metric values for isotropic scaling (green), overall scaling (blue) and no scaling (grey) have been translated vertically by 1, 2 and 3 units in the y-axis in each plot. Only the first 500 atoms of the protein have been included for clarity. Other intermediate doses (not shown) exhibited similar behaviour.

generated in Section 2.5.5. In practice this is achieved by first generating the atom-tagged map using *SFALL* (see the next section), and then using the *GRID* command within *FFT* to force the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map to have identical grid dimensions.

2.5.5 Assignment of a search radius per-atom

For each higher-dose coordinate model, a local region of space to be attributed to each atom can be determined using *SFALL*, in the form of a *tagged atom-map* generated over the unit cell dimensions corresponding to dataset n (identical to that of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map). As discussed in Section 2.5.6, the atom-map can then be used as a template to determine which voxels in each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map should be assigned to each atom.

Each grid point in this map is assigned an atom number rather than an electron density value, but the map has an identical content format to the CCP4-format `.map` $F_{\text{obs,n}} - F_{\text{obs,1}}$ map generated by *FFT*. The main benefit of using *SFALL* to perform this task is that the program automatically calculates a non-overlapping region of space for each atom, which factors in the relative contribution of electron density to that region by each local atom. Here, *SFALL* assumes that electron density is distributed as a Gaussian centred at each atomic position, which is dependent on both element identity and isotropic *B*-factor. The calculation of the tagged atom-map is performed independently for each higher dose dataset $n > 1$.

Prior to running *SFALL*, *RIDL* first calls *PDBCUR* at runtime to removes any hydrogen atoms, anisotropic *B*-factors and secondary conformations from all input coordinate models. In this step, any anisotropic *B*-factors are replaced by equivalent isotropic *B*-factors, as required for *SFALL* to run. By default, for an atom with isotropic *B*-factor $B > 0$, *SFALL* then assigns a maximum search radius of:

$$r = \alpha \cdot \frac{\sqrt{B + 25}}{2\pi} \tag{2.5.4}$$

around the atomic centre (where $\alpha > 0$ is a user-defined scalar of default value 2.5), provided this region does not overlap with that of any adjacent atoms. Figure 2.20 provides an illustration of how such non-overlapping regions are defined, and also how the atom-map is utilised in the *RIDL* pipeline.

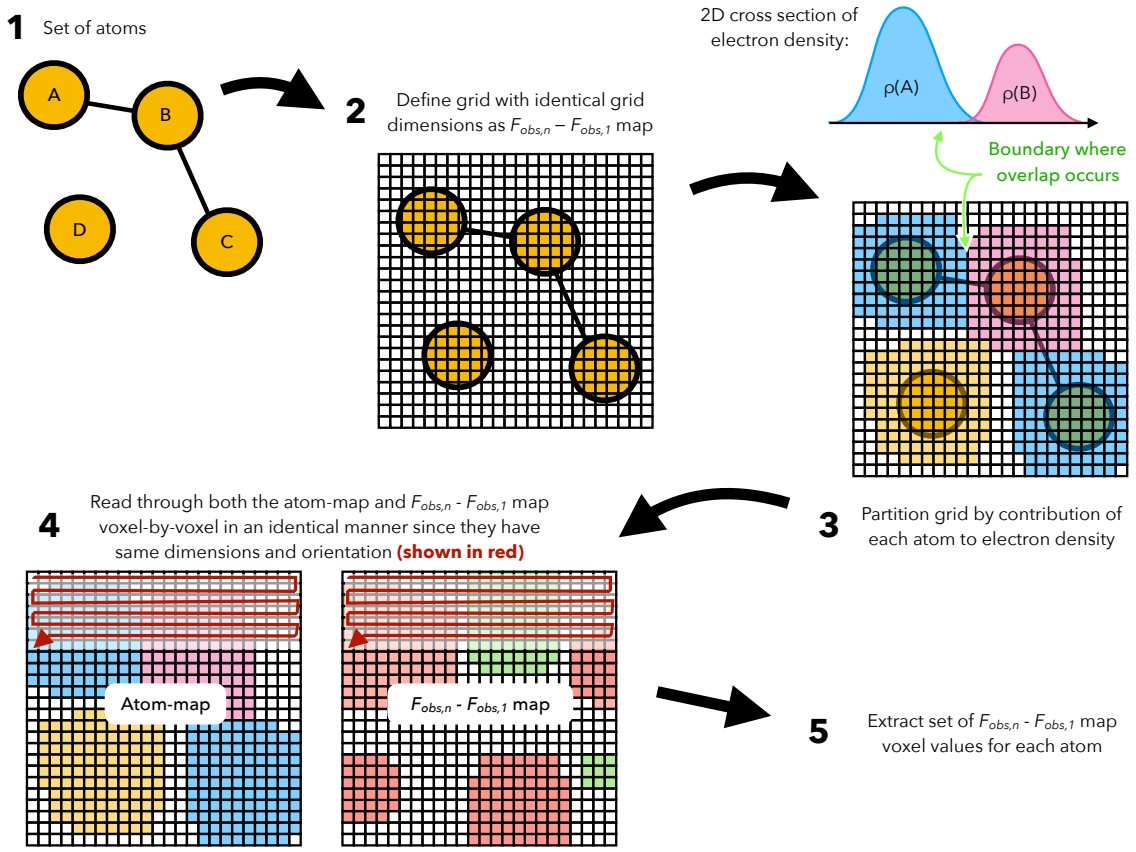


Figure 2.20: A schematic illustrating the role of *SFALL* in the *RIDL* pipeline. *SFALL* incorporates steps 2 and 3, in which a *atom-map* `.map` file is created with identical grid dimensions as the corresponding $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. As illustrated in step 3, a voxel in this atom-map is assigned to a particular atom if the predicted electron density ρ associated with that atom contributes more to that voxel than any other atom in the structure, and the voxel falls within a radius from the atom given by Equation 2.5.4. Voxels may be unassigned to atoms (left blank) if all atoms contribute negligible electron density to it. In step 4, *RIDL* reads the atom-map and associated $F_{\text{obs,n}} - F_{\text{obs,1}}$ map binary files byte by byte. Since the dimensions and orientation of the two maps over the unit cell is identical, at a given position through reading each file, the atom-map file will indicate which atom (if any) a voxel is assigned, and the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map file will indicate the corresponding difference density value at that voxel. A list of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density values per atom can be quickly compiled.

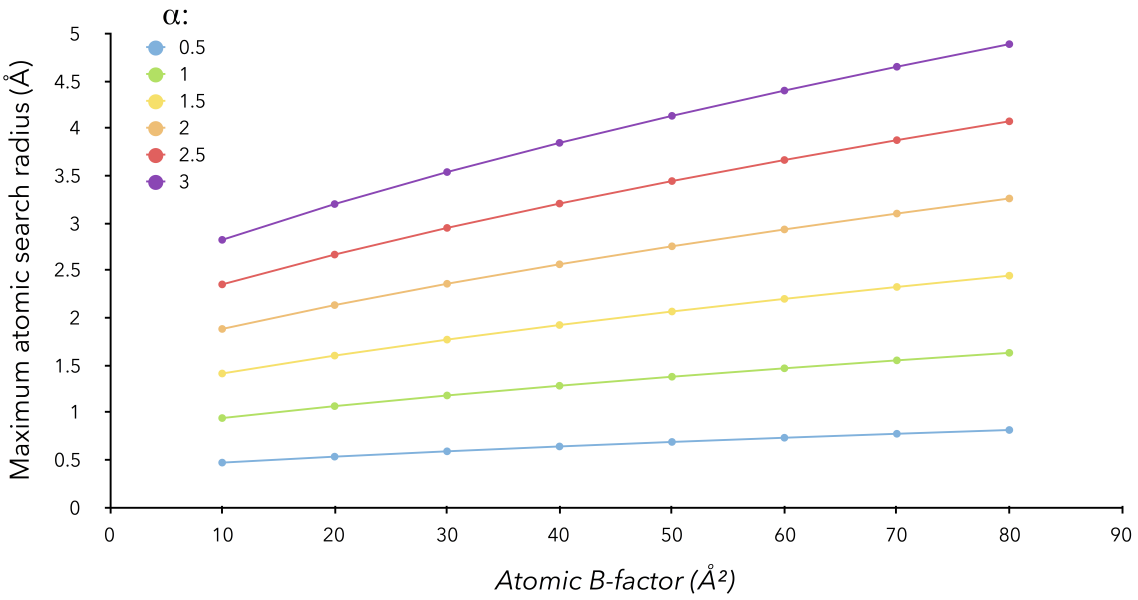
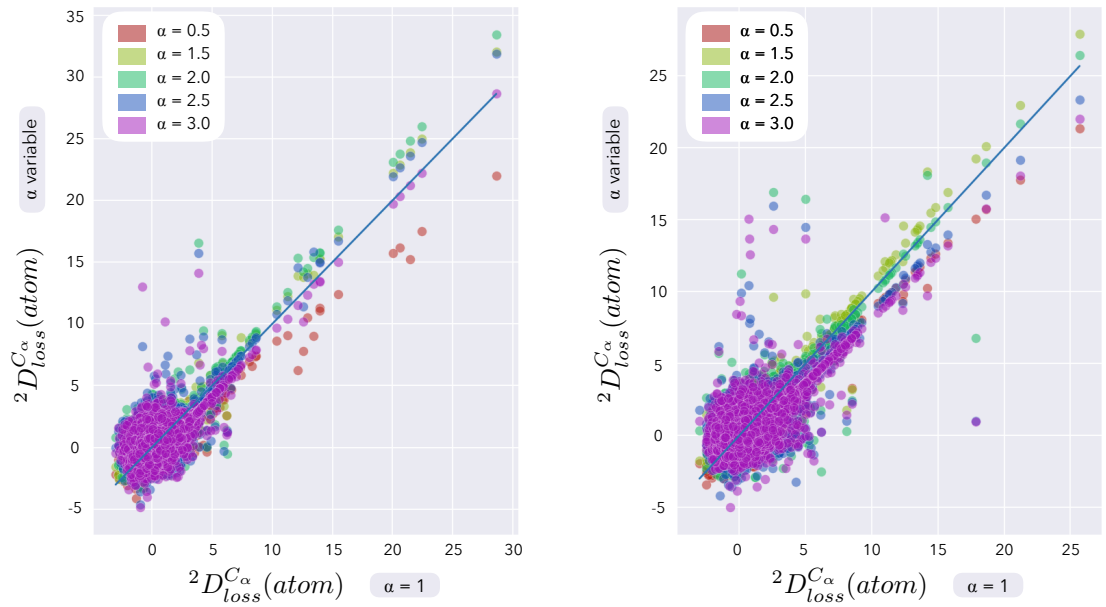


Figure 2.21: The α parameter included in the calculation of the *B*-factor-dependent atomic search radius by *SFALL* (specified in Equation 2.5.4) has been varied and its effect on the search radius for atoms of different *B*-factors monitored.

For the purposes of specific damage detection in *RIDL*, $\alpha = 1$ has been fixed as a default, such that only density changes in close proximity to atomic centres (< 2 Å, Figure 2.21) are considered. Indeed, for the *Dp*Cel7B protein investigated in Section 2.4, the majority of peaks that were visually deemed to correspond to damage were located at a distance below 1 Å from the corresponding atomic centre (Figure 2.11). Figure 2.21 indicates that $\alpha = 1$ is suitable for the inclusion of these peaks, for the range atomic *B*-factors typical of MX coordinate models ($10 - 80$ Å^2). The default α value can be overridden by the user if required.

The effect of varying α on the calculation of the summary metrics is illustrated in Figure 2.22 (see Section 2.5.6 below and Quick reference box 1 for the details of metric construction). For sites of high electron density loss ($D_{\text{loss}}^{C_\alpha} > 5$), the ordering of damage is well conserved for differing α. It is unsurprising that metric values are less conserved for undamaged atoms (with $|D_{\text{loss}}^{C_\alpha}| \leq 2$), since these atoms are more susceptible to the inclusion of background map noise as α is varied. These atoms typically still attain low metric values irrespective of α, however exceptional cases are apparent (with a low x-axis value and high y-axis value) where an increased search radius around an atom has significantly increased the $D_{\text{loss}}^{C_\alpha}$ metric value. It is suggested that the α value should be fixed as a constant value, to ensure that the calculation of summary metrics is



(a) DWD: 3.27 MGy **(b)** DWD: 9.75 MGy

(c) DWD: 16.2 MGy **(d)** DWD: 22.7 MGy

Figure 2.22: For the example case of the *Dp*Cel7B GH7 family protein, the C_α-normalised D_{loss}(atom) metric (Equation 2.5.14, Section 2.5.8) has been calculated using *RIDL* for varying values of the parameter α specified in Equation 2.5.4, as used by *SFALL*. Metric values have been derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps with higher dose dataset n corresponding to 3.27, 9.75, 16,2 and 22.7 MGy. Each x-axis indicates the D_{loss}(atom) metric values across all atoms in *Dp*Cel7B using $\alpha = 1$ (the default value assigned in *RIDL*). The corresponding y-axes indicate metric values at each other tested alpha value between 0.5 and 3. Deviations of points from the line $y = x$ (in blue) indicate atoms for which the metric is sensitive to the choice of α. Other doses for the *Dp*Cel7B damage series (not shown) exhibited similar behaviour.

comparable between different structures or data collection sessions. In the work presented in this thesis, $\alpha = 1$ will be fixed unless otherwise stated.

It is noted that for the *Dp*Cel7B example case, since isotropic *B*-factor refinement has been performed independently for each higher dose dataset, this will lead to a different *B* value in Equation 2.5.4 at each dose for a given atom. Indeed the mean isotropic *B*-factor over all refined atoms in *Dp*Cel7B was reported to deviate between 27.4 Å^2 and 40.2 Å^2 across the 11 datasets in Table 2.2. As indicated later in Figure 2.33 of Section 2.5.9.3, this variability has been tested and it has been deemed that the results of *RIDL* are insensitive to the decision to perform isotropic *B*-factor refinement for the *Dp*Cel7B protein test case.

2.5.5.1 Alternative available approaches

It must be noted that the program *MAPMAN* (Kleywegt and Jones, 1996) could also have been used by *RIDL* to extract electron density values at atomic positions. In order to extract an electron density information, *MAPMAN* will either interpolate, average or integrate over grid points surrounding an atom, but it does not output the set of raw voxel values assigned to each atom. Since a key aim of the current investigation is to design metrics to best quantify site-specific damage (Section 2.5.6), the *MAPMAN* output was thus considered suboptimal. Furthermore, this approach was not chosen for use in *RIDL* since *MAPMAN* only permits a single search radius for all atoms in a coordinate model, that does not account for the spread of electron density as parametrised by individual atomic *B*-factors.

It must also be noted that Tickle (2012) has reviewed the calculations of search radii used by different programs (*MAPMAN* and *SFALL*), and has implemented a more sophisticated search radius in *EDSTATS* that depends on atomic *B*-factor, atom type and map resolution. For a given atom, this involves determination of a radius r_{max} for the radius integral:

$$\int_0^{r_{\mathsf{max}}} \rho_{\mathsf{calc}}(r) dr \tag{2.5.5}$$

such that the integral attains a value that is 95 % of the theoretical value at an infinite radius.

In the above equation, ρ_{calc} is defined as:

$$\rho_{\text{calc}}(\mathbf{x}) = \frac{1}{V}\sum_{hkl} |\mathbf{F}_{\text{calc}}| \cdot \exp(i\phi_{\text{calc}}) \cdot \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}) \tag{2.5.6}$$

where V is the volume of the unit cell, and $|\mathbf{F}_{\text{calc}}|$ and ϕ_{calc} are the calculated structure factor amplitude and phase set derived from the coordinate model corresponding to coordinate model.

In the current development of the *RIDL* pipeline, such a radius determination scheme has not been explicitly implemented. As will be presented later in Section 2.5.7, instead, *RIDL* provides the option to additionally construct damage metrics that weight the importance of each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxel by the relative ρ_{calc} density at that voxel. Full details of this construction will be discussed in Section 2.5.7. This aims to bypass the requirement for a strict search radius, as defined by Tickle (2012), since $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxels located far from an atomic centre will then contribute minimally to a damage metric.

2.5.6 Calculation of per-atom summary metrics

For each dataset $n > 1$, the tagged atom-map and $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps are both cropped to the crystal asymmetric unit corresponding to the dataset n coordinate model with *MAPMASK*, such that resulting maps have exactly the same grid sampling dimensions and orientation in the asymmetric unit. The tagged atom-map can now be used to indicate the exact set of voxels indices, V_{atom}, in the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map that reside in the defined search radius of an atom.

From the collection of local density values, V_{atom}, a multitude of possible summary metrics can derived. Consequently, care must be taken to derive metrics that can suitably quantify radiation-induced electron density changes. Table 2.6 presents the formulation of several metrics and discusses the benefits and limitations of each. As demonstrated in Figure 2.24, several of these candidate metrics appear suitable for the detection of site-specific damage to atoms. Furthermore, as indicated in Figure 2.25, each metric does not entirely agree, and therefore justification is required when deciding on the most reliable damage indicator.

In MX, the radiation-induced cleavage of chemical bonds and disordering of atoms leaves a signature of electron density loss in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps. Consequently, a metric such as D_{loss}

Table 2.6: Construction of possible summary metrics to quantify radiation-induced electron density change to individual atoms. For each metric, a shorthand name has been provided as well as the mathematical formulation for the metric. Here, $\rho_{\Delta}(v)$ denotes the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density value at a voxel $v \in V_{\text{atom}}$. $|V_{\text{atom}}|$ denotes the number of voxels assigned to a particular atom. V^{-}_{atom} and V^{+}_{atom} correspond to the set of voxels attributed to an atom that attain $\rho_{\Delta}(v) < 0$ and $\rho_{\Delta}(v) \geq 0$ respectively.

Metric	Construction	Pros	Cons
D_{mean}	$\frac{1}{\lvert V_{\text{atom}}\rvert} \cdot \sum_{v \in V_{\text{atom}}} \rho_{\Delta}(v)$	D_{mean} takes into account the behaviour of all density values near an atom D_{mean} could provide information on the behaviour of both the -ve and +ve density near an atom	D_{mean} is sensitive to the search radius extent defined by *SFALL* For the scenario of a large -ve and +ve peak near an atom, D_{mean} will be indistinguishable from the scenario were there are no peaks (both will attain $D_{\text{mean}} \sim 0$)
D_{loss}	$\max_{v \in V_{\text{atom}}} [-\rho_{\Delta}(v)]$	D_{loss} is directly comparable to the manual approach of quoting $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peak heights, and therefore in principle should agree with manual inspection D_{loss} allows a direct detection of which atoms have experienced significant density loss D_{loss} is not affected by +ve density as the result of nearby noisy background map behaviour The same D_{loss} value is achieved wherever the maximum -ve density value is situated inside the atomic search radius. As such, D_{loss} does not depend on the perfect positioning of an atom during model building.	D_{loss} only provides a maximum peak height for one single $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peak. If there are other significant nearby peaks, these are not taken into account. D_{loss} does not account for the distance away from an atomic centre, and is therefore susceptible to map noise in the vicinity of the atom D_{loss} is not representative of all voxels inside the search radius of an atom. The maximum value can overestimate the significance of a peak
D_{gain}	$\max_{v \in V_{\text{atom}}} [\rho_{\Delta}(v)]$	See pros for D_{loss}	See cons for D_{loss}
D^{-}_{mean}	$\frac{1}{\lvert V^{-}_{\text{atom}}\rvert} \cdot \sum_{v \in V^{-}_{\text{atom}}} \rho_{\Delta}(v)$	D^{-}_{mean} takes into account the behaviour of all negative density values near an atom and is less sensitive to outliers than D_{loss}	Similarly to D_{mean}, D^{-}_{mean} is again sensitive to the search radius extent defined by *SFALL*
D^{+}_{mean}	$\frac{1}{\lvert V^{+}_{\text{atom}}\rvert} \cdot \sum_{v \in V^{+}_{\text{atom}}} \rho_{\Delta}(v)$	See pros for D^{-}_{mean}	See cons for D^{-}_{mean}

(units of e/Å^3), which is sensitive to electron density loss near atoms, appears to be suitable for the purposes of specific damage detection. D_{loss} shares a direct link with manual inspection of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks; for a negative peak in an $F_{\text{obs,n}} - F_{\text{obs,1}}$ map located in the vicinity of an atom, D_{loss}(atom) will exactly equal the height of the peak in e/Å^3. Indeed, in Figure 2.23, a comparison of the top damage sites flagged by the *PEAKS* pipeline (Section 2.4) with the D_{loss} metric indicates that these metrics exhibit a strong correlation for the *Dp*Cel7B protein test case.

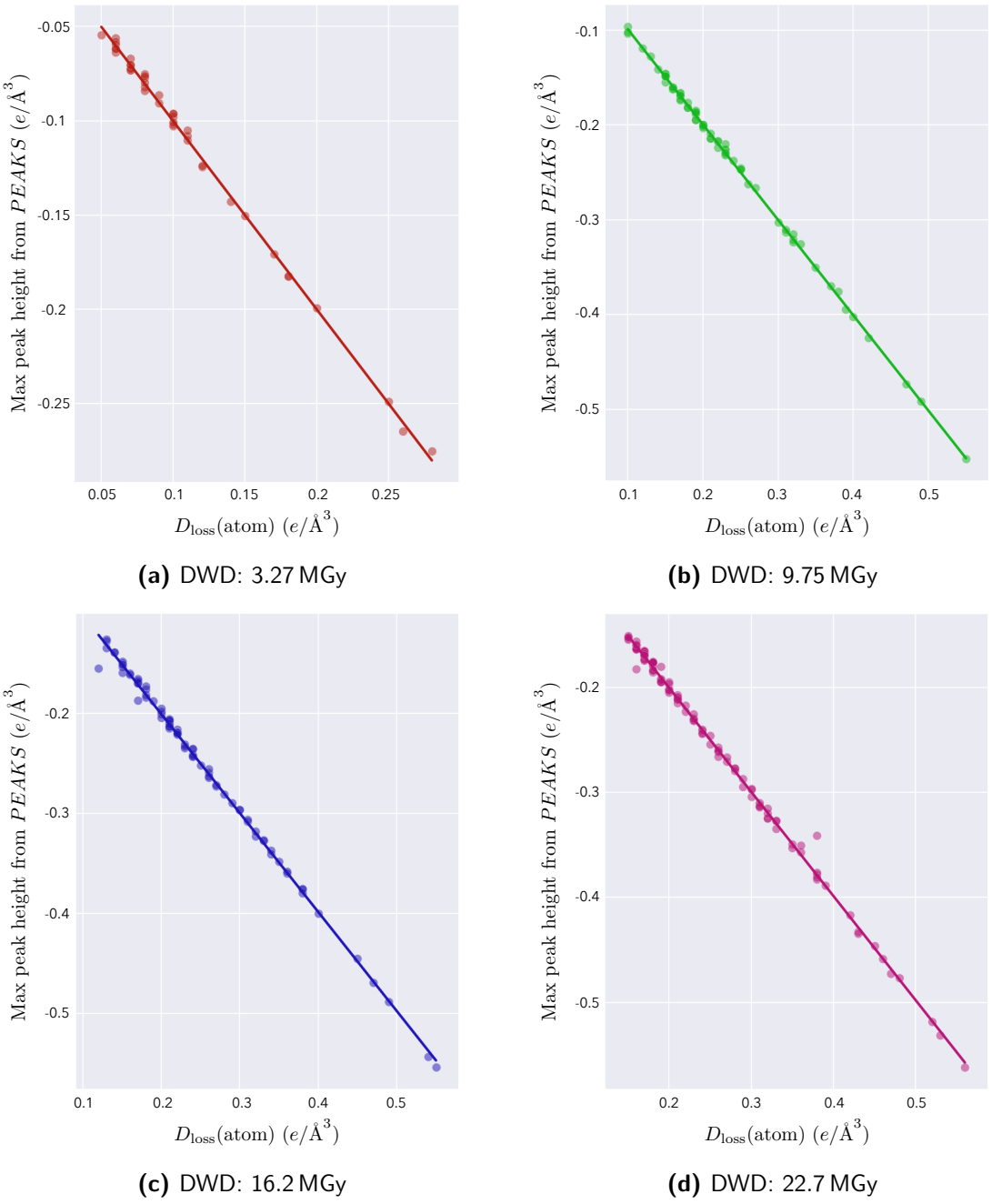
A limitation of the D_{loss} metric is that it is inherently susceptible to false positives, where map noise in the vicinity of an atom can be incorrectly interpreted as damage to the atom. Despite this, the false negative rate associated with D_{loss} is negligible by design, making D_{loss} an ideal indicator of potential damage sites that should be flagged for the attention of a user. In general, D_{loss} provides a worst case scenario for the electron density lost by an atom, and is deemed a useful metric for screening for damage.

Figure 2.24 indicates that the D^{-}_{mean} and D_{mean} metrics also exhibit the power to predict radiation-sensitive sites. A benefit of D^{-}_{mean} over D_{mean} is that the former only averages over voxels for which the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density is negative, and is thus deemed a more suitable quantifier of electron density *loss* around an atom compared to D_{mean}. Unfortunately, metrics that require any averaging or integration of voxel values inside a specified search radius are particularly sensitive to the search radius used, due to the inclusion/exclusion of voxels as the search radius is varied. As discussed in Section 2.5.5, the definition of a *correct* search radius is non-trivial, and consequently these metrics are suboptimal.

2.5.7 Density-weighted metrics

An obvious problem with the D_{loss} metric defined above is that an atom can falsely attain a high D_{loss} value due to noisy background behaviour. This has motivated the development of additional metrics that attempt to mitigate this issue.

For a set of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxels, V_{atom}, assigned to an atom, it was decided to test a scheme in which each voxel is weighted by how much of the atom's electron density contributed to that voxel. In this scheme, a voxel that is located further from an atom (but still inside the atom's search radius) will be down-weighted relative to a voxel positioned at the atomic centre.



(a) DWD: 3.27 MGy

(b) DWD: 9.75 MGy

(c) DWD: 16.2 MGy

(d) DWD: 22.7 MGy

Figure 2.23: For the example case of the *Dp*Cel7B GH7 family protein at 3.27, 9.75, 16,2 and 22.7 MGy, the D_{loss}(atom) metric has been calculated using *RIDL*, and compared directly to the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map negative peak heights (in e/Å^3) detected by the program *PEAKS* (Section 2.4). Only atoms detected to be damaged by the *PEAKS* program have been included. Other doses for the *Dp*Cel7B damage series (not shown) exhibited similar behaviour.

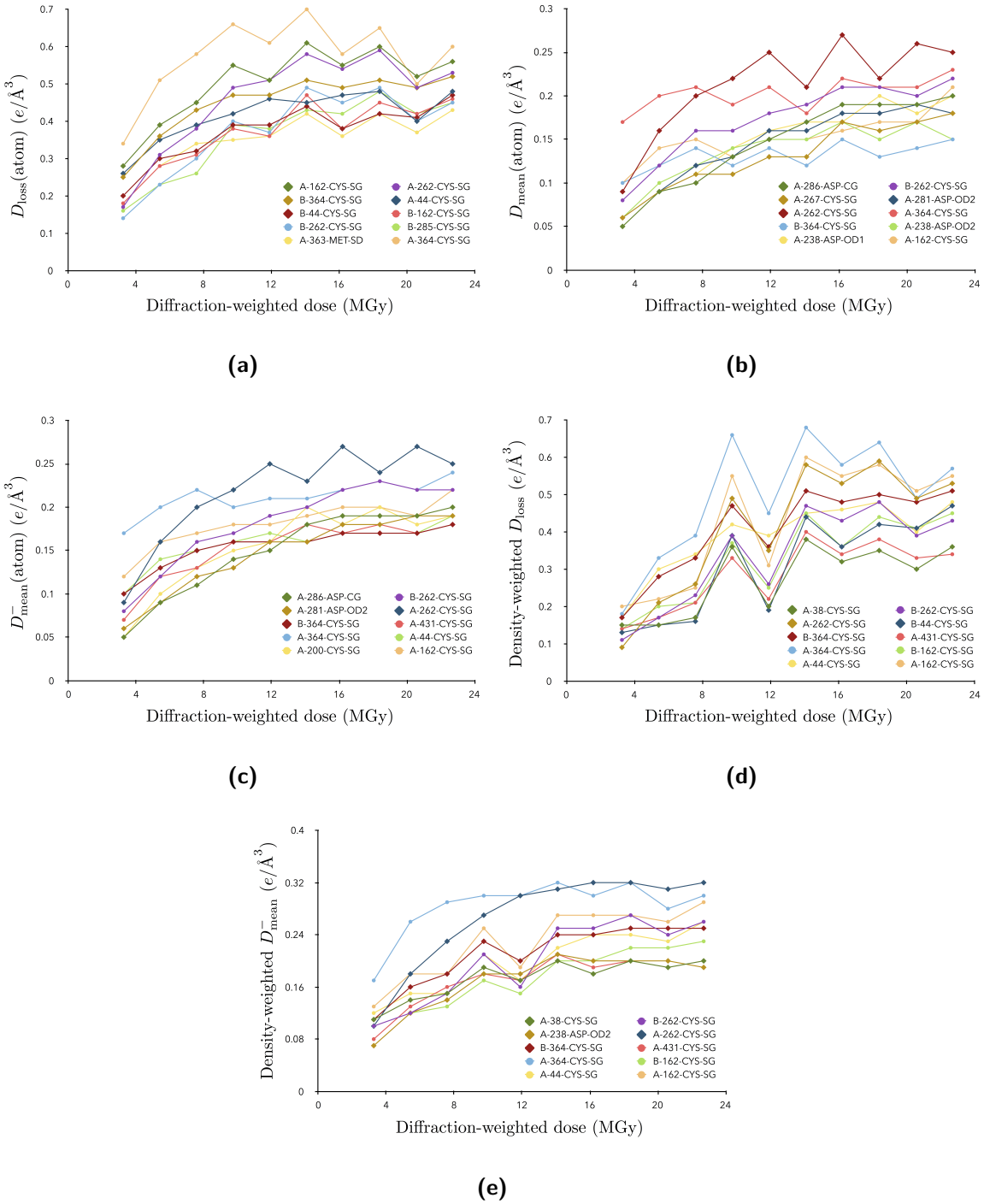
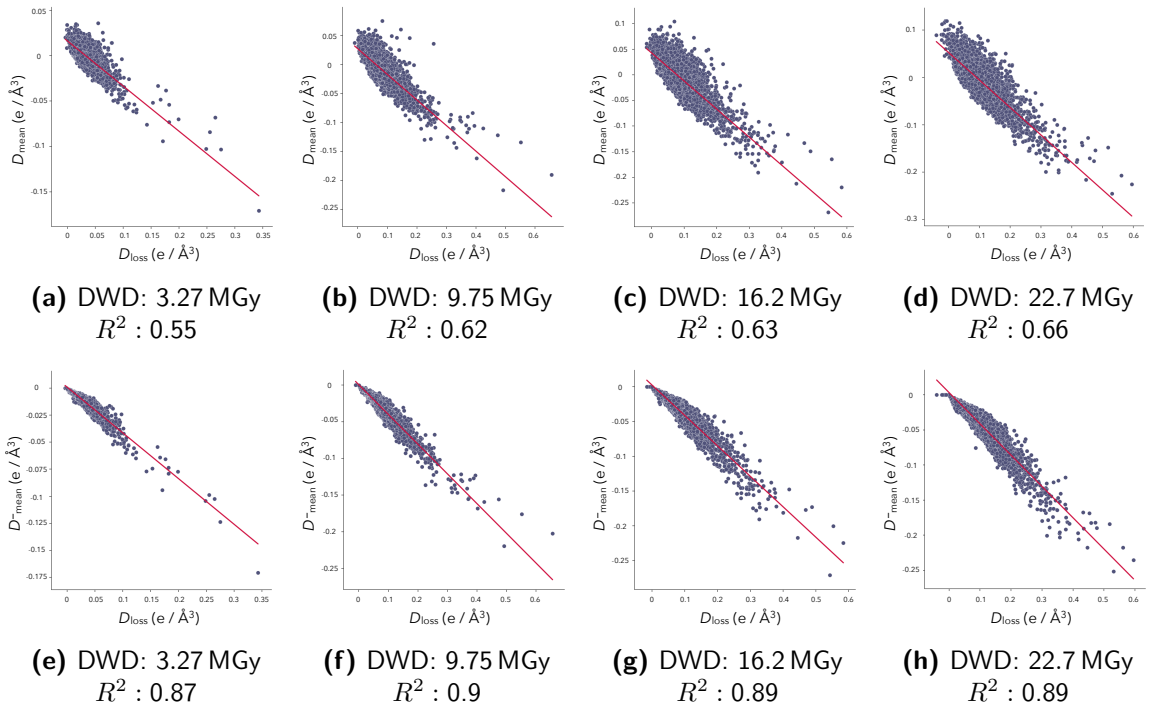


Figure 2.24: For the *Dp*Cel7B protein case study, *RIDL* has determined the top 10 damaged atoms with respect to different summary metrics and these are plotted as a function of dose: (a) D_{loss}, (b) D_{mean}, (c) D^{-}_{mean}, (d) density-weighted D_{loss} and (e) density-weighted D^{-}_{mean}. The top 10 damage sites predicted by each metric differ, however, for all presented metrics, these sites are known to be radiation-sensitive (Cys, Met and Asp).



(a) DWD: 3.27 MGy R^2 : 0.55 **(b)** DWD: 9.75 MGy R^2 : 0.62 **(c)** DWD: 16.2 MGy R^2 : 0.63 **(d)** DWD: 22.7 MGy R^2 : 0.66

(e) DWD: 3.27 MGy R^2 : 0.87 **(f)** DWD: 9.75 MGy R^2 : 0.9 **(g)** DWD: 16.2 MGy R^2 : 0.89 **(h)** DWD: 22.7 MGy R^2 : 0.89

Figure 2.25: For the *Dp*Cel7B protein case study, metric values have been compared for all atoms present in the *Dp*Cel7B coordinate model. In (a) - (d) D_{loss} and D_{mean} have been compared at 3.27, 9.75, 16.2 and 22.7 MGy. In (e) - (h) D_{loss} and D^{-}_{mean} have been compared at the same doses. Other doses (data not shown) exhibit similar behaviour. For each plot, the coefficient of determination, R^2 has been computed using the `SciPy` *Python* package.

A schematic illustrating this weighting scheme has been presented in Figure 2.26.

In practice this was achieved in *RIDL* by additionally using *FFT* to generate a ρ_{calc} map, given as:

$$\rho_{\text{calc}}(\mathbf{x}) = \frac{1}{V}\sum_{hkl} |\mathbf{F}_{\text{calc},1}| \cdot \exp(i\phi_{\text{calc},1}) \cdot \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}) \tag{2.5.7}$$

where V is the volume of the unit cell, and $|\mathbf{F}_{\text{calc},1}|$ and $\phi_{\text{calc},1}$ are the calculated structure factor amplitude and phase set derived from the coordinate model corresponding to the first dataset. As illustrated previously for the example input file in Figure 2.17, in the current implementation of *RIDL*, this information is input using the `phaseLabel` and `FcalcLabel` corresponding to the `mtz3` structure factor file.

For each higher dose dataset n, to ensure that the ρ_{calc} map was compatible with the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map generated in Sections 2.5.4, the grid sampling, orientation and extent was fixed to that of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map using *FFT*. In particular, at each dataset n, the ρ_{calc} map was re-calculated over the unit cell dimensions corresponding to dataset n, to exactly match those of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map.

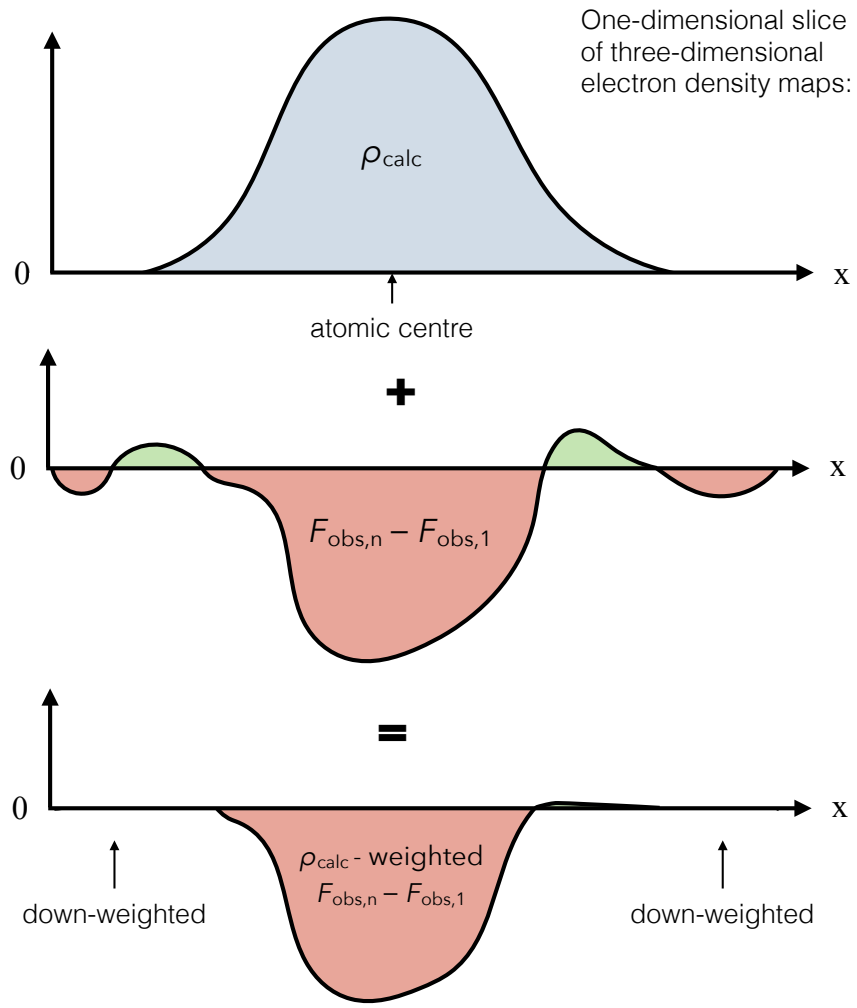


Figure 2.26: Schematic to illustrate the effect of weighting each $F_{obs,n} - F_{obs,1}$ map voxel by the magnitude of the ρ_{calc} density map at each voxel. For clarity, only a one-dimensional slice through three-dimensional Cartesian space is shown. The ρ_{calc} map density (in blue) is highest at positions of well defined atoms built into the coordinate model, and diminishes at distances further from each atomic centre. The $F_{obs,n} - F_{obs,1}$ map takes both positive and negative values (in green/red), however by weighting each voxel value by the the corresponding ρ_{calc} map voxel at that position, $F_{obs,n} - F_{obs,1}$ map values located further from an atomic centre are decreased relative to those situated close to an atom. Note that the '+' operator is not a simple summation of the ρ_{calc} and $F_{obs,n} - F_{obs,1}$ map, but is defined by Equation 2.5.8.

For a given atom each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxel in the set V_{atom} could then be weighted by the corresponding ρ_{calc} map density at that voxel:

$$\text{Density-weighted } D_{\text{loss}} = \frac{\max\limits_{v \in V_{\text{atom}}} -\rho_{\Delta}(v) \cdot \rho_{\text{calc}}(v)}{\max\limits_{v \in V_{\text{atom}}} \rho_{\text{calc}}(v)} \tag{2.5.8}$$

where $\rho_{\Delta}(v)$ has been used to denote the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map value at voxel $v \in V_{\text{atom}}$. This construction (and similarly the construction in Equation 2.5.9 below) resulted in metrics that also had units of e/Å^3.

Figure 2.27a-d illustrates that the density-weighting scheme most significantly affects atoms attaining low D_{loss} values, where the density-weighted D_{loss} values are lower. It is suggested that density-weighting has suppressed noisy background in the vicinity of these atoms. The ordering of top damage sites is conserved between the density-weighted and non-weighted D_{loss} metrics. This is reasonable since for the most pronounced protein damage sites (disulphide bonds), the negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ peak should align directly with each Cys-S_{γ} centre, and density-weighting should have a negligible effect. In Figure 2.27a, a weaker correlation is observed overall than at higher doses; this observation is consistent with the yield of site-specific damage being lower at 3.27 MGy, such that fewer $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks corresponding to true damage events are coincident with atomic centres. As such, a plot of D_{loss} versus density-weighted D_{loss}, as in Figure 2.27a-d, may be useful as a tool to assess the damage signal/noise ratio in a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map at a particular dose.

A density-weighting can also be applied to minimise the search-radius dependence of the average-based metric D^{-}_{mean}:

$$\text{Density-weighted } D^{-}_{\text{mean}} = \frac{\sum\limits_{v \in V^{-}_{atom}} -\rho_{\Delta}(v) \cdot \rho_{calc}(v)}{\sum\limits_{v \in V^{-}_{atom}} \rho_{calc}(v)} \tag{2.5.9}$$

where V^{-}_{atom} is the set of voxels $v \in V_{\text{atom}}$ such that $\rho_{\Delta}(v) < 0$.

Figure 2.28a indicates how the various metrics proposed in this section differ for selected atoms of the *Dp*Cel7B protein case study introduced in Section 2.3. Here, atoms have been arbitrarily selected to represent a range of possible scenarios in the $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference map (Fig-

ure 2.28b-i). The D_{loss} and density-weighted D_{loss} typically agree very well, and both attain the highest metric values overall. For cases where these two metrics differ, (e.g. Ile-35-C_α, Tyr-153-O_H and Asp-238-C_γ) manual inspection indicates that the most negative difference density in the vicinity is not aligned with the atomic centre (Figure 2.28 b, d and e). Due to the construction of D^{-}_{mean}, it is unsurprising that both the D^{-}_{mean} and density-weighted D^{-}_{mean} metrics attain lower values than D_{loss}. Again, the density-weighted and non-weighted D^{-}_{mean} metrics show strong agreement for these selected atoms (and moreover, for the model as a whole, as seen in Figure 2.27e-h). Minor differences do exist (e.g. for Cys-262-S_γ), as expected, due to the weighting scheme in Equation 2.5.9, however the effect of the weighting is more difficult to predict by direct inspection of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map (e.g. Figure 2.28f).

It is crucial to note that the inclusion of a density-weighting scheme will also increase the dependence of the metric on the quality of the coordinate model. This is clear from Equations 2.5.9 and 2.5.8 due to the direct dependence on the ρ_{calc} map. Errors in the model, such as the poor positioning of an atom, the wrong element identity or inaccuracies in either the *B*-factor or occupancy of the atom, will propagate directly to errors in ρ_{calc}. As such, density-weighted metrics should only be trusted when the model is known to be satisfactorily refined.

It is further noted that in the current implementation of *RIDL* the calculated ρ_{calc} maps correspond to the coordinate model for dataset 1. No attempt has been presented to generate ρ_{calc} maps which depend on the refinement of the model at successive higher dose datasets. As an example, the decision on whether to refine atomic *B*-factor values per-dataset would directly impact the local behaviour of the ρ_{calc} map around atoms. To entirely mitigate this variability, the ρ_{calc} map corresponds only to dataset 1 in the current analysis.

2.5.8 The C_α normalisation scheme

A challenge remains as to whether the metrics proposed above can be directly compared between independent MX damage series, particularly if collected on different proteins, at different doses and to different high resolution cut-offs. Furthermore, even for non-radiation sensitive atoms, the metrics designed to quantify electron density loss (e.g. D_{loss}, D^{-}_{mean}) are observed to increase with dose. Figure 2.29a demonstrates how the distribution of D_{loss} values attained for the *Dp*Cel7B

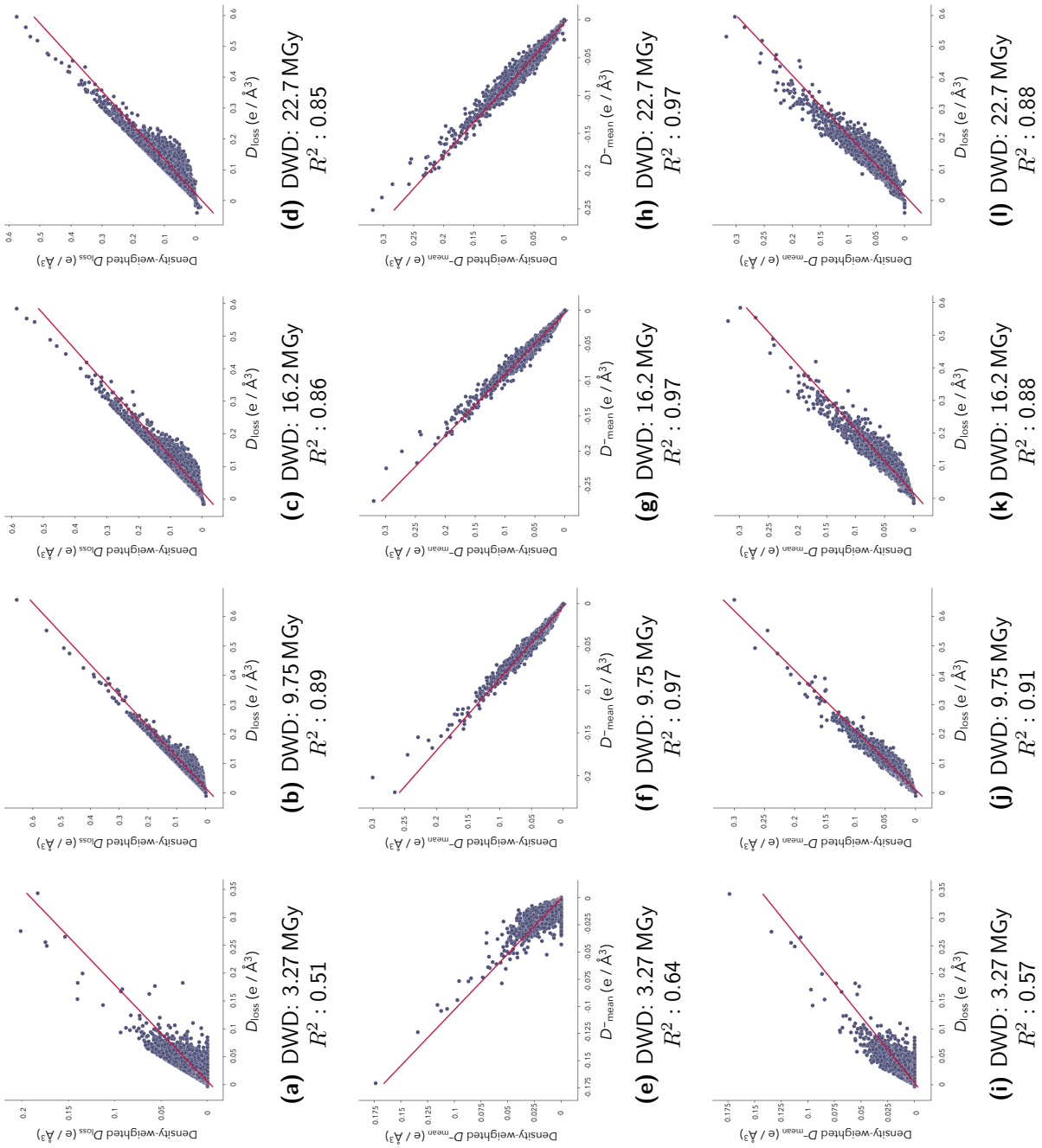
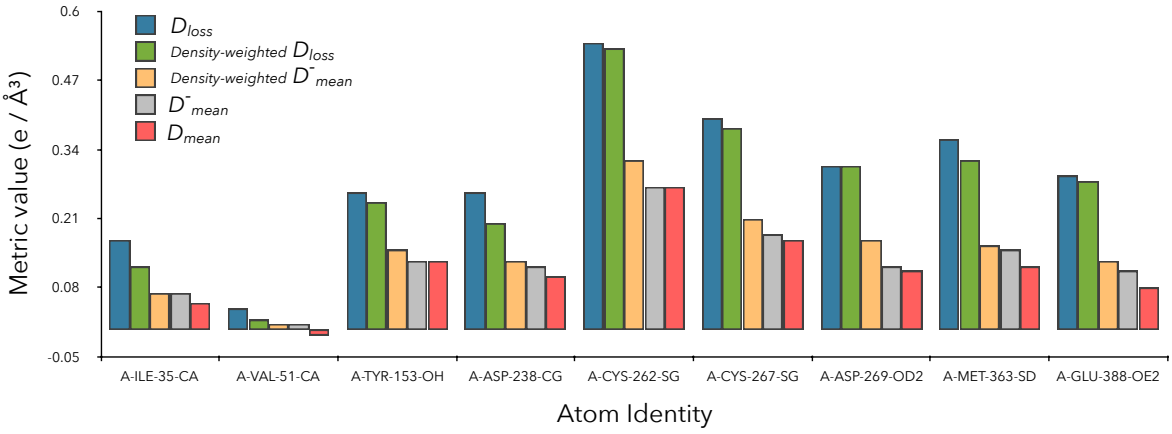
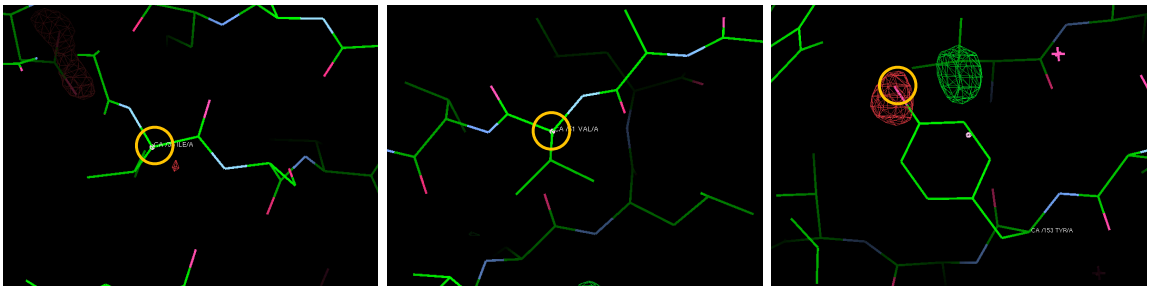


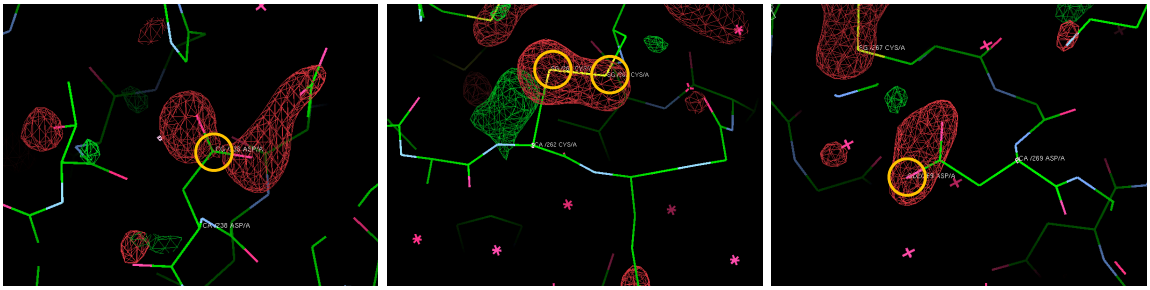
Figure 2.27: For the *Dp*Cel7B protein case study, metric values have been compared for all atoms present in the *Dp*Cel7B coordinate model. In (a) - (d) D_{loss} and density-weighted D_{loss} have been compared at 3.27, 9.75, 16.2 and 22.7 MGy. In (e) - (h) D^-_{mean} and density-weighted D^-_{mean} have been compared at the same doses. In (i) - (l) D_{loss} and density-weighted D^-_{mean} have been compared. Other doses (data not shown) exhibit similar behaviour. For each plot, the coefficient of determination, R^2 has been computed using the `SciPy` *Python* package.



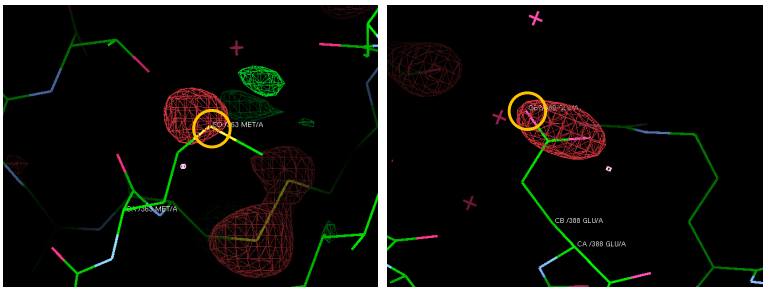
(a)



(b) Ile-35-C_α **(c)** Val-51-C_α **(d)** Tyr-153-O_H



(e) Asp-238-C_γ **(f)** Cys-262-S_γ & Cys-267-S_γ **(g)** Asp-269-$O_{\delta 2}$



(h) Met-363-S_δ **(i)** Glu-388-$O_{\epsilon 2}$

Figure 2.28: For the *Dp*Cel7B protein case study, proposed *RIDL* metrics have been compared for several selected atoms in the protein. Here, metric values have been derived from the $F_{obs,8} - F_{obs,1}$ map only. The atoms were selected in order to represent a range of scenarios, for which either clear difference map peaks attributable to damage are present, or unconvincing difference density is present, as deemed by manual inspection of the $F_{obs,8} - F_{obs,1}$. In (b - i), the selected atoms are indicated by yellow circles and the $F_{obs,8} - F_{obs,1}$ map is contoured at $\pm 4\sigma$ in green/red. Non-rendered screenshots have been taken directly from *COOT* to demonstrate exactly what a user performing manual inspection would observe.

test protein shifts to larger values with increasing dose. This is due to the inevitable increase in $F_{\text{obs,n}} - F_{\text{obs,1}}$ map noise with increasing dose, to which the D_{loss} metric is sensitive. Ultimately, the raw D_{loss}(atom) value of an atom alone is insufficient to determine whether an atom has experienced significant electron density loss, without reference to the rest of the structure.

This section aims to present suitable normalisation schemes, where metric values are compared relative to the reference set of C_α atoms of the protein backbone. The C_α atom set was chosen since it widely considered not to be radiation sensitive at the MGy doses typically reported for MX data collection at 100 K. Two candidate normalisations will be presented here for the case of the D_{loss} metric, and the validity of each discussed. By design, both normalisation schemes (Equation 2.5.10 and 2.5.14) preserve the ordering of damage to atoms as given by the standard D_{loss} metric. For other metrics defined above (e.g. D^{-}_{mean}), the normalisation scheme can be similarly defined.

2.5.8.1 Normalisation scheme 1

For the D_{loss} metric, a C_α-normalised D_{loss}, ${}^{1}D^{C_\alpha}_{\text{loss}}$, can be defined as follows:

$$ {}^{1}D^{C_\alpha}_{\text{loss}}(\text{atom,dose}) = \frac{D_{\text{loss}}(\text{atom,dose}) - \langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in C_\alpha}}{\left|\langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in C_\alpha}\right|} \tag{2.5.10} $$

where $\langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha}$ represents the average D_{loss} value attained by the set of C_α atoms. Unlike D_{loss}, ${}^{1}D^{\text{C}_\alpha}_{\text{loss}}$(atom,dose) is unit-less. Equation 2.5.10 was designed such that the property $\langle {}^{1}D^{\text{C}_\alpha}_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha} = 0$ strictly holds at any dose (i.e. the average metric value attained by the C_α backbone is zero).

The results shown in Figure 2.29b demonstrate that following C_α-normalisation, the majority of atoms in the *Dp*Cel7B protein structure attain a metric value comparable to the C_α backbone at each dose, with ${}^{1}D^{\text{C}_\alpha}_{\text{loss}}(\text{atom,dose}) \approx 0$. Sites of specific damage are located in the skewed right-hand tail of the distribution. A dose-independent criterion can now be established regarding whether an atom is damaged: at any dose, a positive ${}^{1}D^{\text{C}_\alpha}_{\text{loss}}$(atom,dose) value indicates that electron density loss to the atom exceeds that of the C_α-backbone.

Assuming that $\langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha} > 0$, as is the case for the *Dp*Cel7B protein in Figure 2.29b, ${}^1D^{\text{C}_\alpha}_{\text{loss}}(\text{atom,dose}) = n$ indicates that an atom exhibits an $(n+1)$-fold increase in electron density loss, relative to the protein backbone:

$$\frac{D_{\text{loss}}(\text{atom,dose}) - \langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha}}{\langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha}} = n \tag{2.5.11}$$

$$\iff D_{\text{loss}}(\text{atom,dose}) = (n+1)\cdot\langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha} \tag{2.5.12}$$

A benefit of the proposed normalisation is that a dose-independent threshold (e.g. ${}^1D^{\text{C}_\alpha}_{\text{loss}}(\text{atom,dose}) = 1$) could therefore be specified, beyond which an atom is deemed to be damaged.

However, it must be noted that Equation 2.5.10 does not constrain the standard deviation of ${}^1D^{\text{C}_\alpha}_{\text{loss}}(\text{atom,dose})$ values attained by C_α atoms:

$$\sigma_{a\in \text{C}_\alpha}[{}^1D^{\text{C}_\alpha}_{\text{loss}}(\text{a,dose})] = \frac{\sigma_{a\in \text{C}_\alpha}[D_{\text{loss}}(\text{a,dose})]}{\langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha}}, \tag{2.5.13}$$

It is thus unclear how Equation 2.5.13 will behave in general as a function of dose.

2.5.8.2 Normalisation scheme 2

The normalisation scheme presented in Equation 2.5.10 can be modified to:

$${}^2D^{\text{C}_\alpha}_{\text{loss}}(\text{atom,dose}) = \frac{D_{\text{loss}}(\text{atom,dose}) - \langle D_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha}}{\sigma_{a\in \text{C}_\alpha}[D_{\text{loss}}(\text{a,dose})]} \tag{2.5.14}$$

Equation 2.5.14 is strictly a Z-score, as it is the number of signed standard deviations, $\sigma_{a\in \text{C}_\alpha}[D_{\text{loss}}(\text{a,dose})]$, that an atom's D_{loss} value deviates from the mean value attained by the population of C_α atoms.

In addition to the property $\langle {}^2D^{\text{C}_\alpha}_{\text{loss}}(\text{a,dose})\rangle_{a\in \text{C}_\alpha} = 0$ holding, it is also true that $\sigma_{a\in \text{C}_\alpha}[{}^2D^{\text{C}_\alpha}_{\text{loss}}(\text{a,dose})] = 1$ must hold at any dose. As such, Equation 2.5.14 is valid under the assumption that both the mean and standard deviation of D_{loss} values attained by C_α atoms should be dose-independent.

Figure 2.29c demonstrates the effect of applying Equation 2.5.14 to the *Dp*Cel7B test protein. Similarly to Figure 2.29, at each dose, the metric values exhibit a skewed unimodal distribution with mode ~ 0. Compared to the plot in Figure 2.29b, for the majority of atoms (with $|^2D^{C_\alpha}_{\text{loss}}| \leq 5$) there appears to be greater overlap between distributions at different doses (this is most notable by comparing the 3.27 MGy dataset (blue) between Figure 2.29b and c).

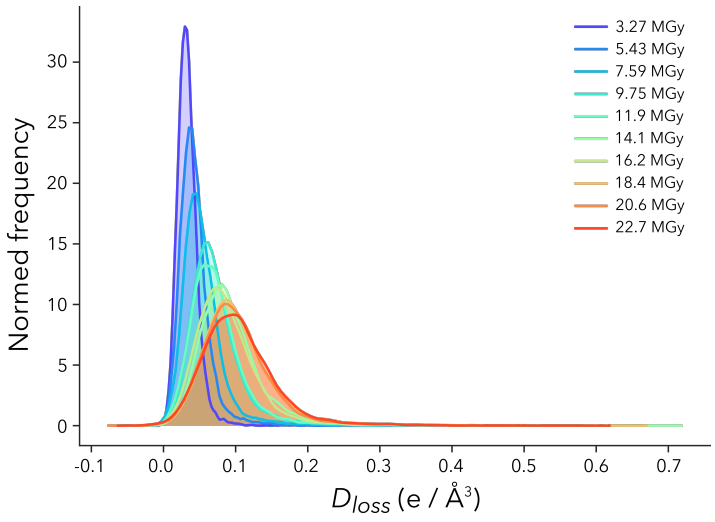
2.5.8.3 Suitability of each normalisation scheme

As shown in Figure 2.30, for both normalisation schemes in Equations 2.5.10 and 2.5.14, the ordering of damaged atom sites (as ranked by D_{loss}) is unaffected by C_α-normalisation at all doses for the *Dp*Cel7B protein test case. Both the $^1D^{C_\alpha}_{\text{loss}}$(atom,dose) and $^2D^{C_\alpha}_{\text{loss}}$(atom,dose) metrics attain consistently lower values at higher doses. This is counter-intuitive, since it should be expected that the yield of damage (e.g. disulphide bond cleavage events) across all units cells in the crystal should increase with dose (assuming a disulphide repair mechanism is lacking). This observation can be explained by the background noise in the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map also increasing with dose, with the results in Figure 2.30 providing a direct indication that the ratio of damage signal to noise in the $F_{\text{obs,n}} - F_{\text{obs,1}}$ is diminishing with increasing dose.

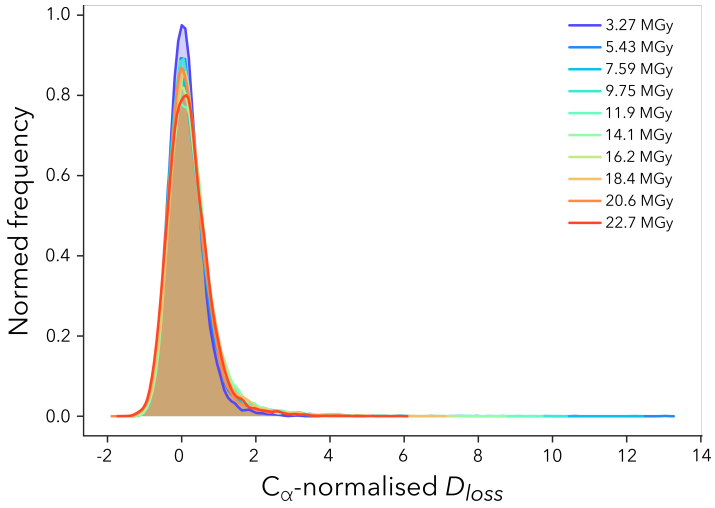
In Section 2.5.9.1 below, it will be demonstrated that the second scheme (Equation 2.5.14) appears to be more stable to varying $F_{\text{obs,n}} - F_{\text{obs,1}}$ map resolution. Consequently, that scheme will be used as the default normalisation unless otherwise specified ($D^{C_\alpha}_{\text{loss}} := {}^2D^{C_\alpha}_{\text{loss}}$).

2.5.9 Testing of the *RIDL* pipeline

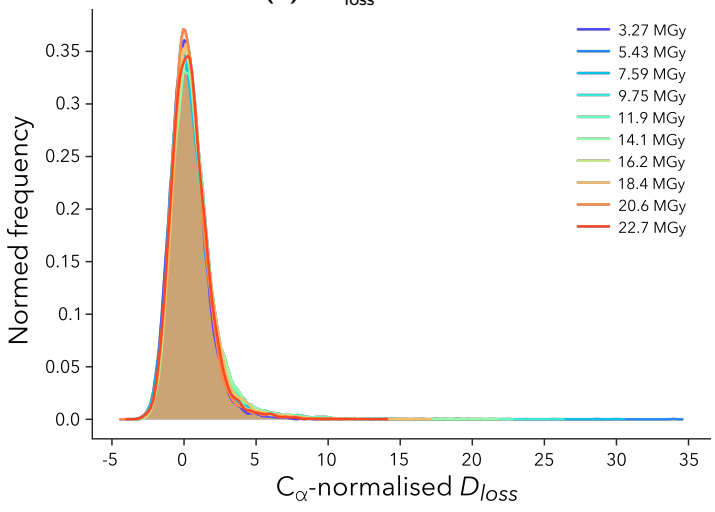
The purpose of this section is to report tests on the robustness of *RIDL* to variations in the quality of input data. Several variables have already been tested in the sections above, including choice of α parameter in *SFALL* (Section 2.5.5, Figure 2.22), and structure factor amplitude scaling protocol in *SCALEIT* (Section 2.5.4, Figure 2.19). These will not be repeated here.



(a) Standard D_{loss} metric



(b) ${}^1D^{C_\alpha}_{\text{loss}}$ metric



(c) ${}^2D^{C_\alpha}_{\text{loss}}$ metric

Figure 2.29: For the *Dp*Cel7B protein damage series, the distribution of (a) D_{loss}, (b) ${}^1D^{C_\alpha}_{\text{loss}}$(atom,dose) metric (defined by Equation 2.5.10), and (c) ${}^2D^{C_\alpha}_{\text{loss}}$(atom,dose) metric (defined by Equation 2.5.14) metric values attained by all (7300) refined atoms in the coordinate model have been shown for each higher dose dataset in the series. The distributions have been presented as kernel density estimation (KDE) plots, i.e. as non-parametric estimates of the probability density function underlying each distribution. Such KDE plots can be regarded here as smoothed fits to each histogram, and present an efficient way to overlap multiple sets of data on a single set of axes.

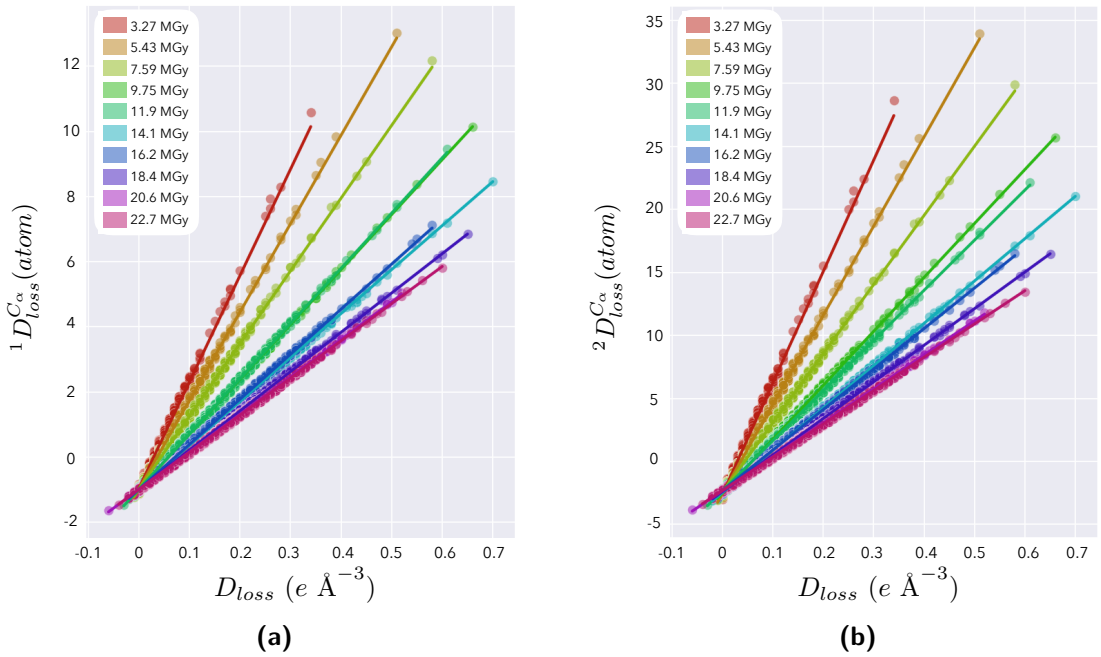


Figure 2.30: For each higher dose in the *Dp*Cel7B protein damage series, the non-normalised D_{loss} metric over all atoms has been compared with (a) the $^1D^{C_\alpha}_{\text{loss}}$(atom,dose) metric (defined by Equation 2.5.10), and (b) the $^2D^{C_\alpha}_{\text{loss}}$(atom,dose) metric (defined by Equation 2.5.14). Metric values have been derived for each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map, with the dose for each high dose dataset n provided in each legend.

2.5.9.1 The effect of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map resolution

It has been widely reported that diffraction resolution diminishes as a function of increasing dose. The maximum resolution to which a crystal diffracts ultimately dictates the resolution of the electron density map that can be achieved. In order to compare *RIDL* metric values between different resolution datasets, whether within a single damage series or between independent crystals, it is essential to be able to predict the effect of resolution.

For the *Dp*Cel7B protein damage series, *RIDL* was run repeatedly, with the maximum permitted resolution of each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map fixed using *FFT* in increasing 0.5 Å increments between 2 Å and 4 Å. As illustrated in Figure 2.31a, the D_{loss} metric appears to be highly dependent on $F_{\text{obs,n}} - F_{\text{obs,1}}$ map resolution, with a greater reduction in map resolution appearing to lead to lower metric values. This effect was observed across the wide dose range of the *Dp*Cel7B series (3.27 – 22.7 MGy).

The resolution-dependence of a metric can be indicated by the extent of deviation of individual points from the diagonal line ($y = x$) in Figure 2.31. C_α-normalisation of the D_{loss} metric

has been shown to partially compensate for the explicit dependence of D_{loss} on resolution. In Figure 2.31b, the normalisation appears to overcompensate, particularly for atom sites attaining high D_{loss}(atom) values. As such, the C_α-normalisation method proposed in Equation 2.5.10 appears unsuitable.

Conversely, Figure 2.31c indicates that the second normalisation scheme (Equation 2.5.14) can substantially remove the explicit dependence of D_{loss} on resolution. Overall, this result implies that if metric values are to be directly compared between independently collected damage series, C_α-normalisation using Equation 2.5.14 must first be performed.

2.5.9.2 The unit cell dimensions used for $F_{\text{obs,n}} - F_{\text{obs,1}}$ map calculations

In Section 2.5.4 above, it has been stated that in the current protocol for *RIDL*, for a given dataset n, a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map is generated over the unit cell dimensions corresponding to the dataset n coordinate model. As such, the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map extent will vary for different n, due to unit cell expansion with increasing dose. Indeed, for the case of *Dp*Cel7B, the unit cell dimensions clearly fluctuate with dose, as shown in Table 2.2 of Section 2.3.

To determine the effect of unit cell variation between datasets on the output of *RIDL*, two strategies have been directly compared:

1. The default case in which each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map (and corresponding atom-map in Section 2.5.5) are generated over the unit cell dimensions for dataset n. Consistent with the discussion in Section 2.5.3.1, rigid-body refinement has been employed for each dataset n to ensure the n^{th} coordinate model unit cell dimensions (and thus the dimensions of the atom-map) match those of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map.

2. The alternative case in which all $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps are generated over the unit cell dimensions for dataset 1. In this case, a single atom-map is derived from the dataset 1 coordinate model and can then be used to extract per-atom data from all $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps.

It is noted that in the first strategy, no isotropic *B*-factor refinement has been incorporated when

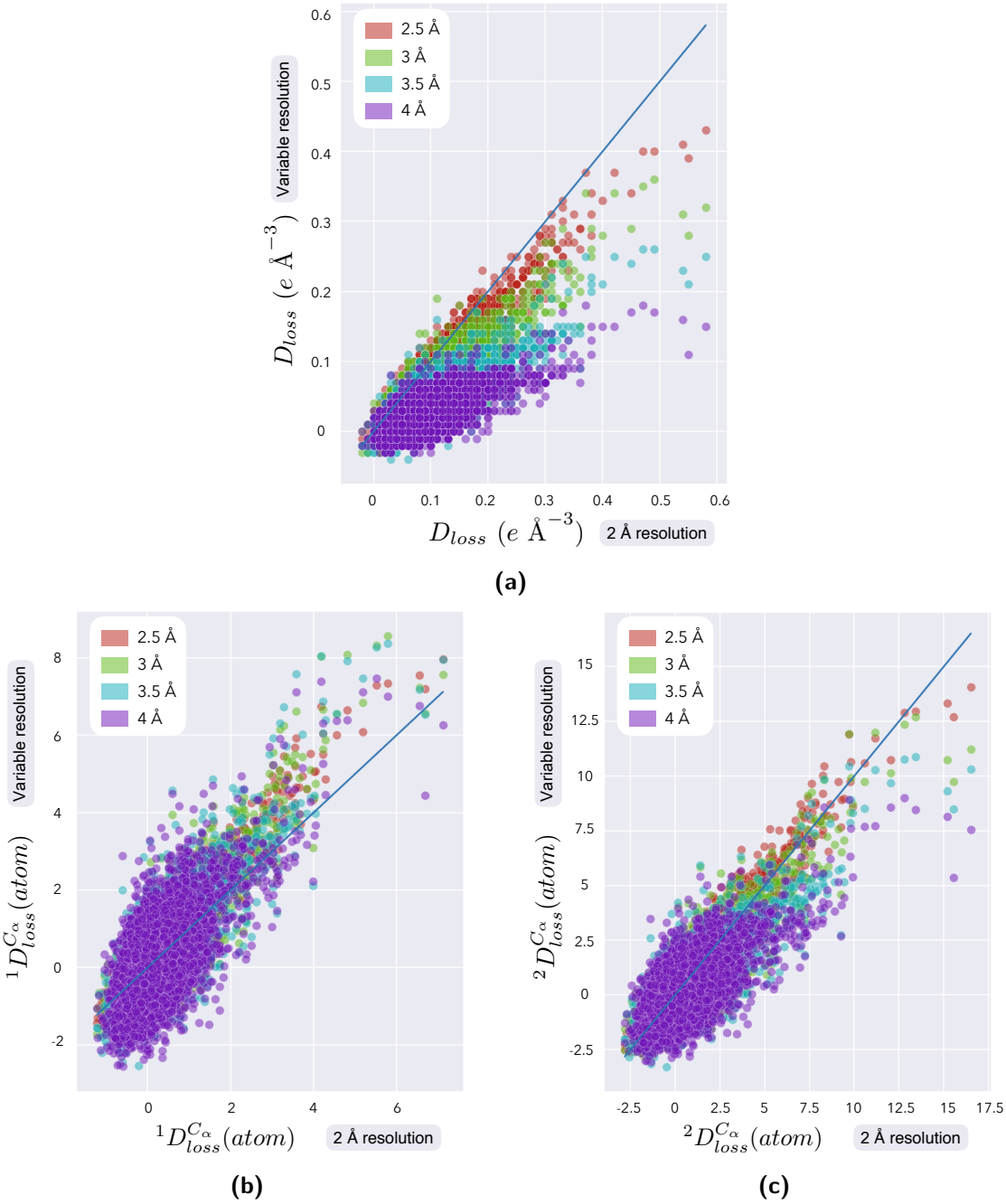


Figure 2.31: The dependence of the C_α-normalised D_{loss} metric on the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map resolution. For the example case of the *Dp*Cel7B GH7 family protein, the effect of maximum $F_{\text{obs,n}} - F_{\text{obs,1}}$ map resolution on *RIDL* metrics is shown for: (a) the standard D_{loss} metric, and C_α-normalised D_{loss} metric calculated by (b) Equation2.5.10 and (c) Equation 2.5.14. Here, metric values have been derived from the $F_{\text{obs,5}} - F_{\text{obs,1}}$ map only, with higher dose dataset corresponding to 9.75 MGy. The maximum map resolution has been varied in 0.5 Å increments between 2 and 4 Å. Each x-axis indicates the D_{loss}(atom) metric values across all atoms in *Dp*Cel7B at 2 Å. The corresponding y-axes indicate metric values at each tested lower resolution cut-off. Deviations of points from the line $y = x$ (in blue) illustrate the resolution-dependence of a metric. Other doses for the *Dp*Cel7B damage series (not shown) exhibited similar behaviour.

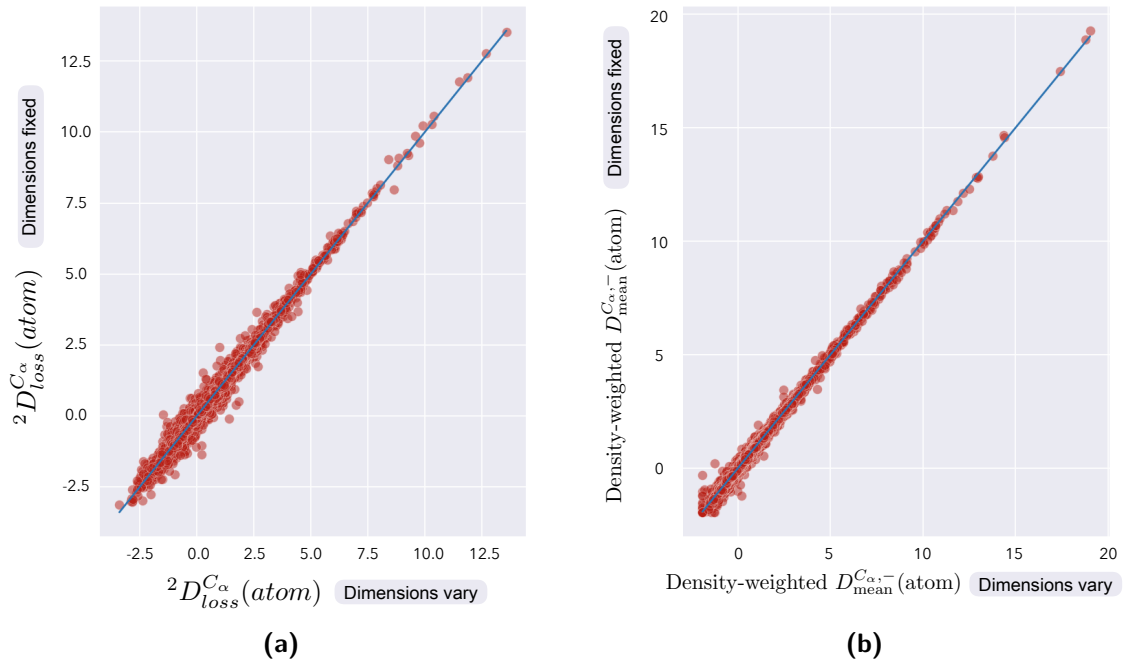


Figure 2.32: The dependence of *RIDL* metrics on the unit cell dimensions used to calculate each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. For the example case of the *Dp*Cel7B GH7 family protein, the (a) C_α-normalised D_{loss}(atom) metric ($^2D^{C_\alpha}_{\text{loss}}$, Equation 2.5.14, Section 2.5.8) and (b) density-weighted & C_α-normalised D^{-}_{mean} have been calculated with *RIDL*. In each case, metric values have been derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps with higher dose dataset n corresponding to 22.7 MGy ($n = 11$). Each x-axis indicates the metric values across all atoms in *Dp*Cel7B such that the $F_{\text{obs,11}} - F_{\text{obs,1}}$ map has been generated over the unit cell dimensions corresponding to dataset 11. The corresponding y-axes indicate metric values for the case where the $F_{\text{obs,11}} - F_{\text{obs,1}}$ map has been generated over the unit cell dimensions corresponding to dataset 1. Deviations of points from the line $y = x$ (in blue) indicate atoms for which the metric is sensitive to the choice of unit cell dimensions. Other doses for the *Dp*Cel7B damage series (not shown) exhibited similar behaviour, and the depicted case ($n = 11$) represents the most extreme scenario available in which the unit cell dimensions differ the most.

refining the dataset 1 coordinate model against each higher dose set of structure factor amplitudes $|\mathbf{F}_{\text{obs,n}}|$, in contrast to the strategy stated in Section 2.5.3.2. This was to ensure that the changing unit cell dimensions was the only variable between the two strategies here, since in strategy 2, *B*-factor values could not be refined at each dose without updating the unit cell dimensions of the coordinate model.

The data in Figure 2.32 indicates that the *Dp*Cel7B test case was highly insensitive to the choice of unit cell dimensions used by *RIDL*. Furthermore, in this figure the output of *RIDL* have been compared using unit cell dimensions for dataset $n = 11$ and dataset 1, which represent the most significant deviation in unit cell dimensions as reported in Table 2.2.

2.5.9.3 The choice of model refinement strategy

In light of Section 2.5.9.2, it is also essential to determine whether variations in the refinement strategy for each higher dose dataset n could lead to deviations in the output of *RIDL*. Indeed, in Section 2.5.3 above, it was suggested that independent restrained refinement of coordinate models corresponding to higher dose datasets may lead to errors due to non-conserved sampling of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density around each refined atom at each dose.

To test the significance of this error, the output of *RIDL* was compared for the *Dp*Cel7B test series, with the higher dose dataset coordinate models refined using four distinct strategies:

1. Rigid-body and isotropic *B*-factor refinement (3 cycles) in *phenix.refine*. Here each of the two chains (A and B) were considered to be a single fixed rigid body. This is the default refinement strategy used for the *Dp*Cel7B test protein presented in this chapter.
2. Rigid-body, isotropic *B*-factor and TLS refinement (3 cycles) in *phenix.refine*.
3. 10 cycles of restrained refinement in *REFMAC*.
4. rigid body refinement (10 cycles) in *REFMAC*.

In all cases, refinement used the refined coordinate model from the first dataset coupled with structure factor amplitudes, $|\mathbf{F}_{\text{obs,n}}|$, for the higher dose dataset $n > 1$.

RIDL was then run for each strategy, using the newly generated higher dose coordinate models to define the location of each atom within each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. It is noted, however, that the actual calculation of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps is not influenced directly by the updated coordinate models, since the calculated phase set, $\phi_{\text{calc,1}}$, corresponding to dataset 1 has been used throughout here. The data in Figure 2.33 compares the metrics from *RIDL* generated using Strategy 1 (the default in the current chapter) with the additional strategies.

The data in Figure 2.33a-c indicate that for the *Dp*Cel7B test case, the $D_{\text{loss}}^{\text{C}_\alpha}$ metric was highly robust across the tested refinement procedures. This trend was also observed for the non-normalised D_{loss} metric (data not shown). In particular, sites of damage consistently attained high $D_{\text{loss}}^{\text{C}_\alpha}$ metric values regardless of refinement strategy. The lack of correlation for atoms attaining low $D_{\text{loss}}^{\text{C}_\alpha}$

values ($D_{\text{loss}}^{\text{C}_\alpha}$<5) was unsurprising, since these atoms are susceptible to the variable inclusion of background $F_{\text{obs,n}} - F_{\text{obs,1}}$ map noise between strategies. However, these atoms typically still attain low metric values between refinement strategies.

For the $D_{\text{loss}}^{\text{C}_\alpha}$ metric, deviations from the diagonal line $y = x$ (in blue) are most significant for the rigid-body refinement strategy (in red in Figure 2.33a-c), as compared to the default strategy (Strategy 1 above). This was expected, since the per-atom search radius defined by Equation 2.5.4 in Section 2.5.5 is *B*-factor-dependent, which are not updated during the rigid-body refinement strategy. Overall, this discrepancy is less significant at higher doses (Figure 2.33b-c), once significant damage is present in each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map, and furthermore this does not disrupt the detection of top damage sites in *Dp*Cel7B.

The data in Figure 2.33d-f suggest that the density-weighted C_α-normalised D_{mean}^{-} was more dependent on the selected refinement strategy, and more so at higher doses (Figure 2.33f). A comparable trend was also observed for the non-normalised metric (data not shown). Most notably, the significant deviation of the rigid-body refinement data (in red) from the diagonal $y = x$ line (in blue) indicates that the decision to use isotropic *B*-factor-refinement (as has been used by default for the *Dp*Cel7B example case presented in this chapter) plays a significant role in the resulting density-weighted D_{mean}^{-} metric values. For the top damage sites in *Dp*Cel7B, it is unsurprising that the rigid-body refinement strategy results in higher metric values than when isotropic *B*-factor refinement is included (each x-axis), since this has the effect of shrinking the volume over which the density-weighted mean is commuted to closer to the centre of the atom.

The decision on whether *B*-factor refinement is appropriate for higher dose datasets is non-trivial; whereas omitting *B*-factor refinement leads to no variation in the search radius of a given atom between datasets, the region of density to which the atom contributes will be underestimated at higher doses if *B*-factor increase is unaccounted for. Fortunately, as illustrated in Figure 2.33d-f, regardless of the chosen refinement strategy, density-weighted D_{mean}^{-} values appear to attain approximately the same ordering of metric values across all atoms. Future development of *RIDL* will aim to determine an optimum refinement approach for such density-weighting schemes.

In the current thesis, such density-weighted metrics have not featured prominently in the remaining chapters. Where presented, due to the slight discrepancy in values between refinement strategies,

it has been ensured that all corresponding models for higher dose datasets have undergone rounds of rigid-body and isotropic *B*-factor refinement (Strategy 1 in this section), to be consistent with the default strategy for the *Dp*Cel7B test case. It is further noted that no conclusions in this thesis have been based solely on the results of such density-weighted metrics.

2.5.9.4 The choice of phase set

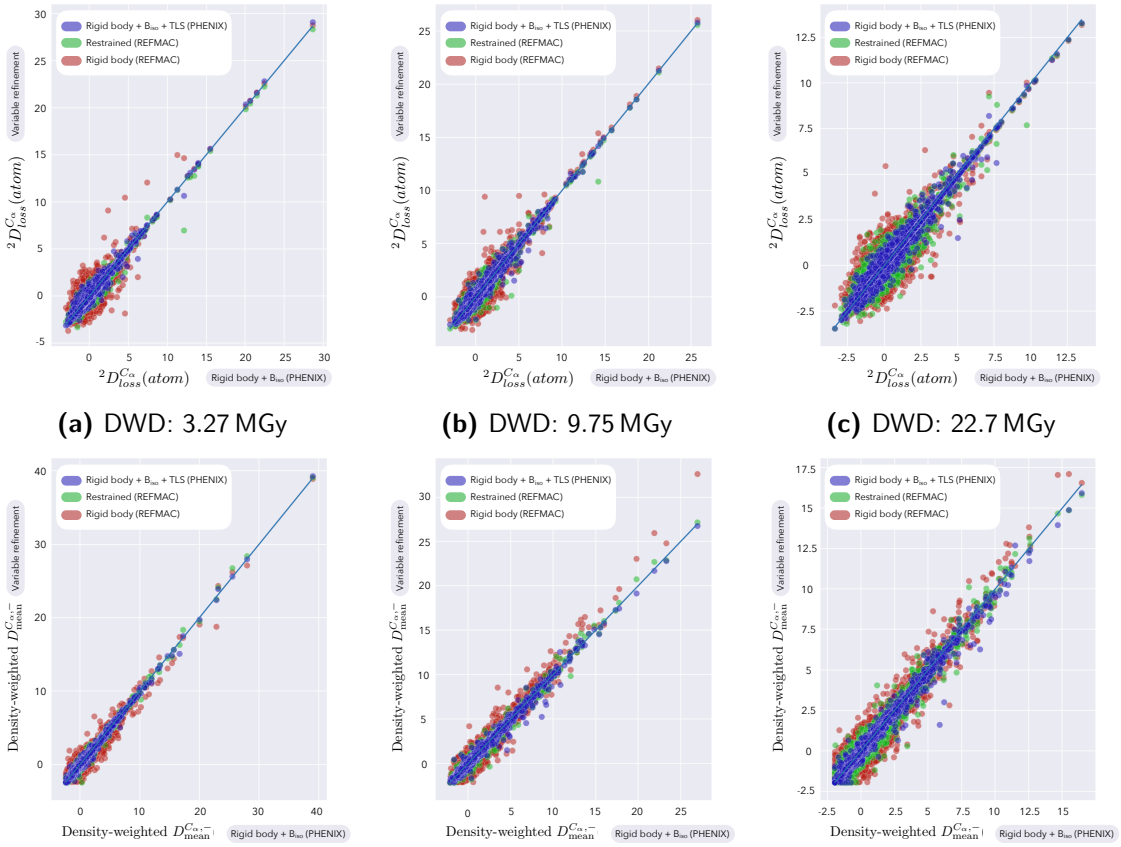
$F_{\text{obs,n}} - F_{\text{obs,1}}$ map generation requires use of a phase set, ϕ_{calc}, that has been generated from a refined coordinate model for the structure. In this chapter, the phase set corresponding to the initial dataset coordinate model, $\phi_{\text{calc},1}$, has been used by default to generate each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. In order to determine the dependence of the results of *RIDL* on the chosen phase set, three different phase sets were testing using *RIDL*:

1. The default case, in which phases, $\phi_{\text{calc},1}$, corresponding to the dataset 1 coordinate model are used for all $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps for $n > 1$.

2. The case in which, for each $n > 1$, the phases, $\phi_{\text{calc},n}$, corresponding to the dataset n coordinate model, are used to generate the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. Here the dataset n coordinate model has been derived following rigid-body and isotropic *B*-factor refinement of the *Dp*Cel7B coordinate model for dataset 1.

3. The case in which the phases, $\phi_{\text{calc},11}$, corresponding to the highest dose coordinate model ($n = 11$) are used for all $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps for $n > 1$. Again, here the dataset 11 coordinate model has been derived following rigid-body and isotropic *B*-factor refinement of the *Dp*Cel7B coordinate model for dataset 1.

The data in Figure 2.34 indicate that the detection of damage sites using *RIDL* is relatively robust to the choice in phase set used to calculate each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. Deviations from the diagonal line ($y = x$, in blue) are present for atoms attaining high $D_{\text{loss}}^{\text{C}_\alpha}$ values (> 5), and these become more significant at higher doses (Figure 2.34c-d). Interestingly, in Figure 2.34a-b, using the $\phi_{\text{calc},11}$ phase set, despite this corresponding to a much higher dataset, agrees well with the use of the $\phi_{\text{calc},1}$ phase set. As such, there is does not appear that the deviations from $y = x$ at



(a) DWD: 3.27 MGy **(b)** DWD: 9.75 MGy **(c)** DWD: 22.7 MGy

(d) DWD: 3.27 MGy **(e)** DWD: 9.75 MGy **(f)** DWD: 22.7 MGy

Figure 2.33: The dependence of *RIDL* metrics on the chosen higher dataset refinement strategy. For the example case of the *Dp*Cel7B GH7 family protein, the (a,b,c) C_α-normalised D_{loss}(atom) metric ($^2D^{C_\alpha}_{\text{loss}}$, Equation 2.5.14, Section 2.5.8) and (c,d,e) density-weighted & C_α-normalised D^{-}_{mean} have been calculated with *RIDL*. Metric values have been derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps with higher dose dataset n corresponding to 3.27, 9.75 and 22.7 MGy. Each x-axis indicates the metric values across all atoms in *Dp*Cel7B where higher dose datasets have been generated by rigid-body and isotropic *B*-factor refinement in *phenix.refine*, using the refined coordinate model from the first dataset coupled with the structure factor amplitudes, $|\mathbf{F}_{\text{obs,n}}|$ for the higher dose dataset $n > 1$. The corresponding y-axes indicate metric values where instead each higher dose dataset is generated following either 3 cycles of rigid-body, isotropic *B*-factor and TLS refinement in *phenix.refine* (in blue), 10 cycles of rigid-body refinement in *REFMAC* (in red), or 10 cycles of restrained refinement in *REFMAC* (in green). In all cases, the first dose coordinate model has been refined against the observed structure factor amplitudes $|\mathbf{F}_{\text{obs,n}}|$ for each higher dose dataset $n > 1$. Deviations of points from the line $y = x$ (in blue) indicate atoms for which the metric is sensitive to the choice of refinement strategy. Other doses for the *Dp*Cel7B damage series (not shown) exhibited similar behaviour.

high doses in Figure 2.34c-d represent any true benefit in using the $\phi_{\text{calc},11}$ of $\phi_{\text{calc},n}$ phase sets for map calculation.

Consequently, the default protocol remains to calculate all $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps using a single fixed $\phi_{\text{calc},1}$ phase set.

2.5.9.5 The effect of phase error

Directly extending from the previous section, the effect of degraded quality of this phase set on the output of the *RIDL* pipeline was next investigated.

Here, it is assumed that the phases are supplied to *RIDL* independently to the refined coordinate model, such that the placement of atoms (and consequently, per atom search radii) are unaffected. Phases were degraded by perturbing all x, y, z atomic coordinates in the lowest dose *Dp*Cel7B protein coordinate model, through a custom *Python* script, followed by the calculation of a new phase set, $\phi_{\text{calc,pert}}$ using *SFALL*. *RIDL* was then run using the degraded phases, but with all other input parameters held constant. A maximum displacement factor $\epsilon \geq 0$ was introduced to perturb each atom, such that the coordinates of an atom become:

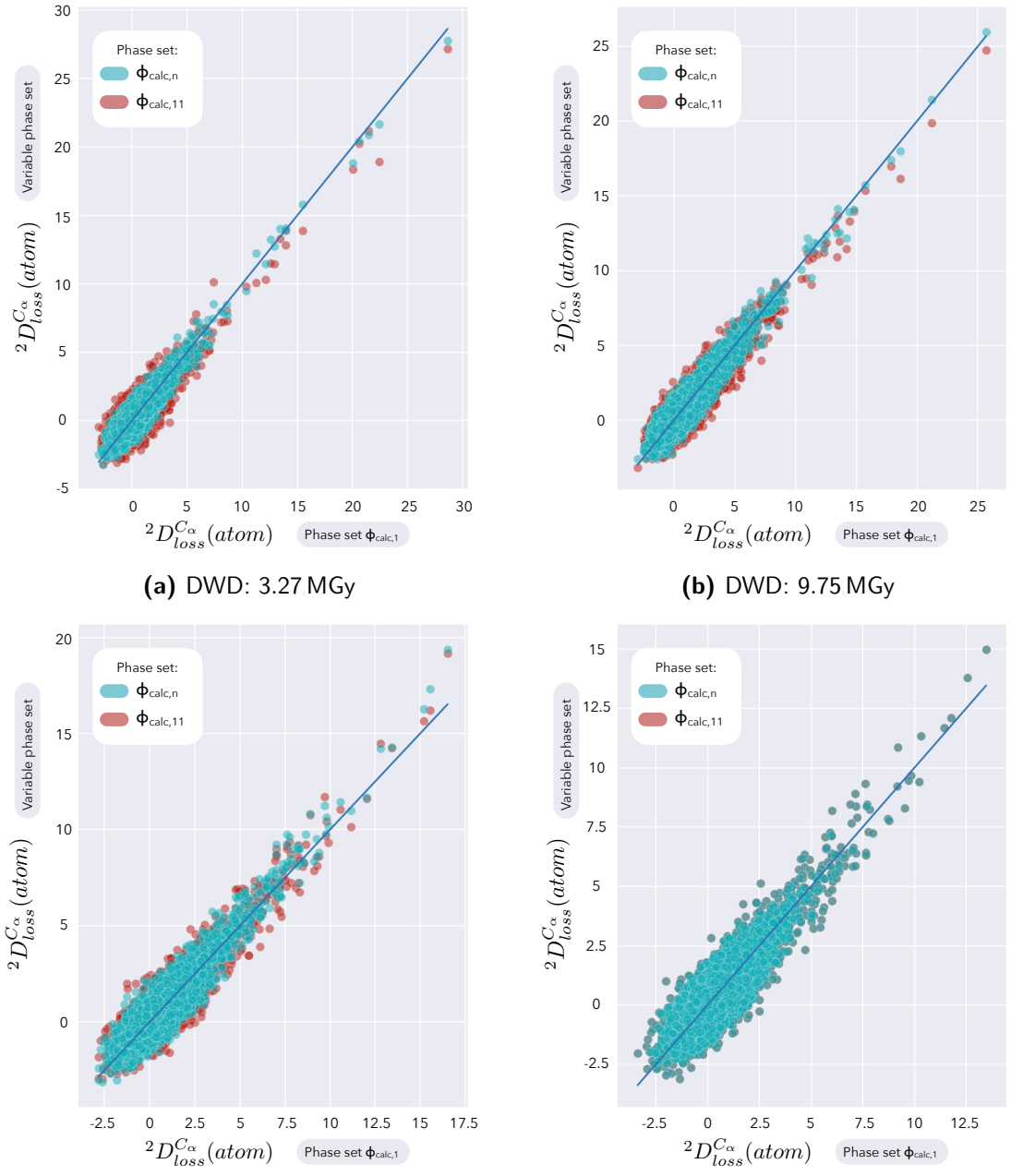
$$x_{\text{pert}} = x + \text{U}[-\epsilon, \epsilon] \tag{2.5.15}$$

$$y_{\text{pert}} = y + \text{U}[-\epsilon, \epsilon] \tag{2.5.16}$$

$$z_{\text{pert}} = z + \text{U}[-\epsilon, \epsilon] \tag{2.5.17}$$

where $\text{U}[-\epsilon, \epsilon]$ is a uniformly distributed random variable between $-\epsilon$ and ϵ. This process was repeated successively for increasing ϵ between 0 and 1. At $\epsilon = 1$, a mean difference in phase of 56.2° from the starting phases was calculated over all hkl Miller indices.

Figure 2.35 indicates that for the *Dp*Cel7B test study, *RIDL* is relatively robust to the effects of phase degradation, with the ordering of detected top damaged atom sites (with high $D_{\text{loss}}^{\text{C}_\alpha}$(atom) values) conserved up to a mean phase error of $\leq$ 42.7°. At 56.2°, and particularly at 22.7 MGy, the introduced phase error becomes too substantial for *RIDL* to correctly identify damage sites.



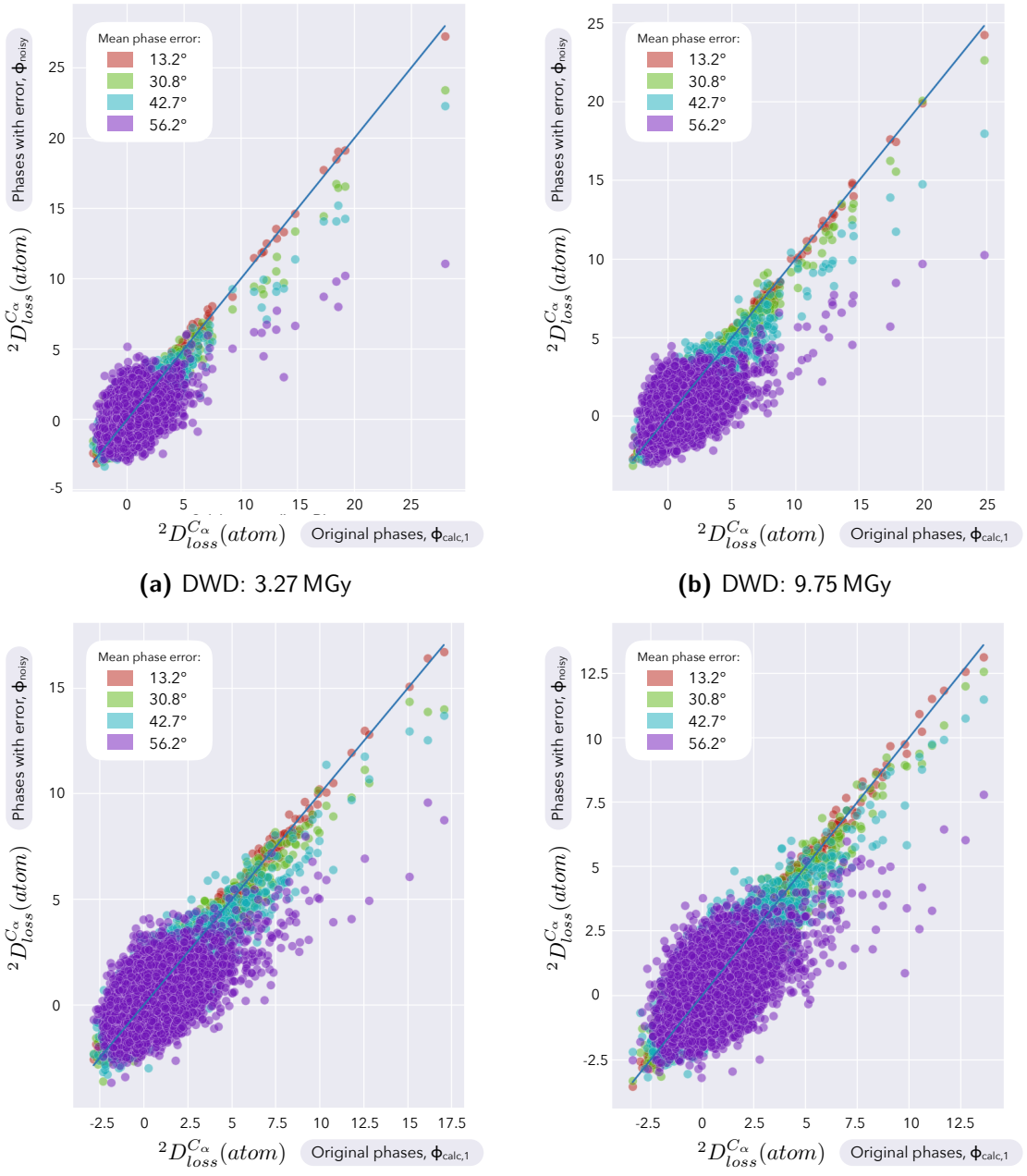
(a) DWD: 3.27 MGy

(b) DWD: 9.75 MGy

(c) DWD: 16.2 MGy

(d) DWD: 22.7 MGy

Figure 2.34: For the example case of the *Dp*Cel7B GH7 family protein, the dependence of the C_α-normalised D_{loss} metric on the choice of phase set used to calculate each $F_{obs,n} - F_{obs,1}$ map. Each x-axis indicates the $D_{loss}^{C_\alpha}$(atom) metric values (Equation 2.5.14, Section 2.5.8) across all atoms in *Dp*Cel7B, using the phases corresponding to the original refined coordinate model for dataset 1. The corresponding y-axes indicate metric values for which phases to calculate each $F_{obs,n} - F_{obs,1}$ map have instead been supplied from the dataset n for (a) $n = 2$, (b) $n = 5$, (c) $n = 8$ and (d) $n = 11$ (in green), or all the highest dose dataset $n = 11$ (in red). As expected, in (d) the two sets of data overlap exactly since $n = 11$. Deviations of points from the line $y = x$ (in blue) indicate atoms for which the metric is sensitive to the choice in phase set ϕ_{calc}. Other doses for the *Dp*Cel7B damage series (not shown) exhibited similar behaviour.



(a) DWD: 3.27 MGy

(b) DWD: 9.75 MGy

(c) DWD: 16.2 MGy

(d) DWD: 22.7 MGy

Figure 2.35: The dependence of the C_α-normalised D_{loss} metric on the $F_{obs,n} - F_{obs,1}$ map phase quality. For the example case of the *Dp*Cel7B GH7 family protein, the C_α-normalised D_{loss}(atom) metric (Equation 2.5.14, Section 2.5.8) has been calculated using *RIDL*, with the calculated phases $\phi_{calc,1}$ used to generate the $F_{obs,n} - F_{obs,1}$ being varied. Metric values have been derived from $F_{obs,n} - F_{obs,1}$ difference maps with higher dose dataset n corresponding to 3.27, 9.75, 16.2 and 22.7 MGy. The mean difference in phase between the original *Dp*Cel7B refined model and perturbed model are given in units of degrees. Each x-axis indicates the $D^{C_\alpha}_{loss}$(atom) metric values across all atoms in *Dp*Cel7B, using the phases corresponding to the original refined coordinate model (as detailed in Section 2.3.3). The corresponding y-axes indicate the metric values with degraded phases between 13.2° and 56.2° mean phase difference. Deviations of points from the line $y = x$ (in blue) indicate atoms for which the metric is sensitive to the quality of the $\phi_{calc,1}$ phase set. Other doses for the *Dp*Cel7B damage series (not shown) exhibited similar behaviour.

2.6 Discussion

In this chapter, two tools (*PEAKS* and *RIDL*) have been developed that aim to achieve a greater insight into site-specific damage, as detectable in $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps. These tools can be run when an MX damage series is available, comprising multiple complete diffraction datasets collected from a single macromolecular crystal at increasing dose states. Both *PEAKS* and *RIDL* have been presented in this chapter, since it is believed that both are essential to provide a complete analysis of radiation damage. The objectivity provided by *RIDL* is highly desirable, however manual inspection of $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference density can still be required to interpret the results.

$F_{\text{obs,n}} - F_{\text{obs,1}}$ map inspection has now become commonplace when performing a specific damage analysis of a protein structure. A major benefit of using $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps to search for damage in proteins is that the dependence on the quality of the coordinate model is minimised; a consideration that becomes particularly important at higher absorbed doses, when the reduced quality of diffraction data can lead to difficulties in the model building and refinement process. This is in direct contrast to the monitoring of atomic *B*-factors or xyz coordinates with increasing dose, which are fundamentally tied to model quality.

PEAKS (Section 2.4) was first designed to assist with the manual inspection process, which for large proteins and noisy $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps, rapidly becomes intractable. *PEAKS* filters the set of detected $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks to a smaller subset of interest to the investigator, and furthermore provides an interface to improve the efficiency at which the remaining peaks can be inspected. Consequently, larger protein systems can be investigated with reasonable ease in a shorter duration of time, as has been demonstrated in this chapter for the *Dp*Cel7B protein test case (Section 2.3). Difference map peak heights that are assigned to atoms as indicating damage are quoted in units of e/Å^3 in *PEAKS*. This is a significant improvement on the more common practice of quoting peak heights in map σ-units, which are map-specific and not directly comparable at different doses. Manual inspection is, however, still an inherent and subjective component of the *PEAKS* program.

The second tool, *RIDL*, was then developed to remove the requirement for manual inspection

entirely, by providing an objective tool to quantify the local behaviour of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference map around individual atoms (Section 2.5). To this end, metrics have been specifically designed that aim to suitably characterise site-specific radiation damage to atoms.

The D_{loss}(atom) metric appears a suitable metric for site-specific damage detection because it provides a worst case scenario for the electron density lost by an atom. As such, D_{loss} can provide the investigator with information on local radiation-sensitive *hotspots*, where the most caution should be taken before structural insight is drawn. It has been demonstrated that the D_{loss} metric generated by *RIDL* is highly consistent with the output of *PEAKS* (Figure 2.23), thus suggesting that the time-consuming process of manual inspection could be avoided entirely by use of this metric.

In general, in order to quantify the total electron density lost by an atom, as opposed to an upper bound that D_{loss} provides, an average must be taken over the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map points surrounding an atom. However, it is unsuitable to simply take an average within a specified limiting radius, since the averaging is highly dependent on the chosen radius. To reduce the strict dependence on limiting radius, a ρ_{calc} map density-weighting scheme can be applied in order to down-weight the importance of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density further from an atomic centre. The resulting density-weighted D^{-}_{mean} metric is then theoretically a more suitable averaging metric for quantifying the magnitude of radiation-induced density loss at an atom. Unfortunately, density-weighting comes with the *caveat* that the resulting metric is now more dependent on the quality of the coordinate model from which the ρ_{calc} map is derived. Due to these limitations, it is suggested that both D_{loss} and the density-weighted D^{-}_{mean} metric should play a role when conducting a specific damage analysis.

Metric normalisation schemes (C_{α}-normalisation to the protein backbone) have also been implemented in *RIDL* in an attempt to allow each per-atom metric to be directly comparable between independent protein systems. Such considerations will be essential for the investigations conducted in later chapters, as will be outlined below.

2.6.0.1 Application of tools to MX damage series

The development of the tools detailed above has enabled the quantitative investigation of multiple independent macromolecular systems, that would not have been achievable by manual inspection of $F_{obs,n} - F_{obs,1}$ difference maps alone. In the work described in Chapters 3,4 and 5, the tools have been used to characterise site-specific damage manifestations in nucleic acids (both co-crystallised with interacting proteins and alone). Nucleic acids have anecdotally been considered radiation-hardy in the previous literature (McGeehan *et al.*, 2007), however a systematic investigation has been lacking. Using *PEAKS*, it has been possible to perform a detailed manual inspection of $F_{obs,n} - F_{obs,1}$ electron difference density for nucleic acids at multiple increasing doses. The *RIDL* program has further permitted a direct quantitative comparison of the radiation-induced changes to nucleic acids with the much better characterised protein components of nucleoprotein complex crystals.

In Chapter 6, work using *RIDL* is reported that has enabled a highly systematic comparison of radiation damage to all protein MX damage series that have previously been deposited in the PDB. In particular, the mechanism of tyrosine radiation damage has been studied across protein systems; whereas clear damage mechanisms for disulphide bond cleavage and decarboxylation have been proposed in the literature, damage to Tyr is less well charactered across proteins. This analysis would be unachievable without the high-throughput methods presented here.

Detection of specific damage using $F_{obs,n} - F_{obs,1}$ maps has classically required the collection of multiple complete diffraction datasets from a crystal. For standard MX data collection, this requirement is not typically achieved, and as such, no quantification of specific damage can be made. In Chapter 7, it will be demonstrated that *RIDL* can be adapted to detect the onset of damage in only a single complete set of diffraction images that would typically all be used to solve the MX structure. This application of *RIDL* could ultimately enable an assessment of radiation damage effects for any solved macromolecular structure deposited in the PDB.

CHAPTER 3

Radiation damage to nucleic acids in complex with proteins. Part I: C-protein DNA complexes

3.1 Introduction

To date, systematic crystallographic investigations regarding radiation-induced changes to DNA (deoxyribonucleic acid), RNA (ribonucleic acid) and the broader class of nucleoprotein complexes have been lacking relative to their well-documented protein counterparts. Nonetheless, nucleic acid-containing structures now comprise approximately 6.4 % of MX-derived structures deposited in the PDB, and this number is anticipated to increase as the structural biology community strives to solve increasingly large ($\geq$ 4000 kDa) and complex macromolecular systems encompassing those bound to nucleic acids. A thorough characterisation of MX radiation damage affecting fundamental non-protein biological components is thus essential to ensure that correct structural interpretations are made at the atomistic scale provided by crystallography.

Crystalline nucleic acids have previously been reported to be radiation-insensitive at the typical doses achieved in MX data collection at 100 K (J. McGeehan, private communications), however the underlying mechanisms behind this robustness are yet to be elucidated. RIP phasing for a DNA/RNA hybrid crystal was only successful due to the radiation-induced cleavage of Br from the C_8 of brominated guanine (Ravelli *et al.*, 2003); the oligonucleotides themselves exhibited no clear site-specific damage. These observations were reinforced by a later MX investigation coupled with Raman spectroscopy (McGeehan *et al.*, 2007), where the breakage of the C−Br bonds in 8-bromo-2'-deoxyguanosine was identified and monitored as a function of dose in non-resonant Raman spectra. Again, DNA debromination was the only chemical change detected, even at the highest accumulated dose tested ($\sim$15 MGy). It is important to note that in both of the above studies, the radiation insensitivity of nucleic acid was a reported byproduct during the demonstration of each respective technique (RIP phasing and Raman spectroscopy), and provides no evidence whether the lack of nucleic acid damage is more widely applicable to non-brominated DNA/RNA crystals.

The suggested robustness of nucleic acids in MX is in striking contrast to the wealth of radiation damage studies on nucleic acids that have been conducted by radiation chemists (e.g. Alizadeh *et al.* (2015); Cadet *et al.* (1999); Michaud *et al.* (2012)), which have indicated that both single and double strand breakages (SSBs and DSBs) readily occur for nucleic acids in solution, and at much lower doses than at those typically inflicted in MX. This discrepancy alone motivates the

need for a systematic study of radiation damage to a variety of oligonucleotide sequences, as will be presented in the current chapter.

For both nucleic acids and nucleoprotein complexes in solution at RT, secondary diffusive hydroxyl radicals produced through solvent radiolysis are known to abstract H atoms from deoxyribose and bases (Michael and O'Neill, 2000; Cadet *et al.*, 1999), and add to unsaturated linkages in both DNA and RNA bases, resulting in SSBs and modified bases and sugars (Chance *et al.*, 1997; Spotheim-Maurizot and Davídková, 2011). Overall it has been estimated that two-thirds of DNA damage in dilute aqueous solution at RT can be attributed to indirect secondary free radical damage from solvent around the DNA (Michael and O'Neill, 2000). However, most radical species (such as hydroxyl radicals (Owen *et al.*, 2012)) are believed to be immobilised below 110 K, and as such, the prevalence of such hydroxyl-mediated damage is not anticipated to be a dominant pathway in cryocooled MX experiments carried out at 100 K.

Low energy electrons (LEEs, with energies of several eV) have been shown to be mobile at temperatures as low as 77 K (Jones *et al.*, 1987), with electron spin resonance (ESR) spectroscopy studies postulating rapid quantum tunnelling along both the protein backbone and nucleic acid stacked base π-systems as the dominant mechanism by which LEEs can relocate from initial ionisation sites at 77 K (Symons, 1997). In MX, after primary photoabsorption of an incoming X-ray photon by both the macromolecule and crystal solvent channels, each ejected high-energy primary photoelectron (as well as associated Auger electrons) is predicted to have enough energy to produce a significant yield of secondary LEEs along its track length (up to $\sim$ 500 secondary LEEs along a $3-4$ µm track length for the incident energies employed in MX (Nave and Hill, 2005; Sanishvili *et al.*, 2011)). These LEEs can then cause further ionisation events as they gradually thermalise throughout the crystal (O'Neill *et al.*, 2002). Ultimately, these LEEs become trapped by electron-affinic sites in both protein and nucleic acid to form electron gain centres in proteins and $T^{\bullet -}$ and $C^{\bullet -}$ centres in nucleic acids. Positive holes are also believed to migrate at 100 K, but to a lesser extent, being more readily trapped, predominantly by deprotonation at protein backbone amide units and at acidic residue side-chain carboxyl groups, and to form $G^{\bullet +}$ centres in nucleic acids.

Extensive experimental studies have been performed to date on such sub-ionisation level LEEs ($0 - 15$ eV) interacting with short DNA oligonucleotides deposited on films at varying degrees

of hydration, and at varying temperatures (predominantly at RT, 4 K, and 70 K) (Alizadeh and Sanche, 2014; Ptasińska and Sanche, 2007). Transient resonant attachment by sub-ionisation-level LEEs to π^* orbitals of DNA bases and the sugar-phosphate backbone, followed by rapid bond dissociation, is now believed to be sufficient for cleavage of both base-sugar N_1C bonds and strong (~ 4 eV, 385 kJ mol^{-1}) covalent bonds of the DNA phosphodiester backbone (Barrios *et al.*, 2002; Berdys *et al.*, 2004; Simons, 2006). Further, computational chemistry calculations have indicated the preferential cleavage of sugar-phosphate C−O bonds over other bonds present in the DNA structure (Théodore *et al.*, 2006). Such bond ruptures are ultimately detected as yields of SSBs, most typically through gel electrophoresis analysis (Boudaïffa *et al.*, 2000). Similar preferential rupture of C−O bonds (in particular the 5' C−O bond) has been reported in a recent X-ray absorption spectroscopy (XAS) study for DNA oligonucleotides in solution under either UVA photon or proton irradiation (Czapla-Masztafiak *et al.*, 2016).

Investigation of radiation damage to interacting protein-nucleic acid complexes is of greater biological relevance than for nucleic acids alone, since nucleic acids are rarely isolated in the cellular environment, instead dynamically interacting with a plethora of proteins such as the nucleosomal histones through to high-order replication, transcription, modification and repair machinery. Indeed, nucleic acids make up some of the fundamental structural and catalytic components of large macromolecular machines such as ribosomes and spliceosomes. Radiation damage solution studies on nucleic acid-containing complexes, typically conducted between $0 - 4\,^{\circ}$C and at much lower doses (< 1 kGy) than those experienced in typical MX experiments, have reported a significant bias towards protein damage as compared with effects on the nucleic acid component. Such studies have typically involved irradiating both the protein and DNA in isolation as well as in the complexed form, and then assessing subsequent binding activity by gel electrophoresis. For the *lac* repressor-operator complex, which is considered to be the paradigm of protein-DNA interacting systems (Barkley and Bourgeois, 1978), γ-irradiation of the complex at ~ 273 K induced peptide chain cleavage and side-chain oxidation by $OH^{\bullet}$ radicals, inducing loss of the repressor protein's ability to bind to the operator DNA sequence of the *lac* operon (Charlier *et al.*, 2002; Eon *et al.*, 2001). The most probable oxidation targets by $OH^{\bullet}$ radicals at the DNA-binding interface of the *lac* repressor protein have subsequently been predicted using a combination of spectroscopy and mass spectrometry (Gillard *et al.*, 2007). It has been demonstrated that bound DNA affords a substantial protective effect in the *lac* repressor-operator complex and that the

naked protein has a relatively higher sensitivity to irradiative damage (Eon *et al.*, 2001). Physical screening of DNA by protein has been suggested to shield the DNA-protein interaction sites from radical damage, yielding an extended life-dose for the complex. The probability of DNA strand breakages is also suppressed by protein-binding, with those regions most protected being situated at protein interfaces (Begusova *et al.*, 2001). The protection of protein by DNA (and vice versa) has similarly been made between the DNA glycosylase Fpg and its abasic DNA target site at 4 °C (Gillard *et al.*, 2004), and for the estrogen response element-receptor complex at 0 °C (Štísová *et al.*, 2006).

The applicability of the aforementioned radiation damage solution studies at $\sim$ 0 °C is unclear, due to the immobility of $OH^{\bullet}$ radicals at 100 K. Radioprotection of DNA by amino acids has also been reported upon exposure to LEEs. In molecular dynamics simulations of interacting thymine nucleobases with free glycine amino acids in both solution and the condensed phase simulations, it has been suggested that either glycine can scavenge the LEEs prior to interaction with the nucleobase, or if the LEE is initially instead captured by the thymine, glycine could donate a proton to minimise the net partial charge on the thymine (Gu *et al.*, 2014). These simulations agree well with experimental reports on $(dT)_{25}$ oligonucleotides complexed with a single-strand-binding protein whilst tethered to a gold surface and irradiated with 3 eV LEEs (Solomun and Skalický, 2008). However Solomun and Skalický (2008) postulated that DNA radioprotection could be due to both physical shielding of DNA by protein and also engagement of the lowest unoccupied molecular orbitals (LUMOs) of each nucleobase through protein binding, such that they are no longer available for LEE attachment (Solomun and Skalický, 2008).

3.1.1 Aims of the current chapter

In comparison to above, to date, no studies have been presented on MX radiation damage to crystalline protein-nucleic acid complexes, and it is unclear how damage processes can affect structural solution of nucleoprotein complexes in MX. To elucidate this, Sections 3.2 and 3.3 will now present two systematic MX specific damage investigations on two independent bacterial protein-DNA systems: the C.Esp1396I-DNA complex and the C.Csp231I-DNA complex. Quantitative controlled comparisons between crystallised protein and nucleic acid damage mechanisms separately is inherently difficult, but such challenges are circumvented through investigating such

naturally forming nucleoprotein complexes. As such, for the selected systems, nucleic acid can be compared directly with well-characterised protein components under identical crystallisation, diffraction data collection and data processing conditions.

The current chapter aims to build on previous radiation damage solution studies by employing techniques that will allow us to investigate radiation induced damage processes to nucleoprotein complexes at the near-atomic level provided by X-ray crystallography.

3.2 The C.Esp1396I-DNA complex

In this section, a well characterized bacterial protein-DNA complex (C.Esp1396I/DNA, $\sim 60\,\mathrm{kDa}$) is employed to probe the effects of X-ray induced radiation damage and its quantitative distribution among the various components at the atomic level. To the author's knowledge, this is the first MX site-specific radiation damage study that has been conducted on a protein-nucleic acid crystal.

The C-protein, C.Esp1396I, is a member of a large class of helix-turn-helix proteins that act as transcriptional regulators of gene expression in both prokaryotic and eukaryotic systems. Restriction-modification (R-M) systems encode both a restriction endonuclease and DNA methyltransferase; the DNA methytransferase modifies host DNA with a specific methylation pattern to then protect it from cleavage from the associated restriction endonuclease. Overall this provides a simple yet powerful protective mechanism for cells to distinguish and eradicate foreign DNA. The C.Esp1396I protein finely regulates the bacterial *Esp1396I* R-M system, providing a delay in endonuclease expression via a concentration-dependent 'on' and 'off' switch (McGeehan *et al.*, 2008; Ball *et al.*, 2009). As Figure 3.1 illustrates, the C.Esp1396I protein achieves this task by binding to a joint C-protein and restriction endonuclease promoter: low yields of C.Esp1396I protein stimulate expression of additional C.Esp1396I protein and endonuclease, whereas as the yield of C.Esp1396I increases, expression is halted. This provides the essential time for host DNA to become methylated prior to interaction with endonuclease.

The C-protein-DNA operator complex investigated here (Figure 3.2) consists of two 'controller' C-protein dimers (chains A and B, chains C and D) bound to a 35 bp DNA operator sequence (chains E and F). This system was chosen as a model candidate for MX site-specific damage analysis

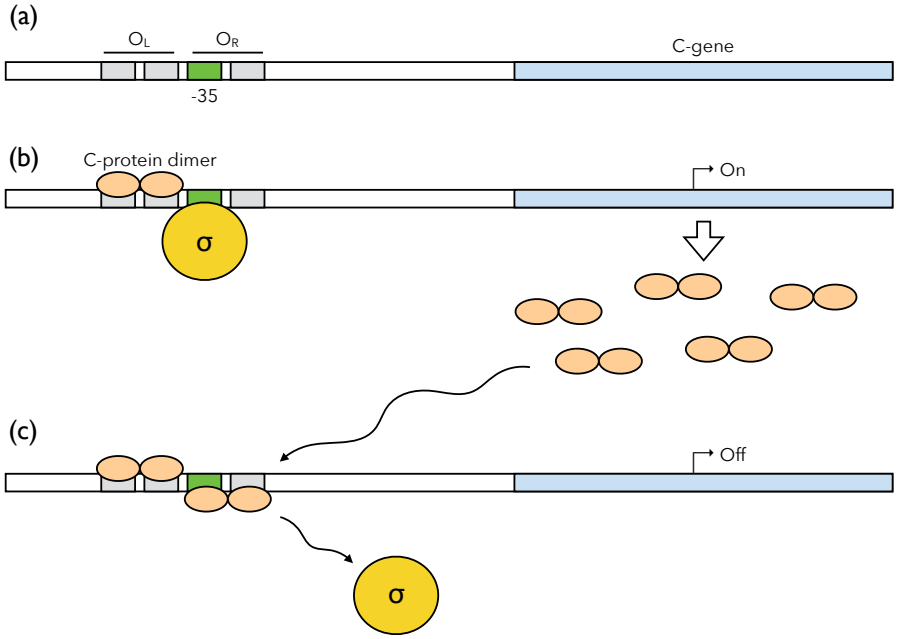


Figure 3.1: The mechanism by which C.Esp1396I protein regulates the timing and expression of the *Esp1396I* bacterial R-M system. Figure has been reproduced and modified from McGeehan *et al.* (2008). (a) The DNA operator sequence includes two binding sites (O_L and O_R) for two C-protein dimers. (b) Low-level C-protein expression occurs from an external weak C-protein-independent promoter (not shown in diagram). C-protein dimers are suggested to bind to the higher affinity O_L site, stimulating transcription of the C-gene (and also the restriction endonuclease gene, not shown in the diagram) via recruitment of RNA polymerase σ^{70} subunit to the -35 site. (c) At high concentrations of C-protein, binding of a second C-protein dimer at the weaker affinity O_R site blocks σ^{70} at the overlapping -35 site, thus switching off expression of both the C-protein and restriction endonuclease genes.

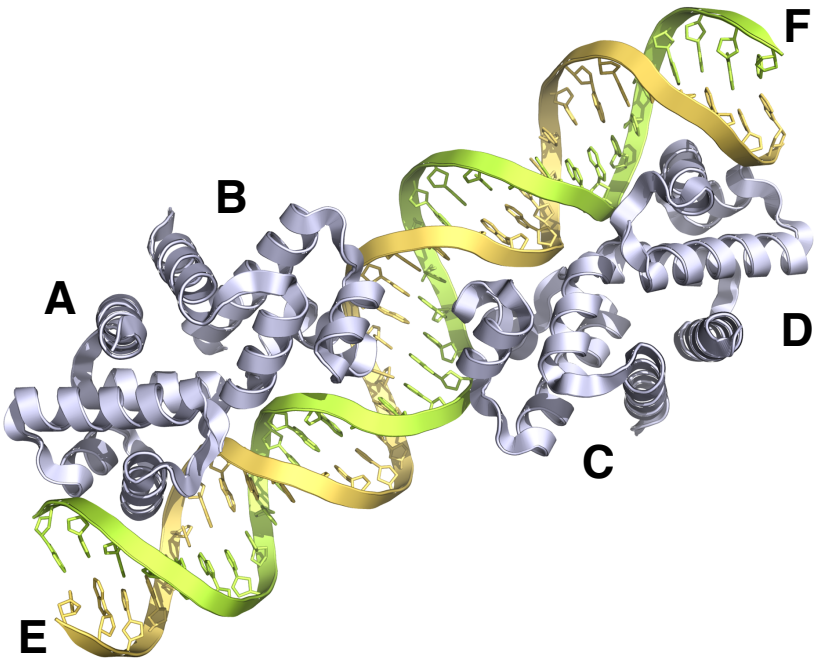


Figure 3.2: The C.Esp1396I-DNA tetrameric complex, rendered in *PyMOL*. PDB: *4x4b*, resolution: 2.8 Å. The complex exhibits a pseudo-dyad axis between bp 17 and 18 of the dsDNA such that chain A rotates $\sim 180^{\circ}$ around this axis onto chain D (and similarly chain B rotates onto chain C). However, non-symmetrical binding interactions between protein chains B and C at the dimer-dimer interface, and a pseudo-symmetrical DNA sequence, prevent true non-crystallographic symmetry (NCS) in the complex in the crystal.

due to the relatively large 35 bp dsDNA (double stranded DNA) component, thus resulting in the number of protein and DNA atoms being of the same order (protein: 2496 refined non-hydrogen atoms; DNA: 1429 non-hydrogen atoms). This feature has enabled a statistically significant comparison to be made between the frequencies of detected specific damage observed on the DNA and protein components.

Furthermore, the dsDNA is distorted in the complex, due to binding of protein dimers to each operator site, resulting in minor groove compression (leading to a 50° bend at each binding site) and large major groove expansion at the DNA sequence centre (McGeehan *et al.*, 2008). An aim of the current investigation was to determine whether there is any relationship between local DNA geometry and site-specific radiation damage.

3.2.1 Crystallisation

All expression, purification and crystallization of the native C-protein-DNA operator complex C.Esp1396I was performed by Prof. J. McGeehan, University of Portsmouth, as described in

McGeehan *et al.* (2008). In short, His-tagged protein was expressed and purified from large-scale cultures of *Escherichia coli* BL21 (DE3) cells bearing the plasmid pET-28b/esp1396IC. Pure C.Esp1396I protein at a concentration of 0.7 mg ml^{-1} was mixed with purified double-stranded 35 bp DNA (sequence: ATGTGACTTATAGTCCGTGTGATTATAGTCAACAT bound to its complementary strand) at a 4:1 molar ratio and then crystallized using a vapour-diffusion sitting-drop method at 293 K with 100 nl nucleoprotein solution and 100 nl mother liquor drop. Crystals grew within one month in 50 mM 2-(N-morpholino)ethanesulfonic acid (MES), pH 7.5, 25 % 2-methyl-2,4-pentanediol (v/v) and 40 mM $MgCl_2$. Crystals were cryoloop mounted and vitrified directly under a 100 K N_2 gas stream, with no additional cryoprotectant agent being added to the mother liquor.

3.2.2 X-ray data collection

All diffraction data were collected at 100 K at beamline ID29, ESRF, by Prof. J. McGeehan, University of Portsmouth, using a wavelength of 0.932 Å (13.3 keV) and an ADSC Q315R mosaic CCD detector. A pinhole designed by R. Ravelli and F. Felisaz (EMBL Grenoble) was utilized to produce a 25 µm circular low-divergence beam at the sample. The beam size before the pinhole was 0.212 mm (vertical) × 0.279 mm (horizontal) (Gaussian profile FWHM). A triangular-wedge-shaped crystal of C.Esp1396I with approximate dimensions of 30 µm × 30 µm × 10 µm) was oriented with the beam direction parallel to the smallest crystal dimension (10 µm), as shown in Figure 3.3. The crystal-detector distance was fixed to 390 mm throughout data collection.

Eight datasets were collected from this C.Esp1396I crystal, each consisting of 100 frames of 1° rotation, each over the same 100° wedge of the crystal, with exposure times per frame and beam attenuation being varied according to the values shown in Table 3.1. The unattenuated beam (at 100 % beam transmission) was included for dataset 7 in order to damage the crystal at a higher rate than in datasets 1-6 (an X-ray 'burn'), to ensure that data had been collected for a large range of absorbed X-ray doses by the last dataset.

The collected diffraction data were then supplied by Prof. J. McGeehan as a series of unprocessed diffraction images. All of the following data processing and analysis has been performed by the author of the current thesis.

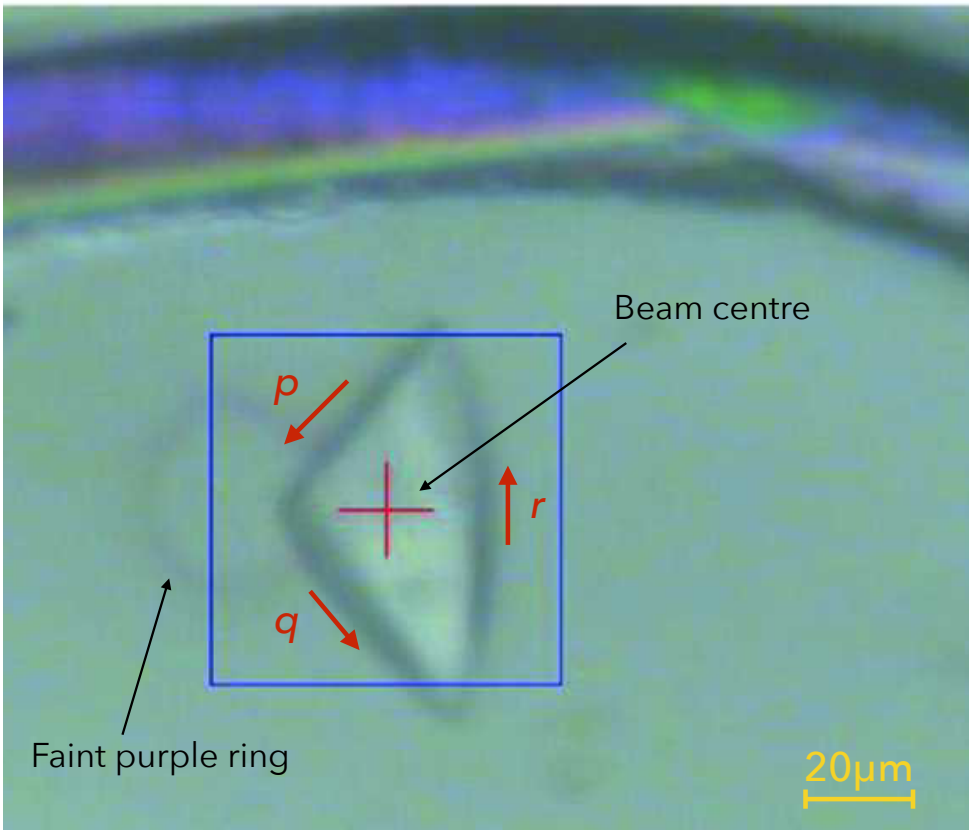


Figure 3.3: The C.Esp1396I crystal in the rayon fibre loop during diffraction data collection. As illustrated, crystal dimensions were approximately $p = 30\,\mu m$, $q = 30\,\mu m$ and $r = 90\,\mu m$, with the crystal positioned in the loop such that dimension $z = 10\,\mu m$ is orientated directly into the page. For reference of scale, the blue box shown around the crystal is $50\,\mu m \times 50\,\mu m$. A faint purple ring can be seen to the left of the crystal, postulated to be formed by the production of solvated electrons upon X-ray irradiation. This was created from a test shot of the X-ray beam prior to crystal centering and provides a visual confirmation of the circular 25 μm beam profile generated from the custom pinhole at beamline ID29, ESRF.

Table 3.1: Diffraction data collection details for each data set of the C.Esp1396I-DNA damage series. All data sets were collected with an angular rotation of $\Delta\phi = 1^\circ$ per image. Each diffraction-weighted dose (DWD) is calculated by *RADDOSE-3D* as described in (Zeldin *et al.*, 2013a).

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

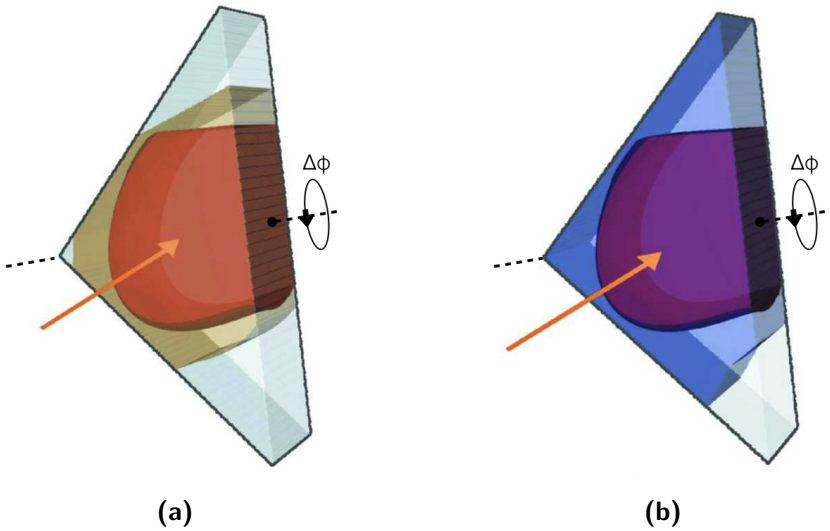


Figure 3.4: *RADDOSE-3D* calculation of dose distributions in the triangular C.Esp1396I crystal after (a) the first dataset and (b) the eighth dataset. In (a) the dose isosurfaces represent 0.16 MGy (clear), 3.3 MGy (brown) and 4.2 MGy (orange) and in (b) 0.16 MGy (clear), 20 MGy (blue) and 47.5 MGy (dark red). The initial direction of the X-ray beam relative to the crystal prior to the start of each dataset collection is indicated with an orange arrow, and the crystal was then rotated about the axis as shown.

3.2.3 Dose calculation with *RADDOSE-3D*

RADDOSE-3D (Zeldin *et al.*, 2013b) was used to calculate the absorbed diffraction-weighted dose (DWD) (Zeldin *et al.*, 2013a) for each of the eight datasets (Table 3.1). Based on the image of the crystal mounted on the beamline (Figure 3.3), the crystal geometry could be approximately modelled using the open source *Blender* software package (*https://www.blender.org*), in order to supply a polyhedral description of the crystal shape (in the form of a set of vertices and planes) to *RADDOSE-3D*. The approximated crystal shape and absorbed dose distributions calculated after the first and eighth datasets are shown in Figures 3.4a and b, respectively.

The intensity profile of the 25 µm-diameter beam was modelled as a Gaussian with FWHM values as detailed in Section 3.2.2. The photon flux prior to beam attenuation was estimated to be 5×10^{10} photons/s throughout the experiment, obtained from the in-beamline ionization gauge I1 which had been cross-checked with a 500 µm thick calibrated silicon PIN diode (Owen *et al.*, 2009) on the day prior to data collection. The calculation of the crystal absorption coefficients in *RADDOSE-3D* included both the water and the heavy-atom content from the crystallization conditions (the sulphur in the MES buffer and the magnesium and chlorine from the $MgCl_2$).

The data shown in Figure 3.5 illustrate the mean intensity values per whole dataset, normalised

to that of the first dataset, and indicate that radiation damage induced decay with increasing dose is approximately linear, in good accordance with previous reports for protein crystals (Owen *et al.*, 2006). For each dataset, $n \geq 1$, the mean intensity, I_n, has been calculated from the list of intensities per resolution bin, as reported by the log file output by *AIMLESS*, as follows:

$$I_n = \frac{\sum_{i=1}^{N_{\text{res}}} \langle I \rangle_i \cdot N_{\text{meas},i}}{\sum_{i=1}^{N_{\text{res}}} N_{\text{meas},i}} \tag{3.2.1}$$

where N_{res} is the number of resolution shells in the diffraction image, and $N_{\text{meas},i}$ $\langle I \rangle_i$ are the number of observed reflections and average intensity, respectively, for resolution shell $i \in \{1, ..., N_{\text{res}}\}$. The DWD to half mean intensity, $D_{1/2}$ is approximately 43 MGy, in good agreement with the $D_{1/2}$ dose limit of 43 ± 3MGy that has previously been experimentally determined as a recommended upper bound for general protein crystals (Owen *et al.*, 2006). There is insufficient evidence to suggest that nucleic acid-containing crystals significantly surpass this established $D_{1/2}$ dose limit for proteins.

3.2.4 Diffraction data processing

For each dataset, each set of diffraction images was independently integrated using *MOSFLM* (Leslie and Powell, 2007), followed by data reduction using *POINTLESS*, *AIMLESS* and *CTRUNCATE* as part of the *CCP4i* GUI (Winn *et al.*, 2011). Since the the crystal used for data collection was found to belong to space group $P6_5$, a consistent indexing convention was enforced between the 8 datasets using *POINTLESS*.

For the first dataset, molecular replacement was then performed using the program *PHASER* (McCoy *et al.*, 2007), utilising the previously deposited C.Esp1396I complex structure PDB entry *3clc* (McGeehan *et al.*, 2008) as the search model (with 100 % protein and DNA sequence identity). In agreement with the original structure, each asymmetric unit contained one tetrameric protein-DNA complex and had a solvent content of 68.7 %. It is noted that due to the high sequence similarity between the search model and target protein, there was no strict requirement for molecular replacement here: instead rigid-body refinement of the *3clc* coordinate model directly against the observed structure factor amplitudes for the current dataset 1 would have been sufficient, having first ensured that a consistent indexing scheme were present between the *3clc*

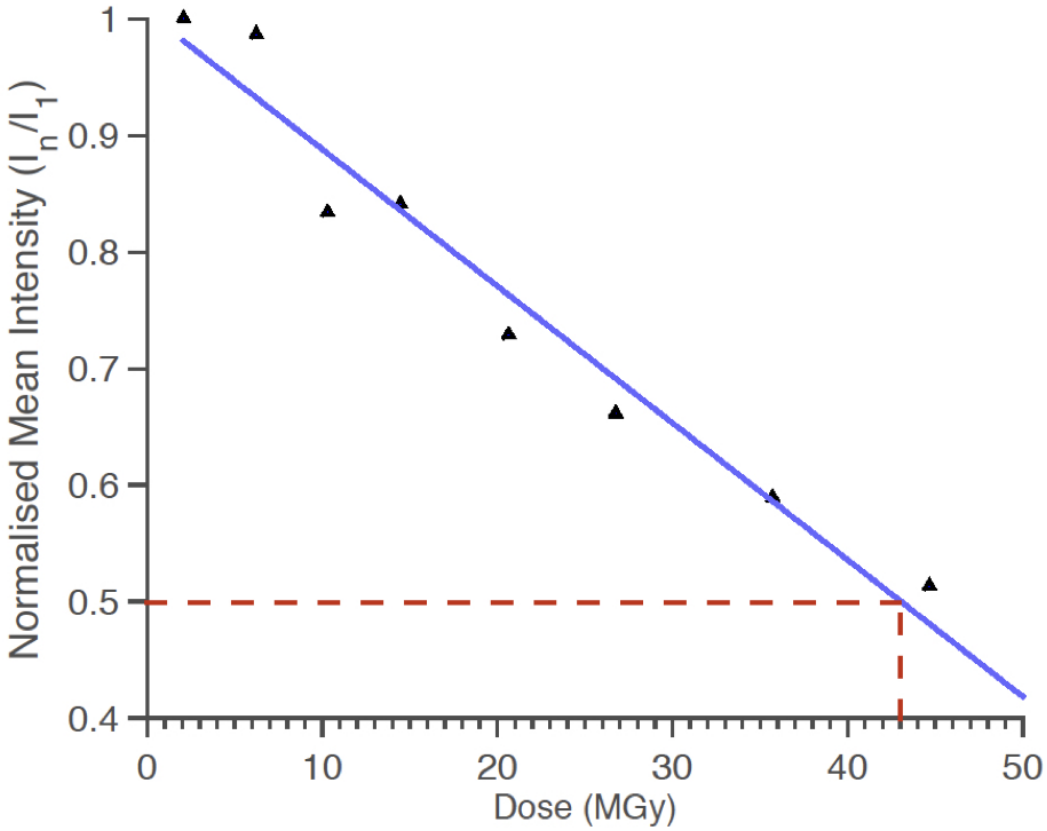


Figure 3.5: Mean intensity decay, I_n, for the 8 C.Esp1396I datasets, normalised to the mean for the first dataset, I_1. The mean intensity is calculated for resolution shells between 69 Å and 2.8 Å. The data are fitted with a least-squares linear fit in *MATLAB* (Matlab, 2010) which gives a $D_{1/2}$ of 43 MGy.

reference data and current data. In the remaining chapters of this thesis, molecular replacement has similarly been used to extract an initial set of phases for each of the subsequent protein case studies, and an identical argument also applies as to why molecular replacement was an unnecessary step. However, the current discussion will not be reproduced in full in later chapters.

The structure obtained from the first dataset was then refined, initially using 10 cycles of rigid-body refinement in *REFMAC5* (Murshudov *et al.*, 2011), then followed by repeated rounds of restrained, TLS and isotropic *B*-factor refinement (typically 10 cycles of each per round of refinement). Here the default TLS groups as defined by *REFMAC5* (1 TLS body per protein/DNA chain, comprising 6 in total) were used during TLS refinement. Refinement was coupled with manual inspection and refinement in *COOT* (Emsley *et al.*, 2010), involving solvent molecule removal and addition. In *REFMAC5*, non-crystallographic symmetry (NCS) restraints were used throughout for the four protein subunits present, excluding the flexible loop region corresponding to residues 43-46. This strategy was in accordance with the refinement strategy used for the previously solved C.Esp1396I structure (Ball *et al.*, 2009), in which variability in this loop region was observed between C-protein monomers. In the current study, the $2mF_{obs,n} - DF_{calc}$ map electron density was poorly defined in these loop regions, however due to the limited resolution (2.8 Å), no clear alternative conformations could be distinguished. Consequently the primary loop conformation from the *3clc* search model was used as a template in these regions. Final rounds of TLS, restrained and isotropic *B*-factor refinement were then performed for the first dataset using *phenix.refine* in the *PHENIX* program suite (Adams *et al.*, 2010), utilizing the integrated *MolProbity4* (Chen *et al.*, 2010) structure validation within the *PHENIX* interface to manually assess Ramachandran, rotamer and bond angle/length outliers (Table 3.3). The final model for the first dataset was refined to a resolution of 2.8 Å.

For the later datasets (2 to 8), molecular replacement was performed with *PHASER* using the refined first dataset final model as the search model. It is noted that this was an unnecessary step, and instead the refined coordinate model for the first dataset 1 would have provided a suitable starting model for refinement, since it had been derived from the exact same crystal with the protein in approximately the same orientation within the asymmetric unit*. For the

*However, it is known that proteins can undergo progressive rotation within the unit cell as a result of increasing X-ray exposure (Ravelli and McSweeney, 2000), and thus there is anticipated to a marginal difference in orientation of the protein between dataset 1 and the later datasets.

later datasets, starting models output from molecular replacement were then refined only using rigid-body and isotropic *B*-factor refinement in *phenix.refine* and were similarly refined to 2.8 Å. This is a similar protocol to that used to refine the higher dose datasets for the *Dp*Cel7B protein damage series solved previously as described in Section 2.3.3. A discussion into why independent restrained refinement of coordinate models at each higher dose was unsuitable has been presented in Section 2.5.3.

Final statistics for model refinements are given in Table 3.2. Coordinates and structure factor amplitudes have been deposited in the PDB with accession codes: *4x4b*, *4x4c*, *4x4d*, *4x4e*, *4x4f*, *4x4g*, *4x4h* and *4x4i*.

3.2.5 $F_{obs,n} - F_{obs,1}$ map analysis

Fourier difference maps ($F_{obs,n} - F_{obs,1}$, for $n = 2, ..., 8$) were calculated to locate sites of significant radiation-induced electron density change in the C.Esp1396I complex. These maps were calculated using the protocol described previously in Section 2.4.1.

Figure 3.6 provides a visual representation of the spatial distribution of $F_{obs,n} - F_{obs,1}$ map peaks across the C.Esp1396I-DNA complex at a range of reported doses. Multiple map thresholds (in units of σ) have been presented to indicate that the background map noise appears to become dominant (and mask true damage signal) above $\pm 3\sigma$ across the range of doses reported. Whereas Figure 3.6 alone can provide no information on whether each observable $F_{obs,n} - F_{obs,1}$ map peak corresponds to a true damage event or noise, nonetheless there is a evident partitioning of $F_{obs,n} - F_{obs,1}$ map density to the protein components over the DNA, implying that the protein component is subject to more radiation damage events.

3.2.5.1 Facilitated manual inspection with *PEAKS*

Since no previous MX site-specific damage investigations are known to exist on protein-nucleic acid crystals, a detailed manual inspection of the electron density was required at each dose.

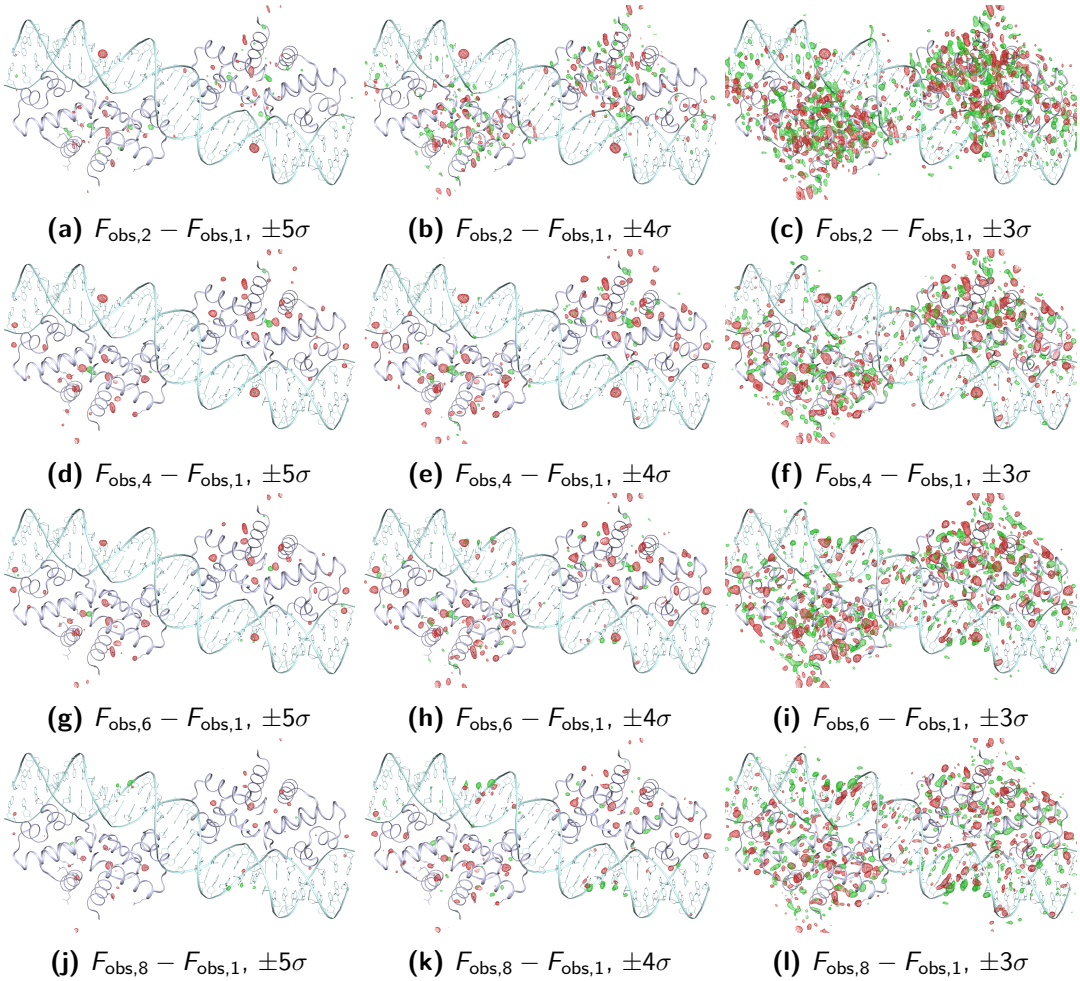
Due to the large size of the complex ($\sim$ 60 kDa) and the resulting large number of detected

Table 3.2: Data processing and refinement statistics for the C.Esp1396I-DNA damage series. Values in parentheses are for the highest-resolution shells. The Wilson *B*-factor was estimated from the Wilson plot. I_n/I_1 is the mean diffraction intensity I_n for dataset n (as defined by Equation 3.2.1), normalised to the mean for the first dataset. For observed F_{obs} and calculated F_{calc} structure factors, $R_{\text{work}} = \frac{\sum |F_{\text{obs}} - F_{\text{calc}}|}{\sum F_{\text{obs}}}$ and R_{free} is the R_{work} formula calculated from a small (5 %) test set of randomly selected reflections. Care was taken to ensure that the same test set of reflections was maintained for all datasets. Unit-cell angles in $P6_5$ are $\alpha = 90°$, $\beta = 90°$ and $\gamma = 120°$ throughout. The resolution range is 69.5 – 2.8 Å with 2.95 – 2.8 Å for the outer shell for all datasets.

Dataset	1	2	3	4	5	6	7	8
PDB code	4x4b	4x4c	4x4d	4x4e	4x4f	4x4g	4x4h	4x4i
Data processing								
Cell dimensions								
$a = b$ (Å)	104.29	104.38	104.33	104.43	104.45	104.45	104.33	104.43
c (Å)	139.18	139.26	139.20	139.21	139.19	139.09	139.01	139.08
No. of observations	128219	129102	128743	128973	128476	129426	126479	126715
No. of unique reflections	21172	21231	21198	21250	21207	21246	21052	21244
Multiplicity	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.0 (6.1)	6.0 (6.1)
Completeness (%)	99.9 (100.0)	99.8 (100.0)	99.8 (100.0)	99.9 (100.0)	99.7 (100.0)	99.8 (100.0)	99.4 (100.0)	99.9 (100.0)
Mosaicity (°)	0.70	0.69	0.68	0.68	0.68	0.68	0.71	0.78
Wilson *B*-factor (Å^2)	55.92	63.72	63.00	60.20	65.34	74.80	70.34	78.05
R_{merge}	0.05 (0.30)	0.05 (0.33)	0.05 (0.41)	0.05 (0.47)	0.05 (0.39)	0.06 (0.82)	0.06 (0.63)	0.08 (1.92)
R_{meas}	0.06 (0.40)	0.06 (0.44)	0.06 (0.50)	0.08 (0.60)	0.07 (0.53)	0.074 (0.989)	0.13 (0.82)	0.17 (2.65)
$CC_{1/2}$	0.994 (0.934)	0.999 (0.929)	0.998 (0.902)	0.996 (0.875)	0.997 (0.894)	0.998 (0.684)	0.974 (0.768)	0.982 (0.193)
$\langle I/\sigma(I) \rangle$	22.1 (5.0)	21.9 (4.7)	21.7 (4.2)	16.9 (3.5)	20.1 (3.9)	17.8 (2.2)	9.6 (2.6)	8.6 (0.8)
I_n/I_1	1.00	0.99	0.84	0.84	0.73	0.66	0.59	0.51
Refinement								
R_{work}	0.205	0.234	0.235	0.240	0.237	0.234	0.246	0.264
R_{free}	0.262	0.279	0.282	0.285	0.281	0.275	0.293	0.296
Mean isotropic *B*-factor (Å^2)	62.18	65.26	62.96	59.62	66.45	72.63	72.00	81.69

Table 3.3: Assessment of final macromolecular geometry for the refined coordinate model of the C.Esp1396I-DNA complex, corresponding to the first dataset, as reported by *Molprobity4* (Chen *et al.*, 2010). Since for the coordinate models corresponding to the higher dose datasets, only rigid-body and isotropic *B*-factor refinement was performed on the refined initial dataset coordinate model coupled with the higher dataset merged structure factor amplitudes, the reported geometry statistics were conserved for the higher dose datasets.

No. non-H atoms:	
Protein	2496
DNA	1429
Water	5
Other	$2\times$ Mg^{2+}
Bond parameters:	
RMSD bond length (Å)	0.011
RMSD bond angle (°)	1.338
Ramachandran analysis:	
Favoured (%)	93.0
Outliers (%)	6.7
Allowed (%)	0.3
Rotamer outliers (%)	3.45
All atom clash-score	10.91



(a) $F_{obs,2} - F_{obs,1}$, $\pm 5\sigma$ **(b)** $F_{obs,2} - F_{obs,1}$, $\pm 4\sigma$ **(c)** $F_{obs,2} - F_{obs,1}$, $\pm 3\sigma$

(d) $F_{obs,4} - F_{obs,1}$, $\pm 5\sigma$ **(e)** $F_{obs,4} - F_{obs,1}$, $\pm 4\sigma$ **(f)** $F_{obs,4} - F_{obs,1}$, $\pm 3\sigma$

(g) $F_{obs,6} - F_{obs,1}$, $\pm 5\sigma$ **(h)** $F_{obs,6} - F_{obs,1}$, $\pm 4\sigma$ **(i)** $F_{obs,6} - F_{obs,1}$, $\pm 3\sigma$

(j) $F_{obs,8} - F_{obs,1}$, $\pm 5\sigma$ **(k)** $F_{obs,8} - F_{obs,1}$, $\pm 4\sigma$ **(l)** $F_{obs,8} - F_{obs,1}$, $\pm 3\sigma$

Figure 3.6: The spatial distribution of $F_{obs,n} - F_{obs,1}$ map peaks over the full C.Esp1396I-DNA complex structure for four increasing dose states: (a)-(c) $F_{obs,2} - F_{obs,1}$, (d)-(f) $F_{obs,4} - F_{obs,1}$, (g)-(i) $F_{obs,6} - F_{obs,1}$, (j)-(l) $F_{obs,8} - F_{obs,1}$. In each row, the map contouring level has been shown at $\pm 5\sigma$, $\pm 4\sigma$ and $\pm 3\sigma$ to provide a qualitative indication for the map contouring level at which background noise becomes dominant.

difference map peaks requiring inspection at each dose[†], the *PEAKS* pipeline (introduced in Section 2.4) was used to facilitate a high-throughput and systematic map inspection at each consecutive dose state.

A lower bound map threshold needed to be specified in *PEAKS*, below which to disregard map peaks prior to manual inspection. Ideally, the threshold should be conserved between different $F_{obs,n} - F_{obs,1}$ maps in the series, in order to preserve consistency. Here, the threshold at which background map noise became dominant in the lowest difference map, $F_{obs,2} - F_{obs,1}$, was taken (corresponding to a value of $\pm 0.04\,e/Å^3$, or exactly $\pm 4\sigma$ for the $F_{obs,2} - F_{obs,1}$ map, as illustrated in Figure 3.6). The same threshold ($\pm 0.04\,e/Å^3$) was used for all later datasets, thus providing a 'worst-case-scenario' lower map threshold at higher doses, since the onset of undistinguishable map noise was typically far higher than $\pm 0.04\,e/Å^3$ (as an extreme example, for the $F_{obs,8} - F_{obs,1}$ map, $\pm 0.04\,e/Å^3$ corresponded to only $\pm 1.22\sigma$). The use of this conservative threshold ensured that no low-lying true damage sites (visible at only low map σ values) were neglected in higher dose datasets. As an unfortunate caveat, the use of such a low map threshold led to a large number of peaks being initially detected by *PEAKS* prior to manual inspection (for the $F_{obs,8} - F_{obs,1}$ map, 7811 peaks were detected initially, as shown in Figure 3.7).

All positive map peaks were excluded in *PEAKS* prior to manual inspection. A net electron density loss has typically been reported as a signature of both X-ray-induced bond rupture and atomic disordering events in proteins at 100 K (Fioravanti *et al.*, 2007; De la Mora *et al.*, 2011), and consequently positive peaks were not of current interest. Here, a custom distance threshold was specified in *PEAKS*, such that peaks further than 2 Å from any atom were immediately excluded. This specification overrides the stricter Van der Waals radii thresholds implemented in Section 2.4.3, the suitability of which have already been demonstrated for the *Dp*Cel7B protein test case in Section 2.4.5. The more relaxed 2 Å distance threshold has been used here to account for the low data resolution (2.8 Å), since there is less certainty that atoms have been correctly positioned. In good accordance with the previous *Dp*Cel7B protein test study (Section 2.4.5), 'by-eye' investigation of $F_{obs,n} - F_{obs,1}$ map peaks for the first two difference map datasets (for

[†] As an example of the impracticality of manual inspection, for the highest dose dataset (44.6 MGy), of the order of 10^3 peaks were detected in the $F_{obs,8} - F_{obs,1}$ map above a threshold of $\pm 3\sigma$ as shown previously in Figure 3.6l. $F_{obs,n} - F_{obs,1}$ map peaks have been detected here using the CCP4 program *PEAKMAX*. As discussed above, a $\pm 3\sigma$ map threshold appeared to be sufficient here as a lower bound for manual inspection of the $F_{obs,n} - F_{obs,1}$ map at 44.6 MGy, since difference map noise became visually dominant at this threshold, as shown in Figure 3.6.

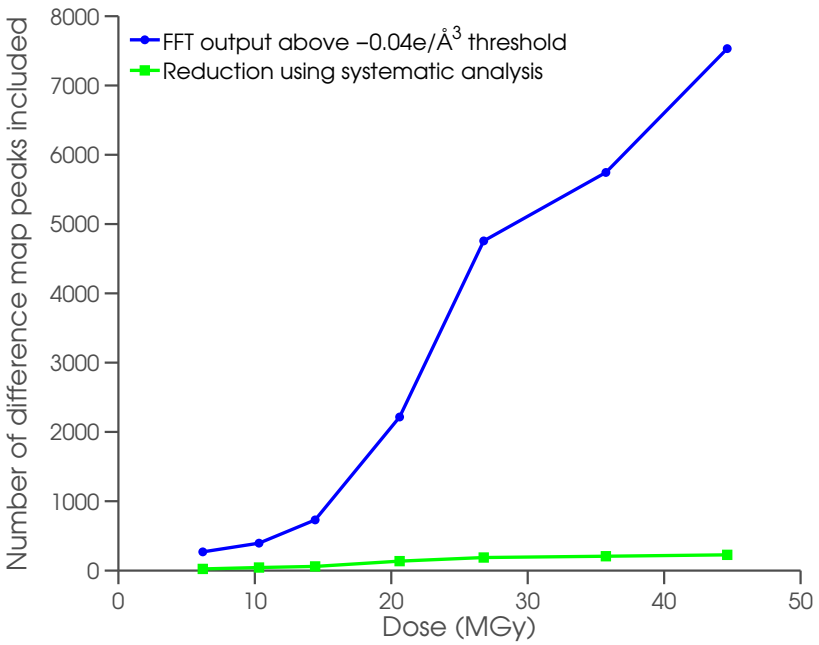


Figure 3.7: Reduction in frequency of peaks detected in $F_{obs,n} - F_{obs,1}$ maps corresponding to potential specific damage following facilitated manual inspection using the *PEAKS* pipeline presented in Section 2.4.

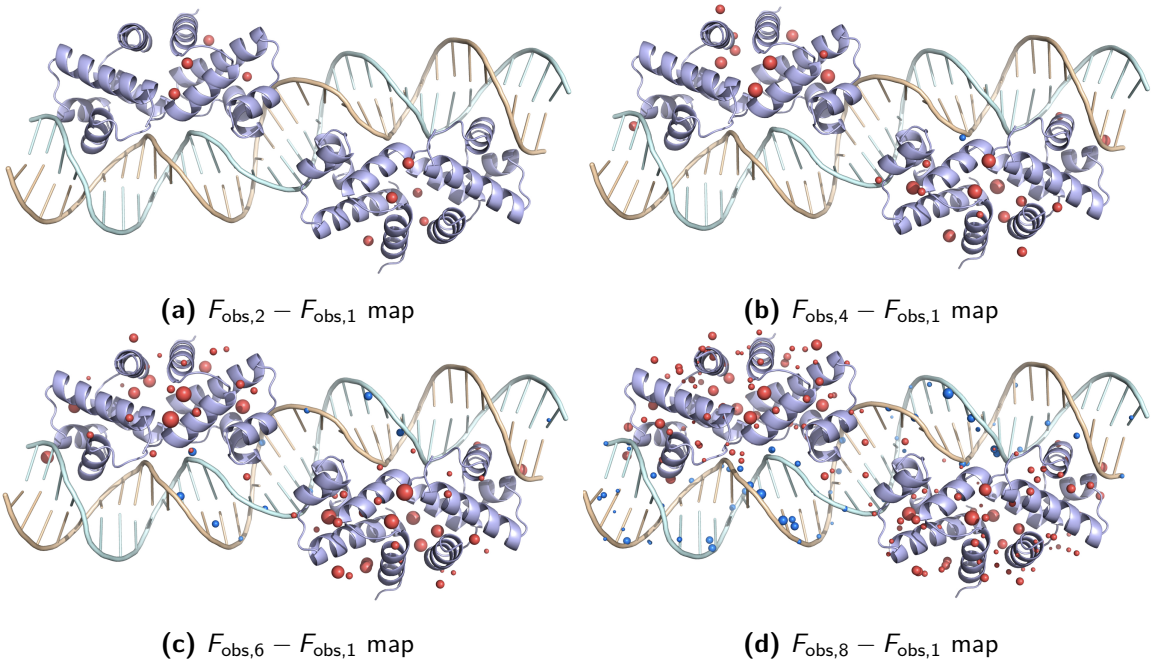
which the number of originally detected peaks was not too high) led to the conclusion that all the difference map peaks corresponding to clear specific damage were below the 2 Å threshold.

The overall reduction in $F_{obs,n} - F_{obs,1}$ peaks at each dose following facilitated manual inspection using *PEAKS* is presented in Figure 3.7. Figure 3.8 provides a visual representation of the spatial distribution of specific damage at each dose that were detected using *PEAKS*. As anticipated, with increasing dose, (a) the frequency of damage site increased, and (b) the magnitude of electron density loss increased at damage sites.

3.2.6 Site specific damage observations using $F_{obs,n} - F_{obs,1}$ maps

3.2.6.1 Protein site-specific damage

Sites of electron density loss consistent with previous reports for native proteins (Weik *et al.*, 2000; Ravelli and McSweeney, 2000; Fioravanti *et al.*, 2007) were detectable by eye above background map noise, even in the lowest dose $F_{obs,2} - F_{obs,1}$ map (Figure 3.8). In particular, multiple instances were located of electron density loss at acidic residue side-chain carboxyl groups (a widely regarded signature of decarboxylation) and around methylthio side-chain groups (indicative



(a) $F_{obs,2} - F_{obs,1}$ map **(b)** $F_{obs,4} - F_{obs,1}$ map

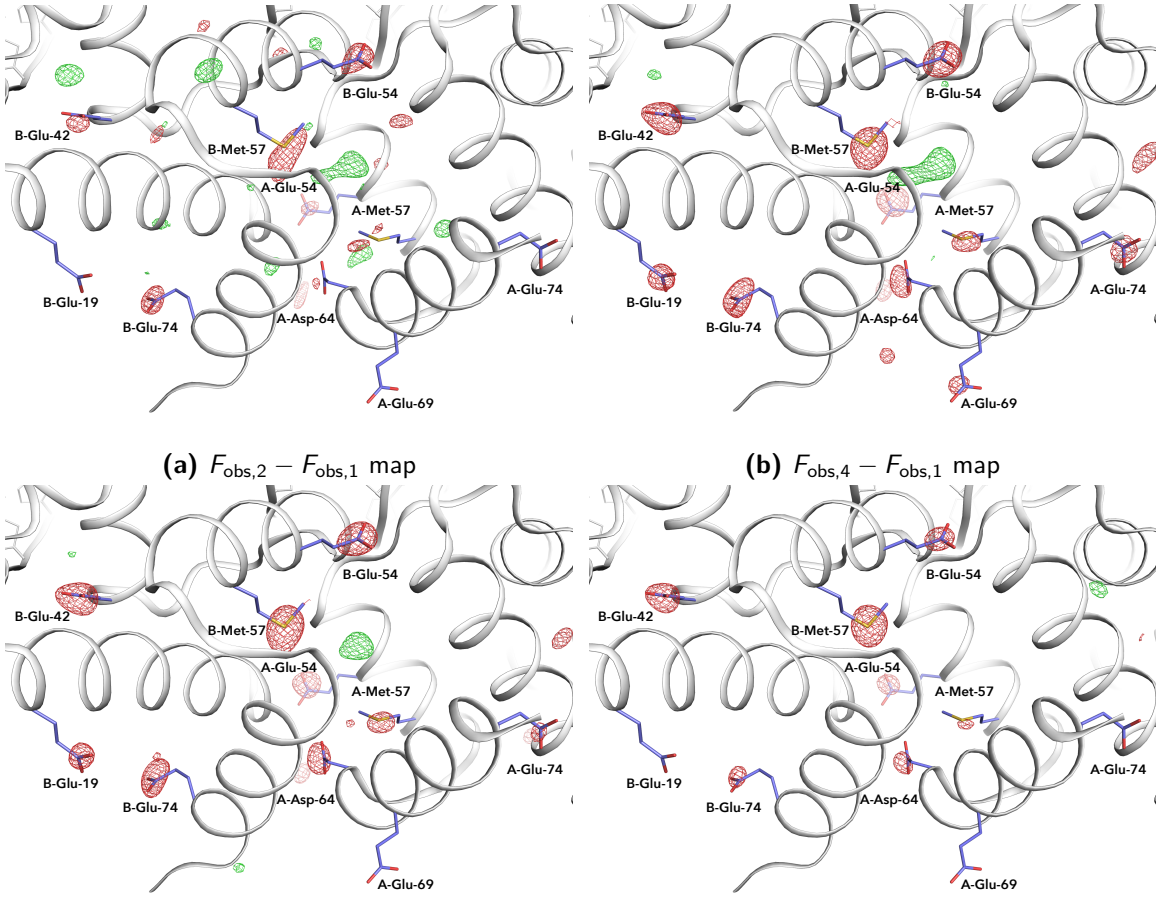
(c) $F_{obs,6} - F_{obs,1}$ map **(d)** $F_{obs,8} - F_{obs,1}$ map

Figure 3.8: The spatial distribution of specific damage sites detected using *PEAKS* throughout the C.Esp1396I-DNA complex derived from $F_{obs,n} - F_{obs,1}$ Fourier difference maps for (a) $n = 2$, (b) $n = 4$, (c) $n = 6$ and (d) $n = 8$. Specific damage sites are represented as spheres, with radii proportional to electron density loss (in e/Å^3). Spheres positioned closer/further than 2 Å to/from the DNA strands are coloured blue/red.

of either side-chain disordering or Met CH_3–S group cleavage). It is noted that the C-protein contains no Cys residues to form disulphide bonds, which are also highly radiation-sensitive in MX. As illustrated in Figure 3.8a, each C-protein dimer interface appears particularly prone to such radiation damage effects. Figure 3.9 provides a representative example of protein damage, here at the dimer interface between chains A and B.

Figure 3.10 compiles the results of manual map inspection across the whole $\sim$ 60 kDa C.Esp1396I-DNA complex. At higher doses ($\geq$ 14.4 MGy), negative difference density was also detectable around many additional residue types: in particular at the main-chain carboxyl group oxygen of Arg and Asn, the side-chain hydroxyl group of Ser the side-chain Cδ atoms of Ile and Val, and the lysyl side-chain group of Lys. The latter observations appear to be consistent with increased side-chain disordering, as opposed to any bond cleavage events.

No significant Tyr residue damage was observable across the wide dose range. This is in strong accordance with the results of Chapter 6, in which it will be concluded that Tyr is not highly prone to radiation damage across MX protein damage series deposited previously in the PDB, contrary to previous reports (Burmeister, 2000). Overall, it was concluded that the protein was



(a) $F_{\mathrm{obs},2} - F_{\mathrm{obs},1}$ map

(b) $F_{\mathrm{obs},4} - F_{\mathrm{obs},1}$ map

(c) $F_{\mathrm{obs},6} - F_{\mathrm{obs},1}$ map

(d) $F_{\mathrm{obs},8} - F_{\mathrm{obs},1}$ map

Figure 3.9: Examples of protein damage sites in the C.Esp1396I-DNA complex detectable in $F_{\mathrm{obs,n}} - F_{\mathrm{obs},1}$ Fourier difference maps, for $n = 2, 4, 6, 8$. Site-specific damage to Glu, Asp and Met residues at the protein-protein binding interface between chains A and B is shown. All Fourier difference maps are contoured at $\pm 5.0\sigma$ in green/red, such that difference map density has been deemed to exceed background noise. Only side-chains corresponding to specific damage sites have been included in the rendering.

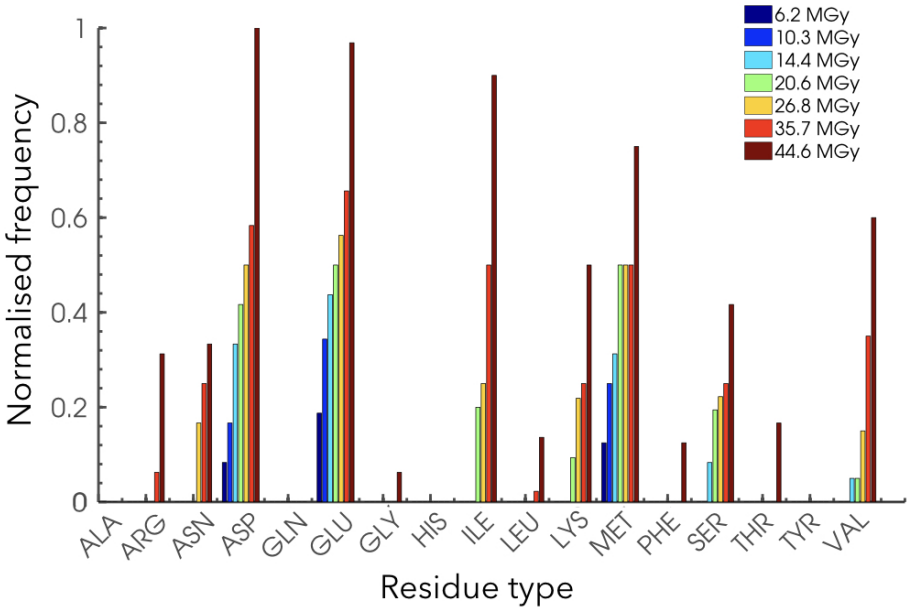


Figure 3.10: Normalised frequency of specific damage against protein residue type over chains A to D for each dose in MGy. Frequencies have been normalised to the number of occurrences of each residue in the structure, since the frequency of damage sites is heavily dependent on the overall number of available damage sites for each residue or nucleotide present. Since the C.Esp1396I protein contains no Cys, Pro or Trp residues, these labels have been omitted.

still highly susceptible to the effects of MX specific damage even when involved in the interaction with nucleic acid.

3.2.6.2 DNA site-specific damage

The observations of MX protein damage presented above indicate that despite the relatively low resolution of the diffraction data (2.8 Å), the generated $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps were of sufficient quality to probe damage effects to the DNA component of the complex.

As Figure 3.8 indicates, during manual map inspection, fewer difference map peaks were categorised as corresponding to potential sites of specific damage for the DNA component across the reported dose range and above the background noise level in each map. It was noted that two large negative difference density peaks were clearly visible, even in the $F_{\text{obs,2}} - F_{\text{obs,1}}$ map, and above $\pm 5\sigma$ in that map (Figure 3.6a); however these peaks have been omitted from the results in Figure 3.8. These two peaks were both located in the solvent region, at approximately 2.1 Å from the C8 atom of the adenine base at position 10 in both DNA chains E and F, such that

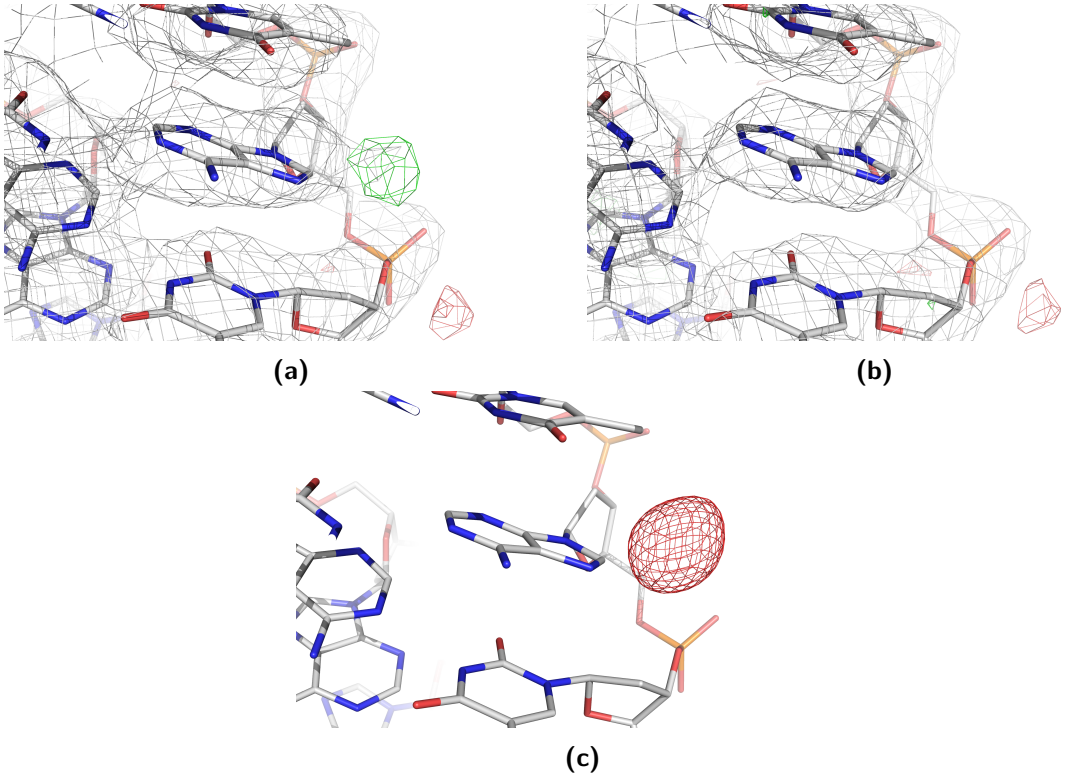


Figure 3.11: Illustration of the unexplained $F_{\text{obs,n}} - F_{\text{obs,1}}$ map negative density near the C8 atom of A-10 in the C.Esp1396I-DNA complex, rendered in *PYMOL*. Only chain E is shown, however chain F is qualitatively similar. In (a) and (b), the $2\text{m}F_{\text{obs,n}} - \text{D}F_{\text{calc}}$ map (grey, contoured at 1.5σ) and $\text{m}F_{\text{obs,n}} - \text{D}F_{\text{calc}}$ map (green/red, contoured at $\pm 3\sigma$) indicate the presence of unexplained positive $\text{m}F_{\text{obs,n}} - \text{D}F_{\text{calc}}$ map approximately 2.1 Å from the C8 atom of adenine for the first dataset (a) $n = 1$, but not the second dataset (b) $n = 2$. In (c) the $F_{\text{obs,2}} - F_{\text{obs,1}}$ map (green/red, contoured at $\pm 5\sigma$) indicates the presence of significant electron density loss at the region where the positive $\text{m}F_{\text{obs,1}} - \text{D}F_{\text{calc}}$ map density was observed in (a). It is postulated that an ion may have coordinated the DNA at the onset of the experiment, yet was lost by the point at which the dose reached 6.19 MGy.

they are symmetrically related about the pseudo 180° near-NCS about the centre of the 35 bp DNA sequence. However, due to the space group symmetry, these peaks were actually much closer since C8 of A10 (chain E) is only 15 Å from the C8 of A10 (chain F) across the crystal contact. As illustrated in Figure 3.11, $2\text{m}F_{\text{obs}} - \text{D}F_{\text{calc}}$ map density for the first dataset (pdb code: *4x4b*) indicated a small extension of the electron density from the C8 atom of each A-10 base, which was not detectable in the later dataset $2\text{m}F_{\text{obs}} - \text{D}F_{\text{calc}}$ maps. There was no indication that these adenine base groups had been modified (J. McGeehan, personal communications), and consequently it has not been possible to associate these peaks with any clear specific damage event to any refined atom built into the C.Esp1396I-DNA complex structure.

At higher doses ($\geq$ 20.6 MGy) there are several locations of suspected site-specific damage localised on the DNA (Figure 3.12). Most distinguishable was electron density loss visible in the vicinity of the phosphodiester bonds comprising the DNA backbone for a subset of nucleotides.

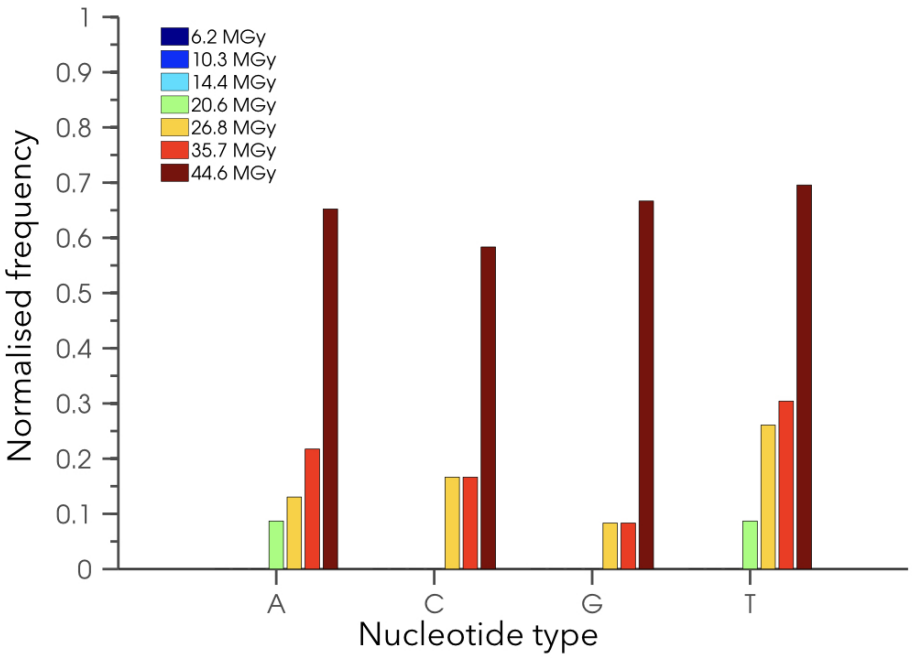
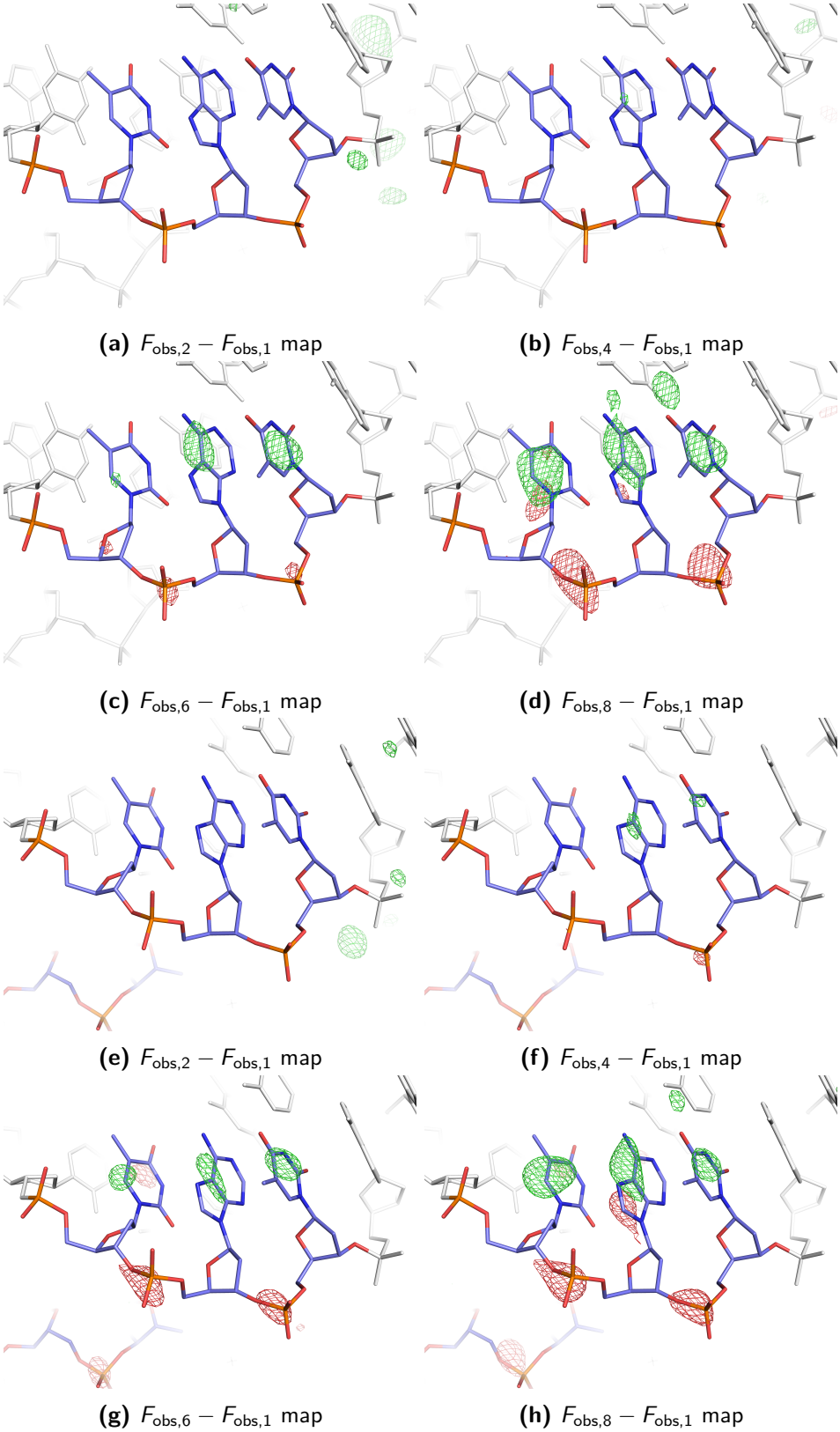


Figure 3.12: Normalised frequency of specific damage against DNA nucleotide type over chains E and F for each dose in MGy. Frequencies have been normalised to the number of occurrences of each nucleotide in the structure, since the frequency of damage sites is heavily dependent on the overall number of available damage sites for each residue or nucleotide present.

Figure 3.13 provides an example of the onset of negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density surrounding the sugar-phosphate C–O bond between two nucleotides (T-24 and A-25 on both chains E and F) that are both at a TATA site between palindromic protein-recognition sequences on the DNA. At lower tested doses, no such radiation-induced changes were observable (Figure 3.13a-b). This could be an initial indication of bond cleavage at this position, which would ultimately result in a single-strand break (SSB).

In addition, positive $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density accumulation was observed proximal to the T-24 and A-25 bases (Figure 3.13c-g). This could be consistent with a mechanism of SSB formation that has previously been proposed in the radiation chemistry literature: the nucleobase first provides a site for LEE attachment, which is followed by transfer of the electron to the anti-phosphate group, leading to rupture of the sugar-phosphate C–O σ bond through dissociative electron attachment (DEA) (Simons, 2006). However, the formation of a transient anion at the base is predicted to have a very short lifetime on the scale of ps – ns at both room and cryogenic temperatures (Sanche, 2005), and it is thus unlikely that it would be a sustained feature visible in a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map. Furthermore, experiments providing evidence for DEA-mediated backbone



(a) $F_{obs,2} - F_{obs,1}$ map **(b)** $F_{obs,4} - F_{obs,1}$ map

(c) $F_{obs,6} - F_{obs,1}$ map **(d)** $F_{obs,8} - F_{obs,1}$ map

(e) $F_{obs,2} - F_{obs,1}$ map **(f)** $F_{obs,4} - F_{obs,1}$ map

(g) $F_{obs,6} - F_{obs,1}$ map **(h)** $F_{obs,8} - F_{obs,1}$ map

Figure 3.13: DNA damage sites in the C.Esp1396I-DNA complex detectable in $F_{obs,n} - F_{obs,1}$ Fourier difference maps, for $n = 2, 4, 6, 8$. Nucleotides T-24, A-25 and T-26 are shown (highlighted in purple) from left to right in each sub-figure for (a)-(d) chain E and (e)-(h) chain F. All Fourier difference maps are contoured at $\pm 4.0\sigma$ in green/red.

cleavage have predominantly been conducted for oligonucleotides in the gas phase, and it is unknown whether a transient anion at the basic subunit could persist long enough under collision with vibrating solvent atoms in a crystalline environment (or aqueous solution) to be transferred to the phosphate unit (I. Carmichael, private communications).

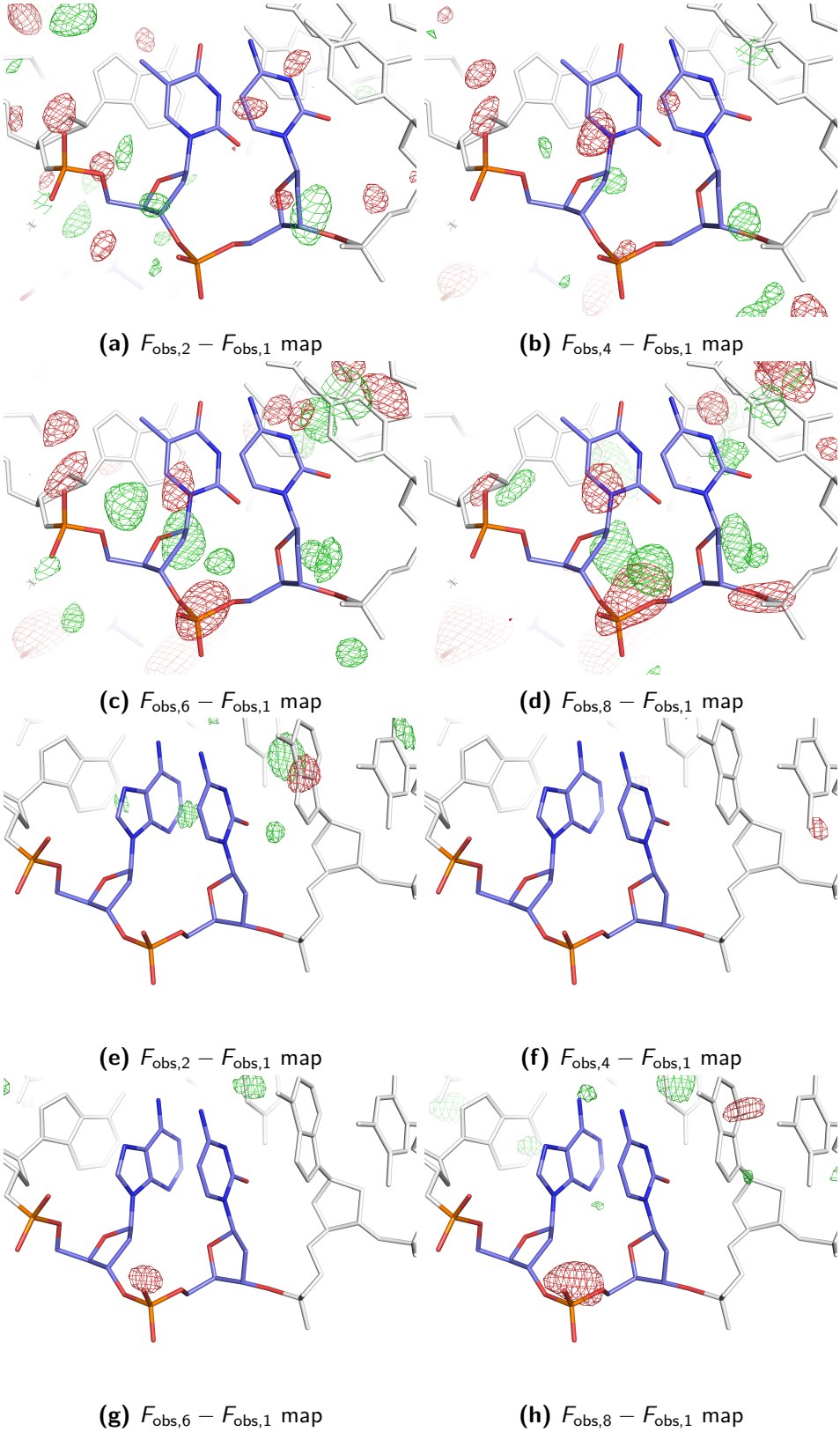
Alternatively, base modification induced by the interaction of close-proximity free radicals from surrounding solvent has also been reported in aqueous studies on nucleic acids (Cadet *et al.*, 1999). Whereas the possibility of such radical interactions cannot be excluded (e.g. hydroxyl radical binding to carbon C6 in T-24), it is widely-anticipated that such radical species should be immobile in the crystal at 100 K (Owen *et al.*, 2012). An alternative hypothesis proposed here is that cleavage of the phosphate backbone permits reorientation or dislocation of the basic subunit and results in a net electron density increase in the vicinity of the nucleobase. Ultimately, the low resolution of the current structure (2.8 Å) limited any conclusion that could be drawn regarding the cause of the positive $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map density.

Figure 3.14 provides additional examples of the onset of electron density loss surrounding the phosphate backbone of nucleotides in chain E at high doses ($>$ 20 MGy). These observations have only been made at lower map contouring levels near the level of background map noise ($\pm 3\sigma$ compared to $\pm 4\sigma$ in Figure 3.13), and therefore provide less convincing evidence for site-specific DNA damage events. The majority of nucleotides present in both chains E and F did not, however, exhibit difference density that could be distinguishable as a bond cleavage event, even with the map contouring threshold set as low as $\pm 3\sigma$.

3.2.6.3 Quantifying the relative damage susceptibilities of protein and DNA

Upon comparison of Figure 3.10 and 3.13, it was evident that the dose for onset of damage that was visually detectable in $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ maps was significantly higher for DNA than for protein ($>$ 20.6 MGy versus $>$ 6.2 MGy, relative to the first dataset at 2.07 MGy). The following work aimed to quantify exactly how robust the DNA component was to X-ray-induced damage compared with the protein.

The program *RIDL* (see Section 2.5) was used to calculate per-atom metrics from each $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map, in order to quantify the radiation-induced electron density loss for each individual



(a) $F_{obs,2} - F_{obs,1}$ map **(b)** $F_{obs,4} - F_{obs,1}$ map

(c) $F_{obs,6} - F_{obs,1}$ map **(d)** $F_{obs,8} - F_{obs,1}$ map

(e) $F_{obs,2} - F_{obs,1}$ map **(f)** $F_{obs,4} - F_{obs,1}$ map

(g) $F_{obs,6} - F_{obs,1}$ map **(h)** $F_{obs,8} - F_{obs,1}$ map

Figure 3.14: Additional potential DNA damage sites in the C.Esp1396I-DNA complex detectable in $F_{obs,n} - F_{obs,1}$ Fourier difference maps for $n = 2, 4, 6, 8$, positioned at (a)-(d) nucleotides T-14 and C-15 of chain E, and (e)-(h) A-32 and C-33 of chain E. Nucleotides of interest have been highlighted in purple. Here, all Fourier difference maps are contoured at $\pm 3.0\sigma$ in green/red, and are thus much closer to the background map noise at each dose.

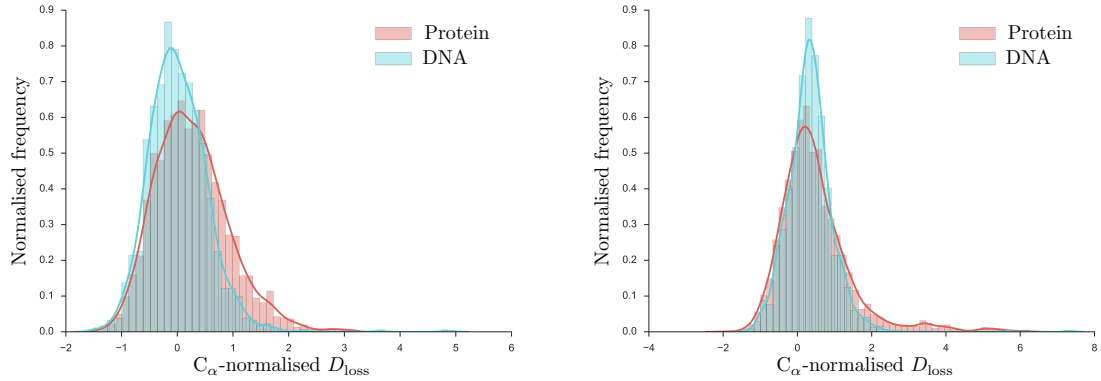
DNA atom in the C.Esp1396I-DNA complex. This provides an objective measure against which the results of manual inspection in Figure 3.8 could be compared. The data in Figure 3.15 provide a direct comparison of the distribution of C_α-normalised D_{loss}(atom) metric values attained by the protein and DNA subunits of the C.Esp1396I system. D_{loss}(atom) metric values for both the protein and DNA have been normalised to the same fixed standard at each dose: the metric values attained by the set of protein C_α backbone atoms, as described in Section 2.5.8.

For each distribution, a right-hand skewed tail (or more generally, the asymmetry about $D_{\text{loss}}^{C_\alpha} = 0$) provided a quantitative measure for the radiation-sensitivity of both protein and DNA at a given dose. Table 3.4 provides a summary of statistics characterising each distribution at each dose. It appears that *skewness* itself is a poor measure of radiation-sensitivity, since it measures asymmetry about the mean, not $D_{\text{loss}}^{C_\alpha} = 0$. This is clearest for the $F_{\text{obs},2} - F_{\text{obs},1}$ map statistics listed in Table 3.4, where the DNA attains a higher skewness value than protein, however the mean $D_{\text{loss}}^{C_\alpha}$(atom) value for protein is considerably higher.

Asymmetry about $D_{\text{loss}}^{C_\alpha} = 0$ could be better quantified by finding the ratio of the area under each distribution between the left and right of $D_{\text{loss}}^{C_\alpha} = 0$. In practice each underlying distribution is unknown, but this ratio can be approximated by calculating the sum of $D_{\text{loss}}^{C_\alpha}$ metric values over all atoms attaining $D_{\text{loss}}^{C_\alpha}(\text{atom}) > 0$ and $D_{\text{loss}}^{C_\alpha}(\text{atom}) < 0$ (denoted as $\Sigma_\uparrow$ and $\Sigma_\downarrow$ respectively) and then calculating the fraction $\Sigma_\uparrow/\Sigma_\downarrow$. This is included as the final column in Table 3.4. For a perfectly symmetrical distribution of values, a asymmetry score of $\Sigma_\uparrow/\Sigma_\downarrow = 1$ would be expected. This *asymmetry score* will be used here as a measure of radiation-sensitivity of each component.

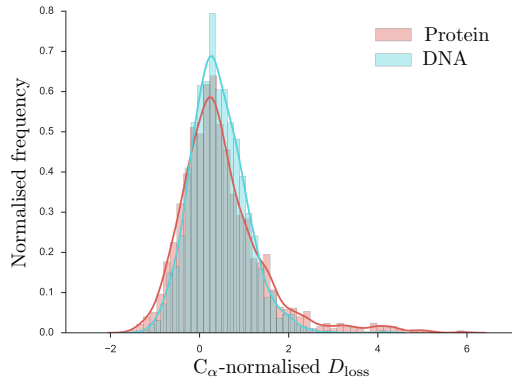
The high asymmetry for the protein component achieved already in the $F_{\text{obs},2} - F_{\text{obs},1}$ map (asymmetry score: 2.78) shown in Figure 3.15a was unsurprising, and was consistent with the observations of protein specific damage presented in Figure 3.9a. In contrast, the DNA component distribution is approximately symmetric about 0 for the $F_{\text{obs},2} - F_{\text{obs},1}$ map (asymmetry score: 0.93). As the data in Figure 3.16 indicate, the asymmetry score for the DNA increased with increasing dose, eventually becoming higher than the protein component (asymmetry score: 7.33 corresponding to the $F_{\text{obs},8} - F_{\text{obs},1}$ map).

Interestingly, the data in Table 3.4 suggest that the DNA was eventually more damaged than the protein component by the final $F_{\text{obs},8} - F_{\text{obs},1}$ map. This result is consistent with the fact

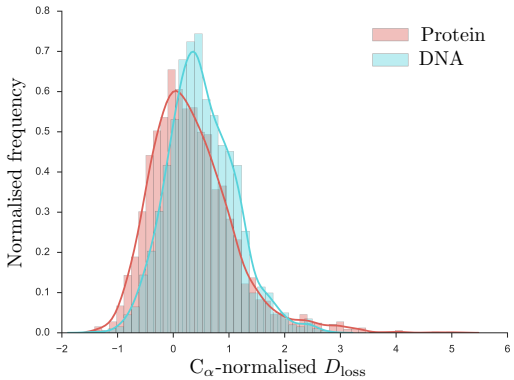


(a) $F_{obs,2} - F_{obs,1}$ map

(b) $F_{obs,4} - F_{obs,1}$ map



(c) $F_{obs,6} - F_{obs,1}$ map



(d) $F_{obs,8} - F_{obs,1}$ map

Figure 3.15: Histogram plots of C_α-normalised D_{loss} values, for protein and DNA components of the C.Esp1396I-DNA complex. Metric values were calculated by the program *RIDL*, as described in Section 2.5, to quantify electron density loss to each atom as detectable in $F_{obs,n} - F_{obs,1}$ Fourier difference maps, for (a) $n = 2$, (b) $n = 4$, (c) $n = 6$ and (d) $n = 8$. D_{loss} metric values for both the protein and DNA atoms have been normalised to the metric values attained by the set of protein C_α backbone atoms at each dose. Each histogram includes the contribution of all refined protein/DNA atoms in the coordinate model. Each histogram is overlaid with a smoothed kernel density estimate (KDE) of the underlying distribution that the metric values for each component obey. Each y-axis is normalised, such that for each histogram, the summed area under all bins is equal to 1. This has been performed to aid comparison between the distribution of metric values attained by the protein and DNA components, despite there being fewer DNA atoms present.

Table 3.4: For each $F_{obs,n} - F_{obs,1}$ Fourier difference map, statistical properties associated with the distribution of C_α-normalised D_{loss} values have been calculated over all atoms in the protein and DNA components of the C.Esp1396I-DNA complex. The *NumPy* and *SciPy* packages in *Python* were used to calculate the mean, standard deviation, skewness and kurtosis values for each distribution. The *asymmetry about 0* score was calculated using a custom script in *Python*, by finding the cumulative sums of $D_{loss}^{C_\alpha}$ metric values above/below $D_{loss}^{C_\alpha} = 0$ (denoted as $\Sigma_\uparrow$ and $\Sigma_\downarrow$ respectively) and then calculating the ratio $\Sigma_\uparrow/\Sigma_\downarrow$.

Map	Component	Mean	Standard Deviation	Skewness	Kurtosis	Asymmetry about 0
$F_{obs,2} - F_{obs,1}$	Protein	0.24	0.64	0.65	0.77	2.78
	DNA	-0.02	0.51	1.19	7.87	0.93
$F_{obs,3} - F_{obs,1}$	Protein	0.32	0.73	1.07	2.83	3.36
	DNA	0.18	0.58	0.89	5.77	2.21
$F_{obs,4} - F_{obs,1}$	Protein	0.48	0.99	1.91	5.96	4.88
	DNA	0.3	0.61	1.73	16.86	4.0
$F_{obs,5} - F_{obs,1}$	Protein	0.4	0.86	1.63	5.09	4.09
	DNA	0.25	0.58	0.7	3.11	3.19
$F_{obs,6} - F_{obs,1}$	Protein	0.4	0.84	1.68	4.77	4.56
	DNA	0.3	0.54	0.6	1.43	4.32
$F_{obs,7} - F_{obs,1}$	Protein	0.33	0.8	1.52	4.25	3.49
	DNA	0.36	0.57	0.42	0.63	5.35
$F_{obs,8} - F_{obs,1}$	Protein	0.27	0.65	1.15	3.03	3.32
	DNA	0.39	0.52	0.39	0.39	7.33

that each phosphate subunit (where damage manifested) comprises a significant proportion of the total DNA composition, whereas the protein damage hotspots (e.g. Glu/Asp carboxyl groups) do not for the protein.

3.2.6.4 Preferential damage to compressed regions of DNA

The above section has indicated that whereas DNA is less susceptible to clear radiation damage effects than protein, at high doses (typically above 20 MGy) radiation-induced changes to local electron density can be observed to both the phosphate backbone and nucleobases of specific nucleotides above the background noise threshold of $F_{obs,n} - F_{obs,1}$ maps.

Figure 3.17 provides a visual representation of the C_α-normalised D_{loss}(atom) metric values (as calculated by the program *RIDL*, Section 2.5) attained over the C.Esp1396I-DNA complex across the reported dose range. Consistent with the visual observations made in Figure 3.13 above, two distinct regions of increased radiation-susceptibility are present at the phosphate backbone around

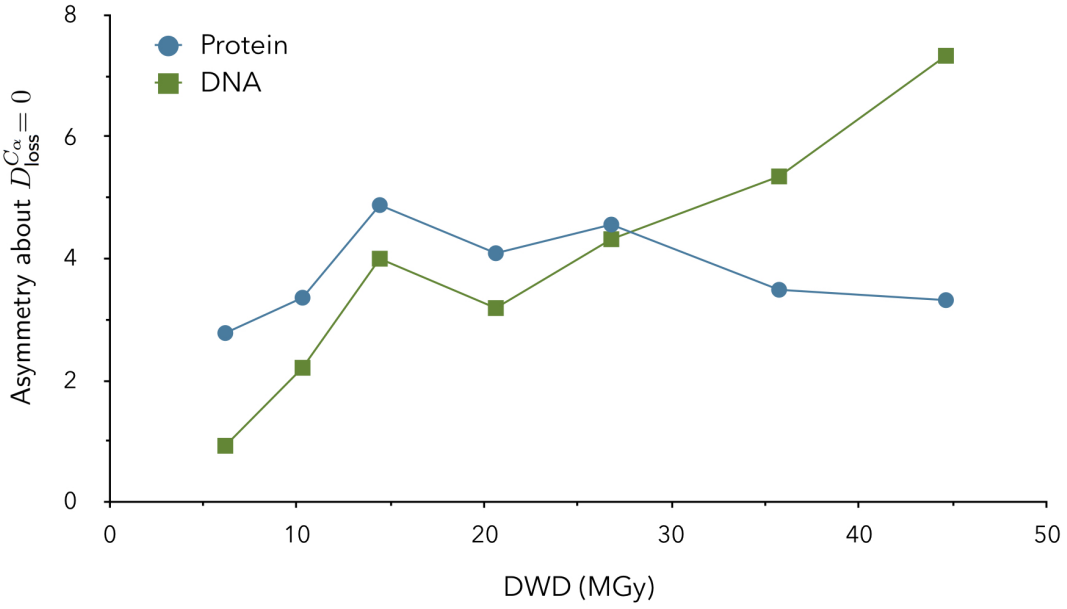
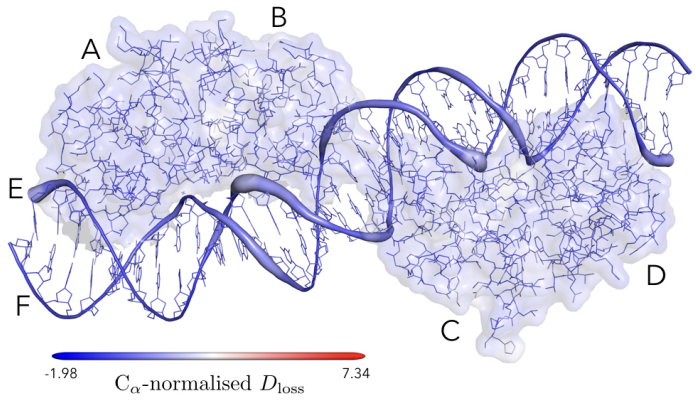


Figure 3.16: The *asymmetry about 0* score calculated for the C_α-normalised D_{loss} metric distributions in Figure 3.15 as a function of increasing dose. Asymmetry was calculated using the formula defined in the caption of Table 3.4.

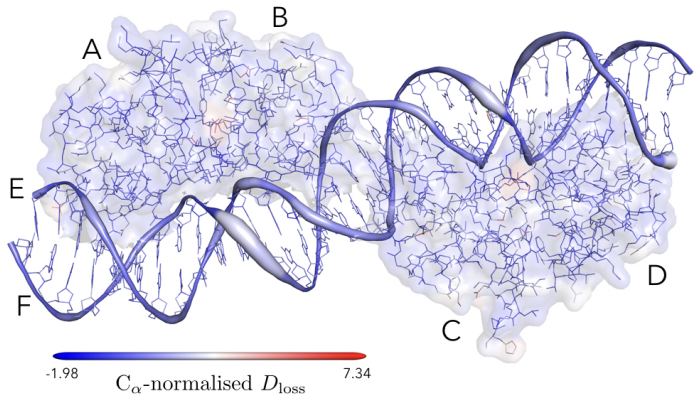
the two TATA sequences (T-24 to A-27 on chains E and F). TATA sequences have been reported to be easily deformable (Dickerson, 1998) in protein-DNA complexes, and as Figure 3.18 indicates, the minor groove is significantly compressed at each of the current TATA regions (to ~ 2 Å). This observation is in strong agreement with previous structural studies that have suggested that the minor groove compression at each TATA site is required to facilitate C.Esp1396I protein binding (McGeehan *et al.*, 2008). It is thus suggested that DNA backbone strain may contribute to the rate at which bond cleavage events can occur at the phosphate group. Indeed, strained cofactor binding sites in proteins have previously been reported to be particularly susceptible to X-ray-induced damage and undergo shifts to relaxed conformations during irradiation (Dubnovitsky *et al.*, 2005).

3.2.7 Mean isotropic *B*-factor analysis

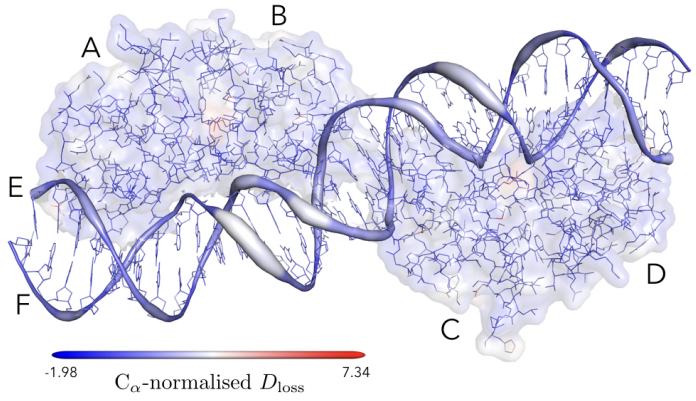
Site-specific damage events can also be detectable by monitoring the change in atomic isotropic *B*-factors with increasing dose (Weik *et al.*, 2000), and this metric provides a useful consistency check to observations made in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps. The data in Figure 3.19 illustrate the variation in the normalised isotropic *B*-factor (averaged over all refined atoms in each chain) against dose.



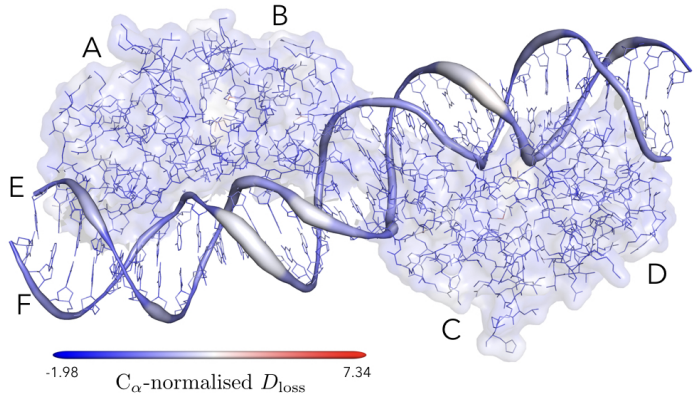
(a) $F_{obs,2} - F_{obs,1}$ map



(b) $F_{obs,4} - F_{obs,1}$ map



(c) $F_{obs,6} - F_{obs,1}$ map



(d) $F_{obs,8} - F_{obs,1}$ map

Figure 3.17: Illustration of the distribution of C_α-normalised D_{loss}(atom) values (see Section 2.5.8 for construction of the metric) across the C.Esp1396I-DNA complex. Metric values have been calculated from $F_{obs,n} - F_{obs,1}$ difference maps at (a) $n = 2$, (b) $n = 4$, (c) $n = 6$ and (d) $n = 8$. Each atom has been shown in *stick representation* using *PYMOL*, with a fixed colour palette ranging between the lowest (-1.98) and highest (7.34) $D_{loss}^{C_\alpha}$ attained across the structure over all displayed datasets. The distribution of $D_{loss}^{C_\alpha}$ values attained by the DNA phosphate backbone has additionally been displayed using the *B-factor putty* representation in *PYMOL* (with width of putty proportional to the $D_{loss}^{C_\alpha}$ metric value at a site), in order to indicate regions of the backbone that are preferentially damaged.

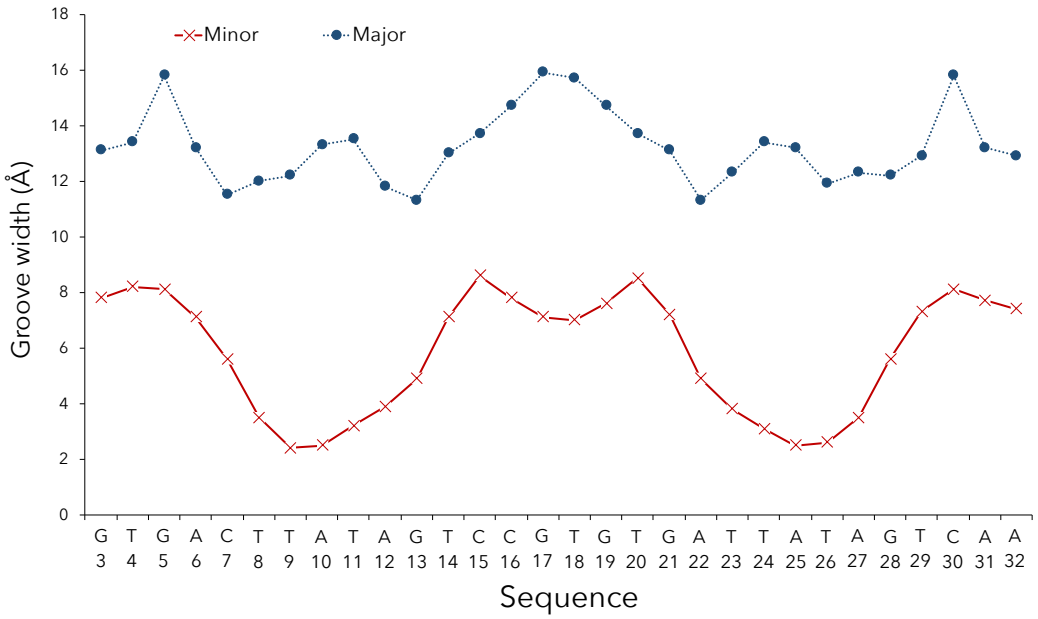


Figure 3.18: The major and minor groove widths (in Å) are shown as a function of position along the 35 bp DNA sequence in the C.Esp1396I-DNA complex. Widths have been calculated with the *CURVES+* online server (Lavery *et al.*, 2009), as the distance between backbone P atoms across each groove. As specified in Lavery *et al.* (2009), to make groove widths more relevant measure for the space available within a helical groove, all widths have been reduced by 2×2.9 Å to account for the average Van der Waals radius of each backbone.

Here the average *B*-factor for each individual protein/DNA chain has been normalised to its value attained at the lowest dose reported (2.07 MGy).

A least-squares linear fit for each nucleoprotein chain in Figure 3.19 indicates a larger rate of normalised average *B*-factor increase for each protein monomer compared to the DNA strands (A: 0.0081 Å MGy^{-1}, B: 0.0102 Å MGy^{-1}, C: 0.0105 Å MGy^{-1}, D: 0.0089 Å MGy^{-1}, and 0.0041 Å MGy^{-1} for both DNA chains E and F). The calculated 90 % confidence intervals indicate that the relative radiation-insensitivity of DNA is statistically significant, in strong agreement with the $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map analysis presented above in Section 3.2.5. Also of note is the fact that the average isotropic *B*-factor dose-dynamics are most comparable between symmetrically-related protein chains (A and D, and B and C) and also the two DNA strands (both attaining 0.0041 Å MGy^{-1}), as a possible manifestation of the rotational near-NCS around the centre of the DNA 35 bp sequence.

Overall these data are highly consistent with the observations made above in Section 3.2.5.

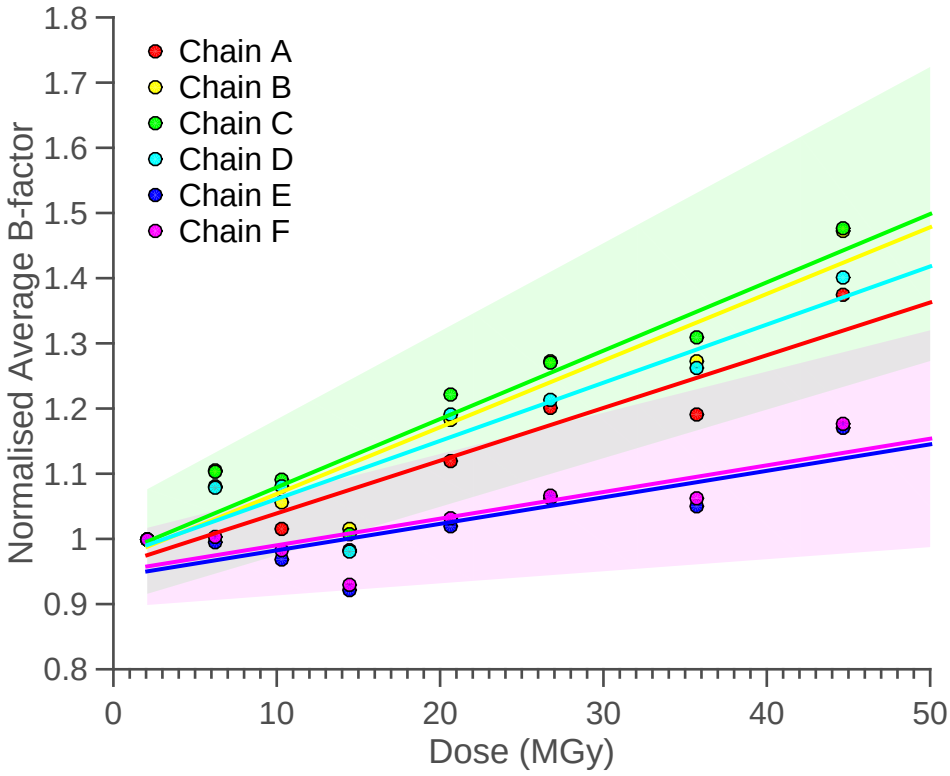


Figure 3.19: The dose dependence of the mean isotropic *B*-factor (in units of $Å^2$) for each chain of the C.Esp1396I-DNA complex (A to D for protein monomers, E and F for the DNA single strands). Each value is normalised to the average *B*-factor attained by each chain in the first dataset for each chain. For each chain, a least-squares linear fit has been performed in *MATLAB*. Green and purple shaded areas denote 90 % confidence intervals on both the intercepts and slopes for chains C and F, respectively (for clarity, the confidence intervals for the other chains have not been included).

3.3 The C.Csp 21 bp DNA complex

The aim of the current section is to address the question of whether the observations of phosphate backbone electron density loss observed for the C.Esp1396I-DNA protein complex in Section 3.2 are reproducible in another protein-DNA complex system. Additionally, it will be determined whether these manifestations again also occur at higher doses than does classical protein site-specific damage.

Here, another C-protein DNA complex has been investigated that also regulates restriction endonuclease expression in a bacterial R-M system: the C.Csp231I system, the structure of which has recently been solved by MX (Shevtsov *et al.*, 2015). Sequence alignment using the Clustal Omega online server[‡] indicated that the C.Esp1396I and C.Csp231I proteins had a low sequence similarity score of only 20.83 %; however these systems share similarities. C.Csp231I also consists of two protein dimers that bind to two operator sites, O_L and O_R. However, whereas in C.Esp1396I the O_L and O_R sites are separated by only 4 bp, in C.Csp231I, this distance is much larger at 18 bp (approximately 60 Å apart). Furthermore, each C.Csp231I protein dimer binds to a palindromic CTAAGN_5CTTAG recognition sequence at each operator, with a larger 5 nucleotide spacer region separating the binding sites of each monomer to the DNA.

A MX damage series has been collected on a single crystal of a C.Csp231I protein dimer bound to the 21 bp dsDNA sequence constituting of only the O_L operator complex. This exact same crystal was used to solve the original structure of the C.Csp231I-O_L complex (PDB: *4jcx* at 2.3 Å) as reported in Shevtsov *et al.* (2015).

3.3.1 Expression, purification and crystallisation

All expression, purification and crystallisation of the C.Csp231I-O_L complex was performed by M. Shevtsov and J. McGeehan, University of Portsmouth. Since this is the identical crystal to that used in Shevtsov *et al.* (2015) and is not the current author's work, the reader is referred to that publication. Of interest to the current investigation, the crystallisation buffer contained 50 mM Tris−HCl pH 8.0, 0.1 M NaCl and 1 mM Na_2EDTA, all of which would have contributed

[‡] *http://www.ebi.ac.uk/Tools/msa/clustalo/*

to the X-ray absorption cross-section of the crystal, and were thus important to factor when calculating the dose absorbed by the crystal. The crystal was cryoprotected by transfer into a solution containing the same components as the well solution with an increase of $3 - 5\,\%$ in the precipitant concentration and the addition of $15 - 20\,\%$ (v/v) glycerol. The crystal was then mounted on a cryoloop and cryocooled in liquid nitrogen.

3.3.2 X-ray data collection

All X-ray data collection was performed by M. Shevtsov and J. McGeehan, University of Portsmouth, at beamline ID14-4 at the ESRF (Grenoble, France) using an incident X-ray wavelength of 0.939 Å (13.2 keV) and a ADSC Q315r X-ray detector positioned at a distance of 394 mm from the sample throughout data collection. Rectangular slits were used to achieve a 50 µm × 70 µm beam size (horizontal × vertical). Unfortunately, the beam shape, including full width half maximum (FWHM) was not recorded during data collection. For the purposes of dose calculations performed in Section 3.3.3, it was thus estimated as 280 µm × 90 µm (horizontal × vertical), as reported in a review article for the ID14-4 beamline (McCarthy *et al.*, 2009). This was consistent with previous reports that the beam-profile on ID14-4 was approximately top-hat shaped (Arzt *et al.*, 2005; Owen *et al.*, 2006).

Again, unfortunately, no recording of the incident photon flux was made during data collection. The online documentation for the ID14-4 beamline[§] has been used to provide an estimation of this value (quoted as $\sim 5 \times 10^{12}$ photons/s at 13.2 keV at 100 % beam transmission). Given that it was recorded that the beam was attenuated to 10.47 % throughout, this corresponded well with the $\sim 2.1 \times 10^{11}$ photons/s integrated counts per-image value provided in `.log` file generated at the beamline alongside image collection. The integrated counts per-image was recorded by an in-line diode at ID14-4, consisting of a 25 µm thick piece of aluminium foil positioned in the beam path at 45° to the incident beam direction. Electrons that are scattered by the foil were then detected by diode, and the induced current was integrated over the duration of image collection.

Collection of the radiation damage series for the C.Csp231I-O_L complex crystal was performed following collection of the diffraction dataset that was used to solve the original structure, as

[§] *http://biosync.sbkb.org/beamlineprofile.do?synch_id=esrf&region=european&bmln_name=ID14-4*

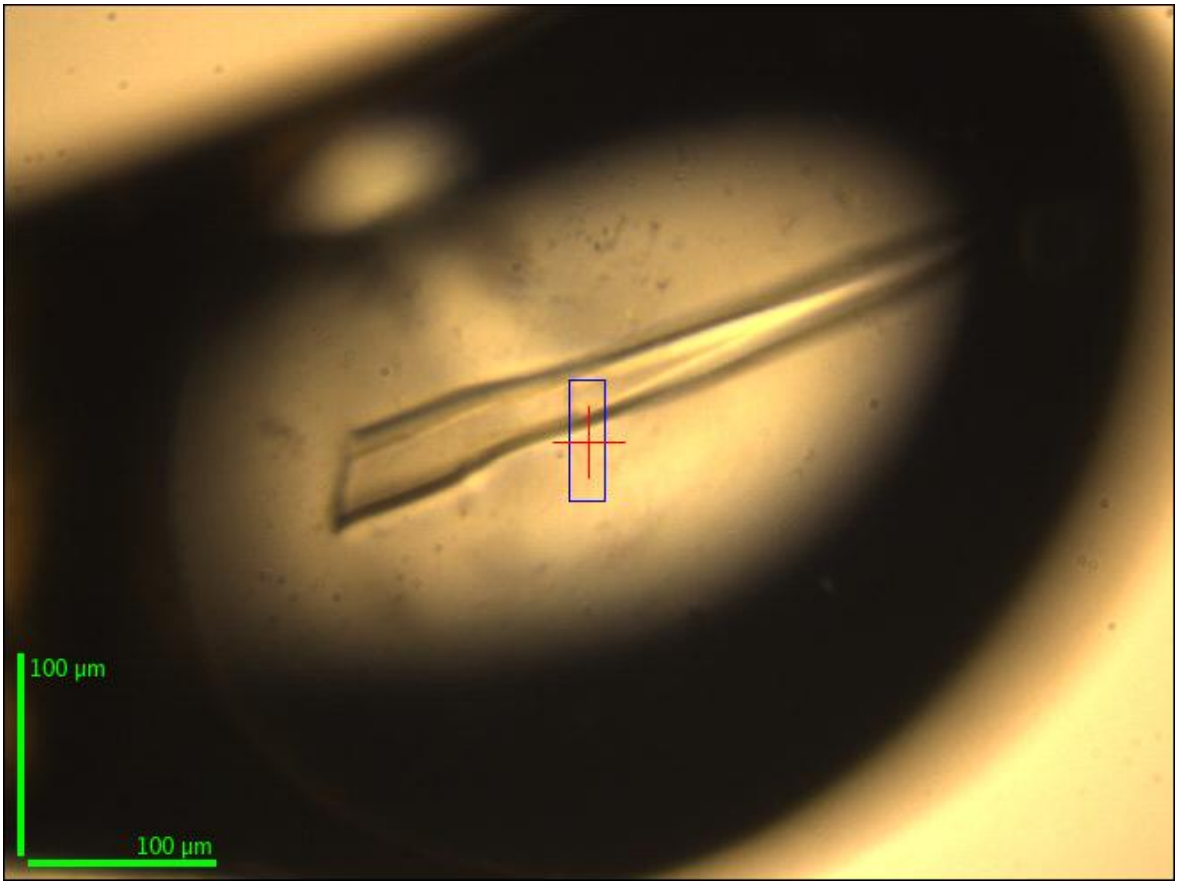


Figure 3.20: The C.Csp231I-DNA crystal mounted in the cryoloop during diffraction data collection. The crystal had approximate dimensions of $x = 20\,\mu m$, $y = 200\,\mu m$ and $z = 20\,\mu m$

reported in Shevtsov *et al.* (2015). The bulky-needle dimensions of the crystal (Figure 3.20) allowed the crystal to be translated to 'fresh' crystal volume prior to collection of the damage series diffraction data (M. Shevtsov, private communications).

Eleven datasets were collected from the C.Csp231I-DNA crystal, each consisting of 90 frames of 1° rotation, each over the same 90° wedge of the crystal with an exposure time of 1 s per frame.

The raw diffraction images were supplied by J. McGeehan, University of Portsmouth, and all subsequent data processing and radiation damage analysis has been performed by the current author.

3.3.3 Dose calculation with *RADDOSE-3D*

Diffraction-weighted dose (DWD) metric values were calculated by *RADDOSE-3D* as specified in Zeldin *et al.* (2013a). The crystal was assumed to be approximately rectangular cuboid-shaped

with dimensions of $x = 20\,\mu m$, $y = 200\,\mu m$ and $z = 20\,\mu m$, which were visually determined from Figure 3.20. In order to accurately estimate the X-ray absorption cross-section of the crystal, including the contribution from protein, DNA and solvent components, the asymmetric unit composition was specified (including unit cell dimensions and number of amino acids and nucleotides per unit cell), in addition to any heavy solvent atoms present in the crystallisation buffer (determined to be Na: 101 mM and Cl: 150 mM, see Section 3.3.1 above). A solvent fraction of 46.9 % was determined from the header information of the refined coordinate model `.pdb` file, as generated in Section 3.3.4 below.

The range of DWD values for the 11 datasets was from 1.2 MGy to 23.7 MGy, with a maximum dose in the crystal volume of 27.5 MGy for the last dataset. Due to uncertainties in the beam FWHM and flux values as discussed in Section 3.3.2, there was inevitable uncertainty in the dose values reported here. To test the suitability of the approximations of these beam parameters, the data in Figure 3.21 show the average intensity decay of recorded reflections versus dose. The DWD half-dose for the crystal is at approximately 19 MGy, and is consequently below the $D_{1/2}$ dose limit of 43($\pm$3) MGy that has previous been experimentally determined and postulated to be an upper dose limit to all protein crystals (Owen *et al.*, 2006)[¶]. This provides an indication that the approximations made above for the beam FWHM and flux parameters appear reasonable.

3.3.4 Diffraction data processing

For each of the eleven datasets in the damage series, diffraction images were independently processed using *DIALS* (Waterman *et al.*, 2016), followed by independent intensity scaling, outlier rejection and conversion to structure factor amplitudes using the standard data reduction pipeline provided by the CCP4i GUI (using *POINTLESS*, *AIMLESS* and *CTRUNCATE*). The space group was determined to be $P6_1$, consistent with the previously published C.Csp231I-O_L complex (Shevtsov *et al.*, 2015). A consistent indexing scheme across datasets was enforced using *POINTLESS*, due to an indexing ambiguity associated with space group $P6_1$. Again consistent with the reports in Shevtsov *et al.* (2015), the crystal was flagged to exhibit partial merohedral

[¶]Whereas 43 MGy has been determined to be an upper dose limit for the mean diffraction intensity to diminish by 50 % of its initial amount, most protein crystals show much lower $D_{1/2}$ values (De la Mora *et al.*, 2011; Zeldin *et al.*, 2013a; Allan *et al.*, 2013). The current data do not suggest that the C.Csp231I system is more radiation-sensitive to global damage than general proteins.

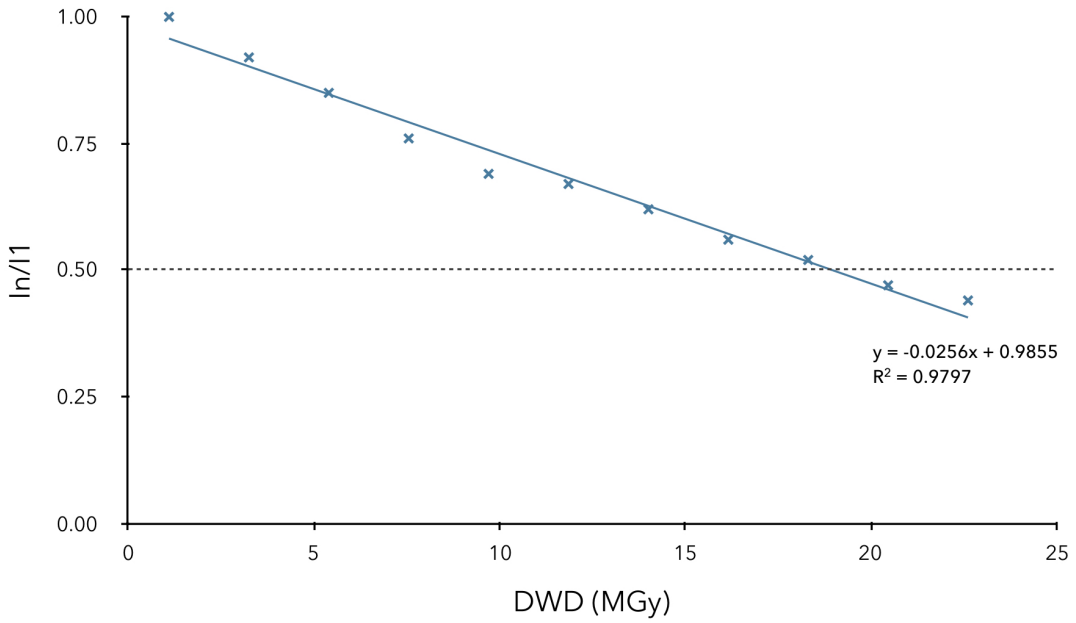


Figure 3.21: The normalised average intensity decay, I_n/I_1, relative to the first dataset, versus diffraction-weighted dose (DWD) for the C.Csp231I-DNA complex damage series. The horizontal dashed line indicates the half-dose, $D_{1/2}$, for the crystal at which the average intensity reached 50 % of the original value in the first dataset. Note that the parameters for the dose calculation were less well established than for the C.Esp1396I-DNA study in Section 3.2.

twinning by the L-test in *POINTLESS* and independently detected in *PHENIX.XTRIAGE*, with twin operator $k, h, -l$. The hemihedral twinning fraction was estimated by the L-test to be as high as 49.1 % but reduced to 46.4 % later during refinement using *REFMAC*. Shevtsov *et al.* (2015) had previously reported a twin fraction 30 % in the region of crystal used to solve the deposited C.Csp231I-O_L structure (PDB: *4jcx*). Care was taken using the program *FREERFLAG* (as part of the CCP4i data reduction pipeline) to ensure that all twin-related reflections acquired the same free R flag, in order to avoid a dependence between the 'free' and 'working' sets of reflections during model refinement below.

For the first dataset, molecular replacement was performed using *PHASER* (McCoy *et al.*, 2007) with the identical structure (PDB: *4jcx* at 2.3 Å) used as a search model. Refinement was performed using *REFMAC*, with the amplitude-based twinning refinement option switched on to account for the partial merohedral twinning discussed above. Initially 10 cycles of rigid-body refinement were performed, followed by repeated rounds of 10 cycles of restrained and TLS refinement, coupled with manual model building in *COOT*. For TLS refinement, the default TLS groups as specified by *REFMAC* (1 TLS body per protein/DNA chain, comprising 4 in total) were used. The final coordinate model was refined to maximum resolution of 2.5 Å. Table 3.5 reports

the data processing statistics for the C.Csp231I-O_L complex.

For each later dataset, similarly as performed for the C.Esp1396I-DNA complex as described in Section 3.2.4, 10 cycles of rigid body refinement were performed in *REFMAC*, using the refined coordinate model from the first dataset coupled with the structure factor amplitudes of each later dataset. No restrained refinement was performed for later dataset models.

3.3.5 Site-specific damage observations using $F_{obs,n} - F_{obs,1}$ maps

Similarly to the previous C.Esp1396I-DNA complex investigation presented in Section 3.2, $F_{obs,n} - F_{obs,1}$ Fourier difference maps were generated (as described in Section 2.4.1), in order to assess radiation damage effects in the C.Csp231I-O_L complex. It was unclear on how to rigorously calculate $F_{obs,n} - F_{obs,1}$ maps whilst accounting for the presence of partial merohedral twinning. Each intensity I_{hkl} now contains the contribution of each twin domain:

$$I^{1+2}_{h,k,l} = \alpha I^{1}_{h,k,l} + (1-\alpha) I^{2}_{k,h,-l} \tag{3.3.1}$$

where α is the twin faction of the smaller domain. As a result, each structure factor term in the Fourier difference summation:

$$F_{obs,n} - F_{obs,1}(\mathbf{x}) = \frac{1}{V} \sum_{hkl} \left[|\mathbf{F}_{obs,n}| - |\mathbf{F}_{obs,1}| \right] \cdot \exp(i\phi_{calc,1}) \cdot \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}) \tag{3.3.2}$$

includes the contribution of two reflection intensities from each twin domain.

Due to the high approximated twinning faction ($\alpha > 46\,\%$), no attempt was made to detwin the data (as could be performed using either *SFCHECK* or *DETWIN* in the CCP4 program suite, but not recommended for high twinning factions exceeding $\alpha = 45\,\%$, due to the high error in evaluating the true reflection intensities $I^{1}_{h,k,l}$ and $I^{2}_{k,h,-l}$ as α becomes close to $50\,\%$). Instead, structure factor amplitudes used to calculate $F_{obs,n} - F_{obs,1}$ maps were taken directly from *CTRUNCATE* with twinning untreated.

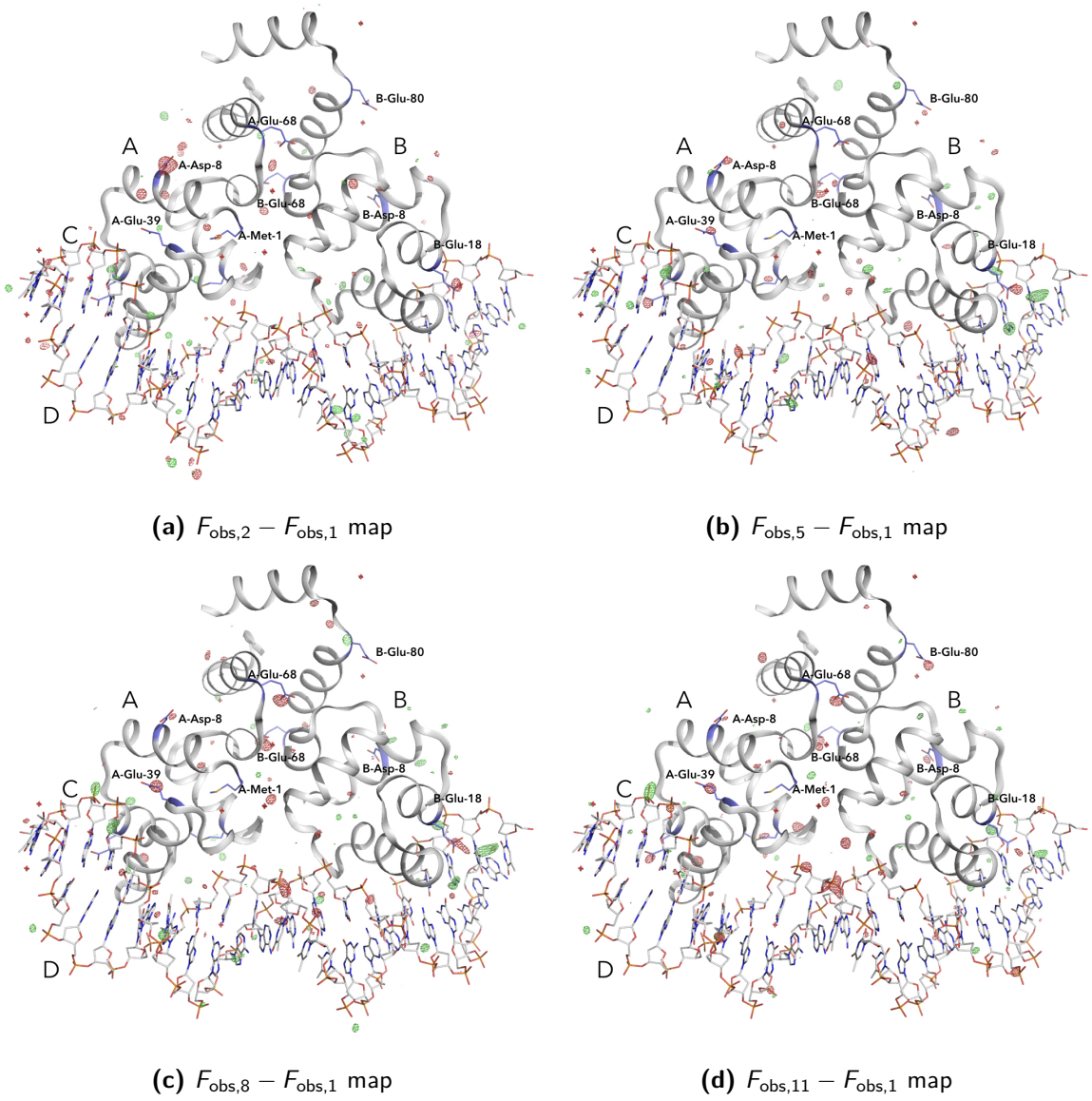
High levels of noise were present in the generated $F_{obs,n} - F_{obs,1}$ difference maps that could not be associated with clear radiation damage events, and this was suspected to be a consequence of the

Table 3.5: Data processing and refinement statistics for the C.Csp231I-O_L complex damage series. Values in parentheses are for the highest-resolution shells ($2.6 - 2.5$ Å). The *B*-factor is estimated from the Wilson plot. For observed F_{obs} and calculated F_{calc} structure factors, $R_{\text{work}} = \frac{\sum |F_{\text{obs}} - F_{\text{calc}}|}{\sum F_{\text{obs}}}$ and R_{free} is the R_{work} formula calculated from a small (5 %) test set of randomly selected reflections. An identical test set of reflections was used for all datasets. Unit-cell angles in space group $P6_1$ are fixed at $\alpha = 90°$, $\beta = 90°$ and $\gamma = 120°$ throughout.

Dataset	1	2	3	4	5	6	7	8	9	10	11
Dose (MGy)	1.17	3.43	5.69	7.94	10.20	12.45	14.71	16.96	19.22	21.47	23.73
Data processing											
Resolution range (Å)	2.5-51.0	2.5-51.0	2.5-50.9	2.5-50.8	2.5-50.8	2.5-50.8	2.5-50.6	2.5-50.7	2.5-50.6	2.5-50.6	2.5-50.5
Cell dimensions											
$a = b$ (Å)	62.45	62.39	62.35	62.27 62.23	62.20	62.06	62.09	62.00	61.99	61.86	
c (Å)	151.93	151.81	151.75	151.55	151.45	151.38	151.03	151.12	150.89	150.87	150.57
No. of observations	64907 (7322)	64778 (7331)	64665 (7353)	64410 (7366)	64329 (7346)	64221 (7309)	63705 (7276)	63779 (7284)	63495 (7212)	63187 (7024)	62016 (6389)
No. of unique reflections	11631 (1298)	11604 (1300)	11591 (1306)	11554 (1310)	11533 (1306)	11514 (1299)	11425 (1291)	11444 (1295)	11398 (1283)	11396 (1285)	11316 (1269)
Multiplicity	5.6 (5.6)	5.6 (5.6)	5.6 (5.6)	5.6 (5.6)	5.6 (5.6)	5.6 (5.6)	5.6 (5.6)	5.6 (5.6)	5.6 (5.6)	5.5 (5.5)	5.5 (5.0)
Completeness (%)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100.0 (100.0)
R_{merge}	0.09 (0.91)	0.09 (1.06)	0.10 (1.25)	0.11 (1.65)	0.11 (2.11)	0.12 (2.68)	0.12 (3.33)	0.132 (5.06)	0.15 (5.72)	0.15 (4.95)	0.155 (2.96)
R_{meas}	0.10 (1.01)	0.10 (1.16)	0.11 (1.38)	0.12 (1.83)	0.12 (2.32)	0.13 (2.95)	0.14 (3.67)	0.15 (5.58)	0.16 (6.31)	0.17 (5.47)	0.17 (3.31)
$CC_{1/2}$	0.999 (0.803)	0.999 (0.745)	0.999 (0.673)	0.999 (0.573)	0.999 (0.455)	0.999 (0.366)	0.999 (0.20)	0.999 (0.193)	0.998 (0.158)	0.998 (0.109)	0.998 (0.093)
$\langle I/\sigma(I) \rangle$	13.6 (1.5)	13.2 (1.3)	12.9 (1.1) 13.2 (0.8)	12.3 (0.7)	10.0 (0.4)	9.4 (0.3)	8.7 (0.2)	8.1 (0.2)	5.8 (0.2)	7.2 (0.3)	
I_n/I_1	1.00	0.92	0.85	0.76	0.69	0.67	0.62	0.56	0.52	0.47	0.44
Wilson *B*-factor (Å^2)	34.1	35.9	38.8	42.7	45.8	53.0	58.6	72.5	74.9	76.3	83.1
Refinement											
R_{work}	0.202	0.231	0.232	0.234	0.239	0.247	0.247	0.254	0.262	0.267	0.277
R_{free}	0.216	0.242	0.242	0.245	0.256	0.260	0.251	0.265	0.273	0.281	0.288

Table 3.6: Assessment of final macromolecular geometry for the refined coordinate model of the C.Csp231I-DNA complex, corresponding to the first dataset, as reported by *Molprobity4* (Chen *et al.*, 2010). Since for the coordinate models corresponding to the higher dose datasets, only rigid-body refinement was performed on the refined initial dataset coordinate model coupled with the higher dataset merged structure factor amplitudes, the reported geometry statistics were conserved for the higher dose datasets.

No. non-H atoms:	
Protein	1454
DNA	855
Water	13
Other	0
Bond parameters:	
RMSD bond length (Å)	0.01
RMSD bond angle (°)	1.04
Ramachandran analysis:	
Favoured (%)	99.43
Outliers (%)	0.00
Allowed (%)	0.57
Rotamer outliers (%)	1.32
All atom clash-score	0.79



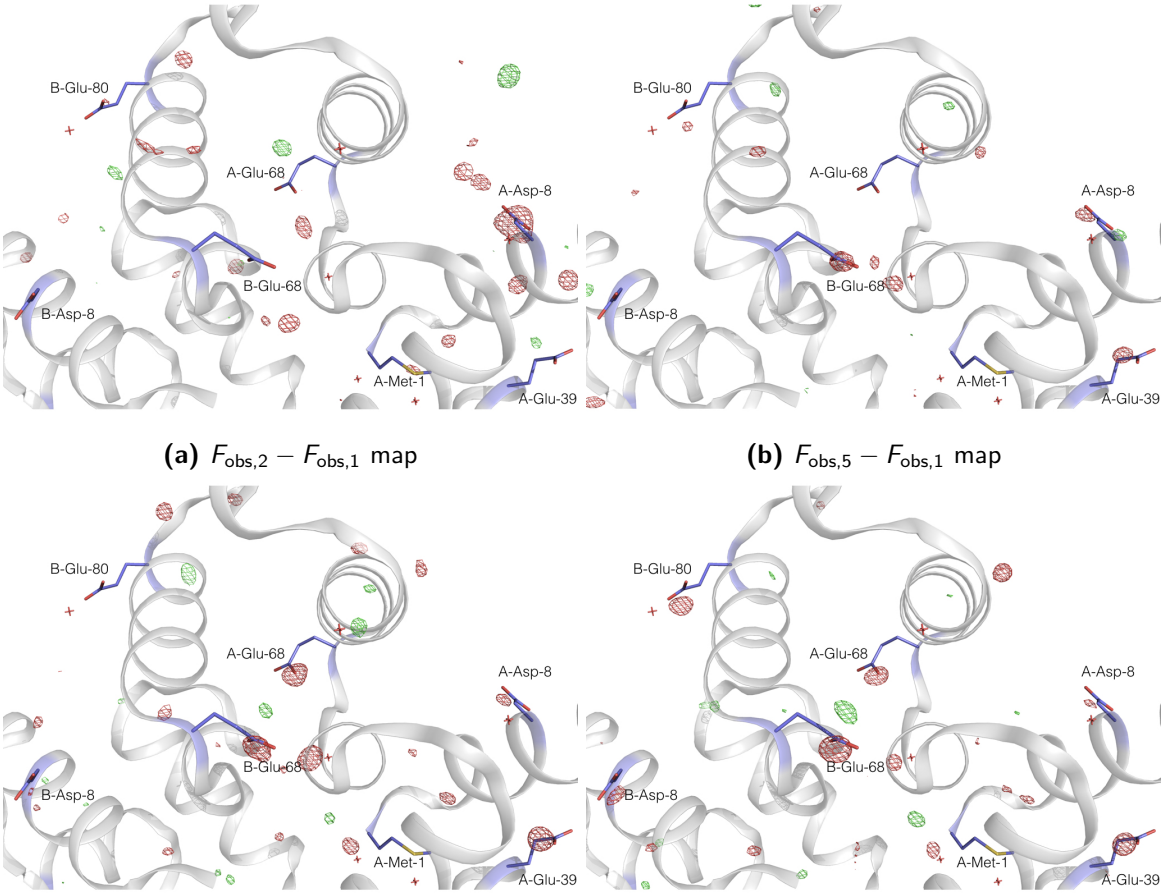
(a) $F_{obs,2} - F_{obs,1}$ map

(b) $F_{obs,5} - F_{obs,1}$ map

(c) $F_{obs,8} - F_{obs,1}$ map

(d) $F_{obs,11} - F_{obs,1}$ map

Figure 3.22: The spatial distribution of $F_{obs,n} - F_{obs,1}$ map peaks over the full C.Csp231I-DNA complex structure for four increasing dose states: (a) $F_{obs,2} - F_{obs,1}$, (b) $F_{obs,5} - F_{obs,1}$, (c) $F_{obs,8} - F_{obs,1}$ and (d) $F_{obs,11} - F_{obs,1}$. For each map, the map contouring level has been fixed at $\pm 3.5\sigma$ in green/red. Protein residues for which significant electron density loss has been detected have been shown in *stick representation* in *PYMOL*.



(a) $F_{\text{obs},2} - F_{\text{obs},1}$ map **(b)** $F_{\text{obs},5} - F_{\text{obs},1}$ map

(c) $F_{\text{obs},8} - F_{\text{obs},1}$ map **(d)** $F_{\text{obs},11} - F_{\text{obs},1}$ map

Figure 3.23: Examples of suspected damage sites to the protein component of the C.Csp231I-DNA complex detectable in $F_{\text{obs,n}} - F_{\text{obs},1}$ Fourier difference maps, for $n = 2, 4, 6, 8$. Site-specific damage to Glu, Asp and Met residues at the dimer interface between chains A and B has been shown. Residues of interest have been shown with the *stick representation* in *PYMOL*, with all other residues omitted. All Fourier difference maps are contoured at $\pm 3.5\sigma$ in green/red, which was deemed to be at a level exceeding the background noise of the map.

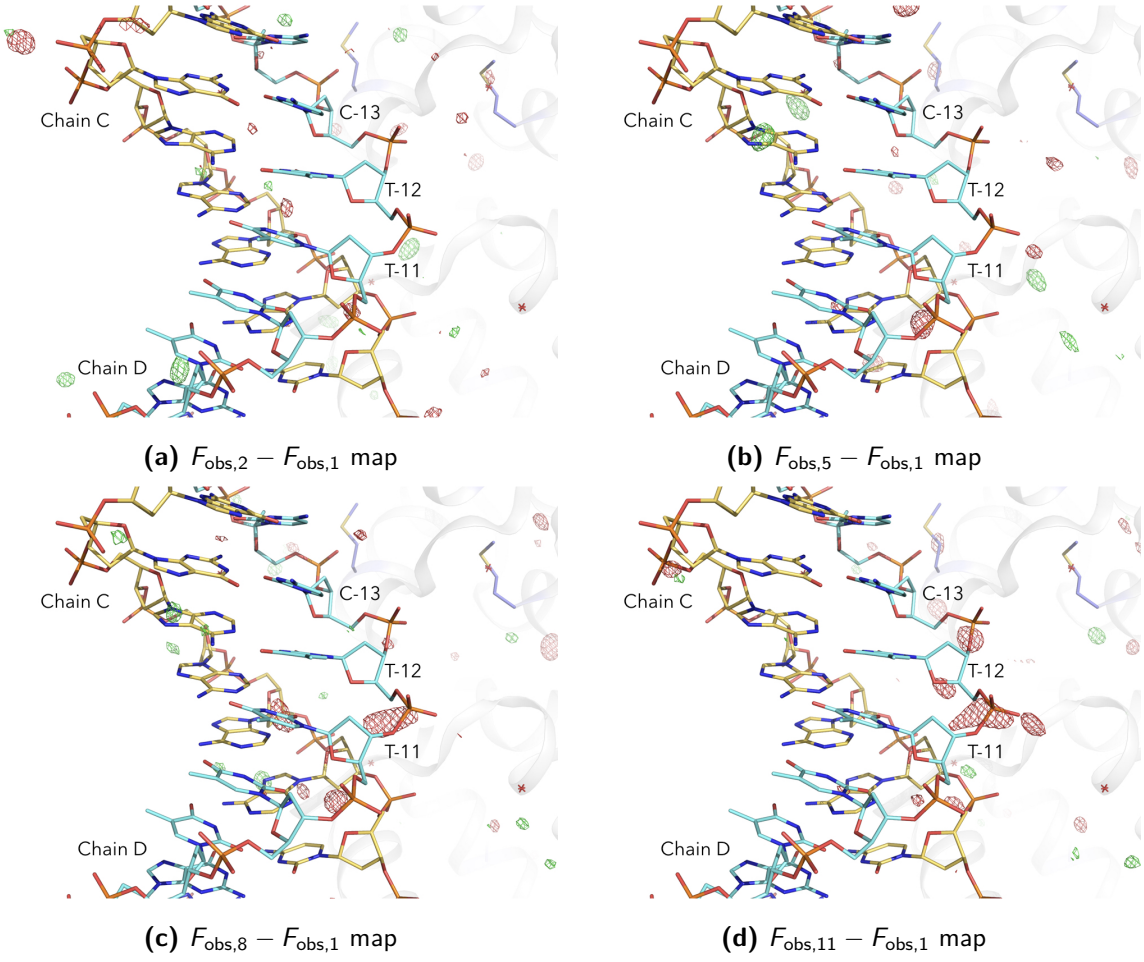
twinned data. Nonetheless, manual inspection of $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps did reveal sites of electron density loss that were consistent with protein site-specific damage, most notably multiple sites of potential decarboxylation of Glu and Asp residues, in addition to electron density loss associated with the methylthiol group of several Met residues (Figure 3.23). As with the C.Esp1396I protein, C.Csp231I contains no disulphide bonds.

However, as indicated in Figure 3.23, it was noted that several $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks were not exactly coincident with their attributed residue (e.g. A-Glu-39, B-Glu-80). This could either be a consequence of poor placement of each side-chain during model building, or the effect of partial twinning on $F_{\text{obs,n}} - F_{\text{obs,1}}$ map quality. Additionally, $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks did not always persist at higher dose datasets once detected, as indicated in Figure 3.23 for A-Asp-8.

Consistent with the previous observations on the C.Esp1396I-DNA complex presented in Section 3.2.6, electron density loss sites were detectable at several regions on the phosphate backbone of the DNA, and furthermore, these effects became more pronounced in higher dose $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps (typically for $n \geq 8$, corresponding to $\geq 17.0\,\text{MGy}$). As illustrated in Figure 3.24, nucleotides T-11 and T-12 of chain D appeared to be particularly prone to electron density loss; these nucleotides occupy the central spacer region of the 21 bp O_L operator sequence between the two protein-DNA interaction sites, and the DNA backbone is suspected to be most distorted and strained at this spacer region (Shevtsov *et al.*, 2015). Overall, these electron density loss sites did not typically coincide exactly with either the phosphate group or the C−O bond either side of the phosphate, and thus could not be clearly assigned as corresponding to a specific bond rupture event or disordering of the individual backbone atoms.

Many regions of unexplainable positive and negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density coincided with both the protein and DNA components, in addition to occupying nearby solvent regions (Figure 3.22). As such, the validity of the manual interpretation of each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map was questionable. As an example, Figure 3.25 demonstrates the significant unexplainable positive $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density located at the basic subunits of several adjacent nucleotides (T-2, A-3 and C-4 of chain D), at the same $F_{\text{obs,n}} - F_{\text{obs,1}}$ map threshold ($\pm 3.5\sigma$) as used in Figure 3.23 and 3.24.

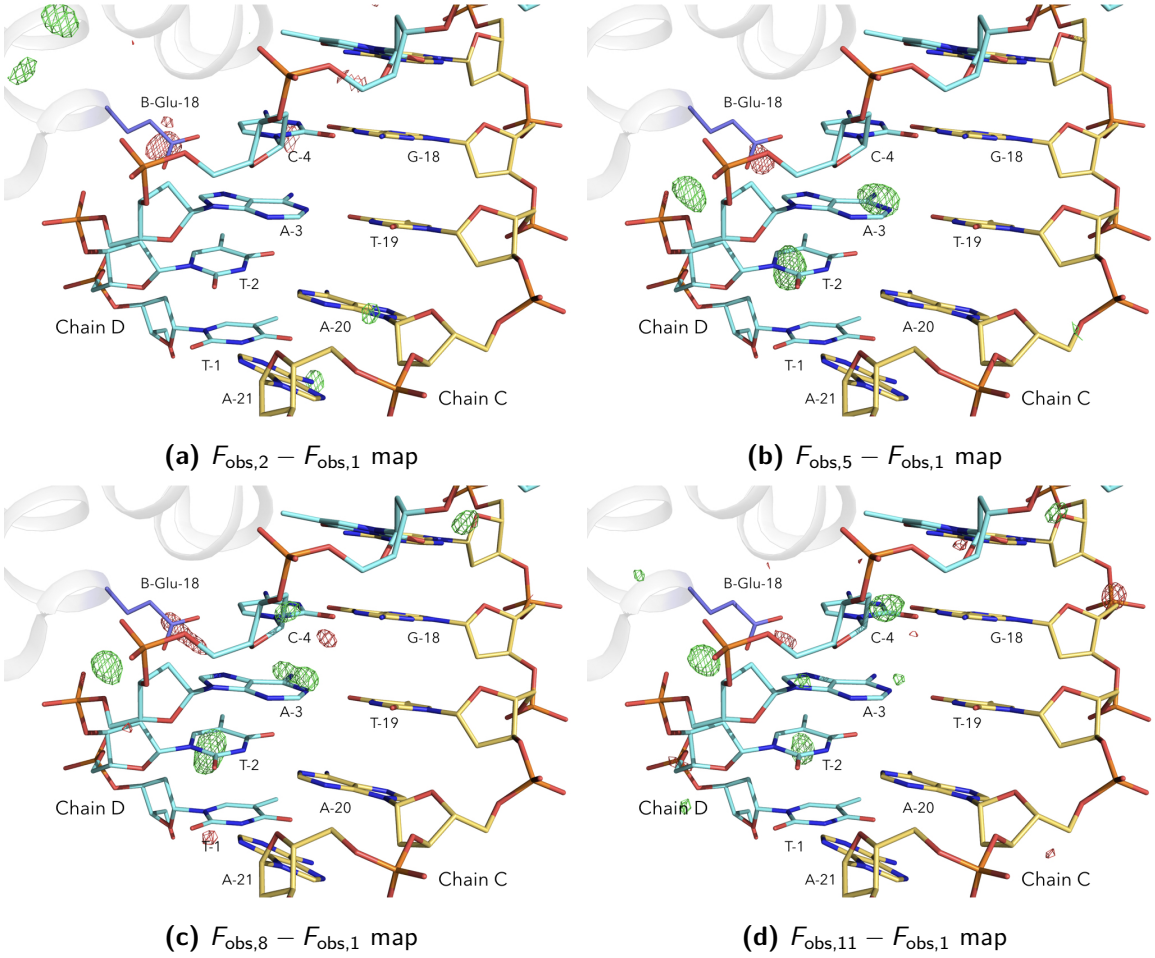
Ultimately, as a consequence of the poor quality of the $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps, further map investigation using either the *PEAKS* (Section 2.4) or *RIDL* (Section 2.5) program was deemed



(a) $F_{\text{obs},2} - F_{\text{obs},1}$ map **(b)** $F_{\text{obs},5} - F_{\text{obs},1}$ map

(c) $F_{\text{obs},8} - F_{\text{obs},1}$ map **(d)** $F_{\text{obs},11} - F_{\text{obs},1}$ map

Figure 3.24: Examples of suspected damage sites to the DNA component of the C.Csp231I-DNA complex detectable in $F_{\text{obs,n}} - F_{\text{obs},1}$ Fourier difference maps, for $n = 2, 4, 6, 8$. Electron density loss localised at the phosphate backbone of nucleotides T-11 and T-12 of chain D has been illustrated. All Fourier difference maps are contoured at $\pm 3.5\sigma$ in green/red to be consistent with Figure 3.23, and this was deemed to be at a level that exceeded the background noise level in the map.



(a) $F_{obs,2} - F_{obs,1}$ map **(b)** $F_{obs,5} - F_{obs,1}$ map

(c) $F_{obs,8} - F_{obs,1}$ map **(d)** $F_{obs,11} - F_{obs,1}$ map

Figure 3.25: Examples of unexplainable positive $F_{obs,n} - F_{obs,1}$ map density located on the DNA component of the C.Csp231I-DNA complex, in particular at the basic subunits of T-2, A-3 and C-4 in chain D, and in the solvent region proximal to the phosphate group of T-2 in chain D. $F_{obs,n} - F_{obs,1}$ maps have been shown for (a) $n = 2$, (b) $n = 4$, (c) $n = 6$ and (d) $n = 8$. All Fourier difference maps are contoured at $\pm 3.5\sigma$ in green/red to be consistent with Figure 3.23, and this was deemed to be at a level at which protein site-specific damage events could be distinguished above the background noise in each $F_{obs,n} - F_{obs,1}$ map.

to be unreliable and unproductive, and has thus been omitted from this report.

3.4 Conclusion

MX site-specific radiation damage investigations on nucleic acids have been severely lacking to date. It is unclear how crystalline nucleic acids can be modified under X-ray irradiation at either RT or 100 K, and ultimately where errors due to radiation damage can manifest in a nucleic acid structure solved by MX. The investigation of self-assembling and functionally-relevant protein-nucleic acid complexes presented in the current chapter allows the radiation sensitivity of DNA to be directly and quantitatively compared to its bound protein counterpart, for which a multitude

of prior MX radiation damage studies exist.

Section 3.2 has presented novel findings of site-specific damage at 100 K to the DNA component in a well-characterised bacterial protein-DNA complex, the C.Esp1396I system. In particular, X-ray induced electron density loss has been detected at the phosphate backbone subunits of DNA, and moreover, there appears to be preferential electron density loss at compressed AT-rich regions of each DNA strand. These observations are consistent with bond rupture events located at the phosphate backbone, however due to the limited resolution of the C.Esp1396I study (2.8 Å), this could not be assigned to any specific bond at each phosphate group. The use of further complementary experiments (e.g. time resolved infrared spectroscopy) in conjunction to X-ray data collection would be a suggested approach to directly identify the bond cleavage event. This proposed experiment falls beyond the time frame of the current project, and has instead been suggested as future work.

In the current experiment it has been demonstrated that DNA in the C.Esp1396I system is particularly radiation-resistant: whereas classical protein damage effects (decarboxylation and electron density loss to Met side-chains) was distinguishable in the lowest $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference map calculated (providing a comparison between dose states 2.07 MGy and 6.19 MGy), electron density loss to the DNA component only became prevalent at much higher doses (above 20.6 MGy relative to the 2.07 MGy base state). These results are in direct contrast to radiation damage studies on freeze- or vacuum-dried films of oligonucleotides typically irradiated at room temperature, for which strand cleavage events (both SSBs and DSBs) are detectable in significant yields at much lower doses (in the mGy – kGy regime) following the interaction of LEEs (Alizadeh *et al.*, 2015). Whereas it could be argued that crystal packing may limit the ability of a full SSB event in the current systems, the initial rupture of the C–O bond at the phosphate backbone should not be affected by this. Indeed, for the C.Esp1396I-DNA complex, as a result of the $P6_5$ crystal symmetry, the surfaces of the DNA component that are not in direct contact with either C-protein dimer are highly solvent exposed, and such a cleavage event should thus not be sterically hindered (Figure 3.26).

Furthermore, as detailed in Section 3.2.6.2, the additional presence of positive density accumulation in the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map at basic subunits near to sites of phosphate electron density loss could be interpreted as the reorientation of the basic subunit following backbone bond cleavage

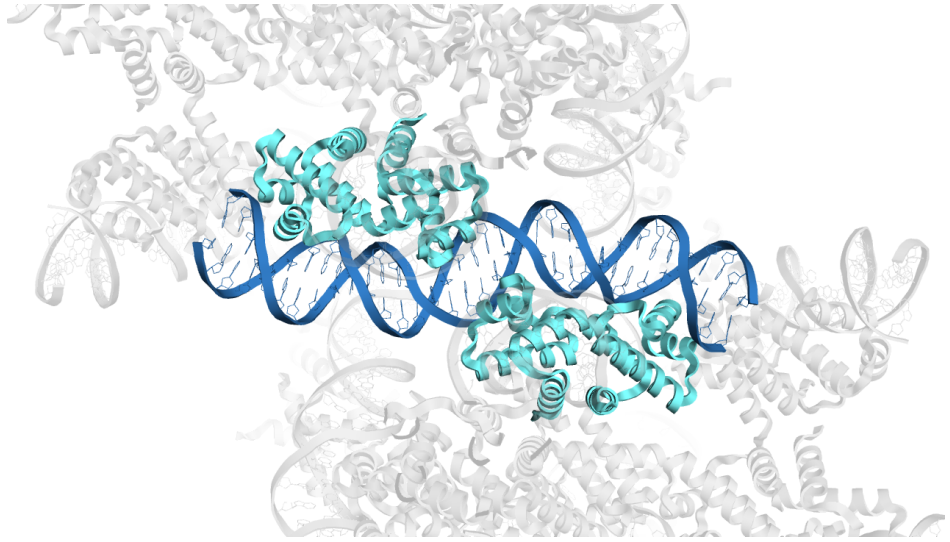


Figure 3.26: Rendering of the crystal packing for the C.Esp1396I-DNA complex. Symmetrically-related subunits have been generated using the *symexp* command in *PYMOL*. The bottom left and top right surfaces of the DNA are highly solvent exposed.

at high dose values. To test this hypothesis computationally, it is suggested that a QM/MM (quantum mechanics/molecular mechanics) approach could be utilised to simulate the motion of a nucleotide at a site of simulated phosphodiester bond cleavage. Due to the crystalline environment this simulation is non-trivial, and care would need to be taken to account for both the crystal packing and the modelling of the solvent arrangement in the bulk solvent channels of the crystal. This study falls outside the scope of the work presented in this chapter and has instead been suggested as a future extension of the current work.

To complement the work conducted in Section 3.2, site-specific damage to a further C-protein DNA complex (the C.Csp231I protein bound to 21 bp O_L DNA operator sequence) has been investigated. Unfortunately, due to significant merohedral twinning detected in the data (with twin fraction of $\sim 46\,\%$), both data processing and subsequent inspection of $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps was challenging. As such, the success of damage analysis on the C.Csp231I system has been severely limited.

Nonetheless, despite the resulting poor quality of $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps, manual inspection revealed sites of Glu/Asp decarboxylation and electron density loss proximal to the phosphate backbone of the DNA at a subset of nucleotides. In particular, the observations of electron density loss at the phosphate backbone only became defined in the highest $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps generated. These observations provide evidence supporting the conclusions drawn in Section 3.2.

Most interestingly, for this independent C-protein-DNA system, the most dominant electron density loss site situated on the DNA backbone again appeared to be located at an AT-rich region that is under the most strain due to protein binding (in C.Csp231I at the 5 bp spacer region between the two protein recognition sequences). Unfortunately, the prevalence of significant unaccountable noise in all $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps limited any further conclusions that could be drawn in the current study on the C.Csp231I system.

Overall, this chapter has demonstrated the radiation-robustness of DNA for two example systems, both on double stranded DNA in complex with proteins. However it is unclear whether nucleic acids are intrinsically radiation-insensitive in MX, since it can be speculated that both the double helix conformation and bound proteins could stabilise the DNA, thus contributing to its observed robustness. In order to investigate whether such radiation-insensitivity persists for more dynamic single-stranded nucleic acids, Chapter 4 will next investigate a protein-nucleic system comprising of single-stranded RNA bound to protein. Again, the presence of protein will provide an ideal standard against which to compare the RNA.

Alternatively, due to the prevalence of protein site-specific damage in the C.Esp1396I and C.Csp231I -DNA complexes, it could be postulated that bound protein may offer protection to nucleic acid. It is suggested that protein may scavenge mobile electrons at 100 K that would otherwise interact with the DNA, similarly to that performed by small molecule additives (radioprotectants) soaked into protein crystals (De la Mora *et al.*, 2011). To further investigate whether bound protein mitigates DNA damage, this question can only be addressed by investigating isolated nucleic acid crystals in the complete absence of protein. To this end, Chapter 5 will conduct a further investigation to evaluate the radiation-sensitivity of DNA in isolation.

CHAPTER 4

Radiation damage to nucleic acids in complex with proteins. Part II: The TRAP-RNA complex

4.1 Introduction

In Chapter 3, radiation damage analysis across a well characterised bacterial protein-DNA complex (C.Esp1396I, PDB: *3clc*, resolution: 2.8 Å (McGeehan *et al.*, 2008)) was described, and it was found that the bound C-protein component was far more damage susceptible relative to DNA over the wide dose range investigated ($2.1 - 44.6$ MGy over eight datasets). However, the conclusions drawn for the C.Esp1396I-DNA complex in Chapter 3 were ultimately limited by the low resolution of the available MX damage series (at 2.8 Å), such that a reliable per-atom analysis of damage was unachievable. Moreover, from the results presented in Chapter 3, there was insufficient evidence to indicate whether DNA was intrinsically less susceptible to X-ray induced damage in MX for crystalline complexes such as C.Esp1396I, or whether protein-binding provided protection to DNA, either by scavenging damaging radical species away from the DNA subunit, or by physically stabilising the DNA upon binding. Ideally, to truly elucidate this question would also require a non-protein-bound C.Esp1396I DNA component within the same crystal to act as a direct control under identical crystallisation and data collection conditions. Unfortunately, due in part to the unpredictability of crystallisation, it has not been possible to secure any available protein and nucleic acid-containing crystals that contain a free nucleic acid component.

The investigation presented in the current chapter aims to address the questions raised in Chapter 3 and has focused on the specific radiation damage to crystals of *Trp* RNA-binding attenuation protein (TRAP) bound to the 53 bp RNA sequence $(GAGUU)_{10}GAG$. To the author's knowledge this is the first site-specific damage analysis to be conducted on a complex containing single-stranded RNA. In comparison to the relatively rigid double helix dsDNA conformations exclusive to complexes presented in Chapter 3, it is also unclear whether the radiation-insensitivity of nucleic acid should persist over a comparatively flexible RNA strand. For the current TRAP-RNA system, diffraction data have been collected to ~ 1.98 Å, thus permitting substantially more reliable conclusions to be drawn regarding site-specific damage to individual atoms.

The TRAP-RNA system was chosen primarily since crystals grew with each asymmetric unit containing one RNA-bound TRAP macromolecule, and one TRAP molecule with no RNA bound to it (Figure 4.1). This configuration was ideal for allowing the effect of RNA-binding on damage to each protein molecule to be probed, by direct comparison between equivalent residues of the

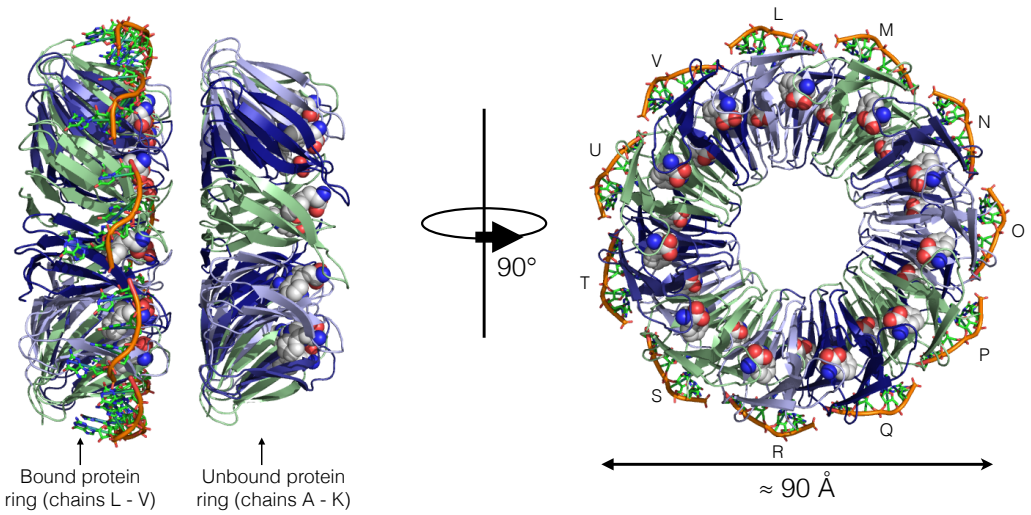


Figure 4.1: The TRAP-RNA complex asymmetric unit, rendered in *PYMOL* (PDB: *1gtf*, resolution: 1.75 Å). For each TRAP ring ($\sim$ 91 kDa), eleven bound L-tryptophan ligands have been represented as spheres. The RNA strand has been shown bound to the exterior of one TRAP ring (chains L-V). TRAP crystallises in a 1:1 ratio of the RNA-bound and non-bound TRAP ring (chains A-K) stacked in the asymmetric unit.

RNA-bound and non-bound protein. It could therefore be tested whether nucleic acid-binding to a protein could alter the susceptibility of radiation sensitive residues to radiation damage, and consequently, whether or not protein appeared to scavenge damage in the RNA-bound form. Furthermore, each TRAP functional assembly consists of eleven identical protein chains (each of 74 amino acids, of total size $\sim$ 91 kDa), forming an undecamer ring structure with 11-fold rotational symmetry about an axis through the ring centre. By employing this high symmetry, a thorough statistical quantification of the radiation damage effects of RNA binding to TRAP was achievable.

4.1.0.1 *Trp* RNA-binding attenuation protein

TRAP regulates the transcription of tryptophan biosynthetic genes in *Bacillus subtilis* and several other bacteria through an attenuation mechanism (Shimotsu *et al.*, 1986). TRAP becomes functional in Trp-rich intracellular environments; once activated through the binding of L-tryptophan ligands at eleven intersubunit hydrophobic pockets symmetrically spaced around the TRAP ring, TRAP can then bind to a specific RNA sequence within the 203 bp 5' leader segment of the nascent *trpEDCFBA* operon mRNA. RNA-binding induces the formation of a downstream *terminator* hairpin conformation in the RNA, which leads to termination of transcription before the

region encoding structural genes is reached. In the absence of TRAP-binding, an *anti-terminator* hairpin instead forms that overlaps with the TRAP-binding site, and *trpEDCFBA* mRNA is transcribed (Antson *et al.*, 1999; Hopcroft *et al.*, 2002).

Interestingly, in *B. subtilis*, TRAP is also known to regulate translation of the *trpE* (Du and Babitzke, 1998), *trpG* (Yang *et al.*, 1995) and *yhaG* genes (Sarsero *et al.*, 2000). In the former, TRAP binds to the same *trpEDCFBA* operon mRNA leader sequence, to cause another downstream RNA hairpin to form, sequestering the *trpE* ribosome-binding site. For the *trpG* and *yhaG* genes, TRAP competes directly with the ribosome for an overlapping binding site (Sarsero *et al.*, 2000). The preferential binding interactions exhibited by TRAP are thus well documented.

TRAP binds with high affinity ($K_d \sim 1$ nM) to specific RNA sequences containing eleven (G/U)AG trinucleotide repeats, each separated by several less-specific spacer nucleotides (Elliott *et al.*, 2001) (see fig. 4.2 reproduced from Antson *et al.* (1999) for the binding mechanism). The highly conserved trinucleotide repeats bind to eleven symmetric binding pockets around the TRAP ring, predominantly through base interactions with the protein, via direct and water-mediated hydrogen bonds to residues Glu-36, Lys-37, Asp-39, Lys-56 and Arg-58, in addition to multiple van der Waals interactions (Hopcroft *et al.*, 2002). No direct hydrogen-bond interactions are made between TRAP and the repeated RNA spacer nucleotides (Hopcroft *et al.*, 2002), and these nucleotides are relatively disordered.

4.2 Materials and methods

4.2.1 Crystallisation and X-ray data collection

All protein expression and purification, crystallisation and data collection was performed by Prof J. McGeehan and Dr. M. Shevtsov, University of Portsmouth, UK. Since this was not the author's own work, only a brief summary has been provided in the current section, containing information that is most relevant for subsequent data handling and dose calculations.

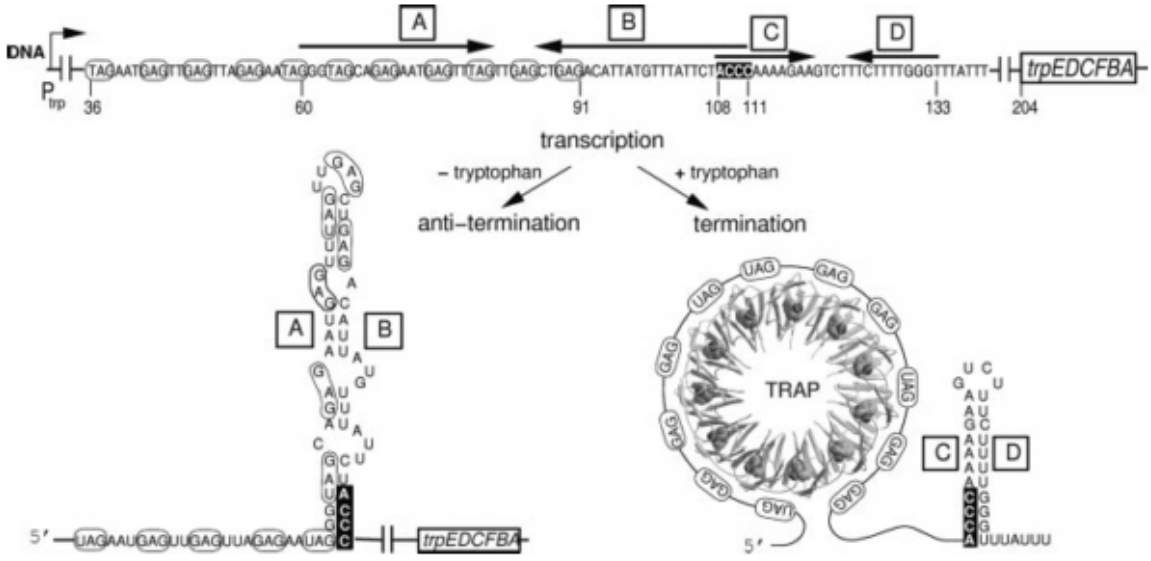


Figure 4.2: The transcription attenuation in *B. subtilis* involving TRAP. The 203-nucleotide leader region that precedes the structural genes of the *trpEDCFBA* operon of *B. subtilis* is shown with numbering corresponding to the start of transcription. The inverted repeats predicted to form the terminator (nucleotides 108−133) and the anti terminator (nucleotides 60−111) stem-loops in the RNA transcript are designated by large boxed letters with complementary strands shown by arrows above the DNA. GAG and TAG triplets (or UAG in the resulting RNA transcripts) have been circled. TRAP is shown as a ribbon model bound to the 36−91 region of the RNA transcript containing the eleven GAG and UAG triplets. This figure has been reproduced directly from Antson *et al.* (1999).

4.2.1.1 RNA synthesis and protein preparation

The 53-base RNA $(GAGUU)_{10}GAG$ was synthesized by *in vitro* transcription with T7 RNA polymerase and gel-purified (the reader is referred to Hopcroft *et al.* (2002) for full details). TRAP from *B. stearothermophilus* was overexpressed in *E. coli* and purified as previously described (Chen *et al.*, 1999).

4.2.1.2 Crystallisation

TRAP-RNA crystals were prepared using a previously established hanging drop crystallisation protocol (Antson *et al.*, 1999). By using a 2:1 molar ratio of TRAP to RNA, crystals successfully formed from the protein-RNA complex ($\sim 15\,\mathrm{mg\,ml^{-1}}$) in a solution containing 70 mM potassium phosphate pH 7.8 and 10 mM L-tryptophan.

The reservoir contained 0.2 M potassium glutamate, 50 mM triethanolamine pH 8.0, 10 mM

$MgCl_2$ and $8 - 11\,\%$ monomethyl ether PEG 2000. In order to accelerate crystallisation, a further gradient was induced by adding 0.4 M KCl to the reservoir after 1.5 µl protein solution was mixed with an equal volume of the reservoir solution. As in Hopcroft *et al.* (2002), crystals grew within 3 days and were vitrified and stored in liquid nitrogen immediately after growth. The cryosolution contained 12 % monomethyl ether PEG 2000, 30 mM triethanolamine pH 8.0, 6 mM L-tryptophan, 0.1 M potassium glutamate, 35 mM potassium phosphate pH 7.8, 5 mM $MgCl_2$ and 25 % MPD.

4.2.1.3 X-ray data collection

An MX damage data series was collected at 100 K at beamline ID14-4, ESRF by Prof. John McGeehan, University of Portsmouth, using an incident wavelength of 0.940 Å (13.2 keV), and an ADSC Q315R mosaic CCD detector. The crystal-detector distance was fixed at 304.5 mm throughout data collection, which gave a resolution of 1.98 Å in the largest inscribed circle. The beam size was recorded to be 0.100 mm × 0.160 mm (vertical × horizontal), achieved using rectangular slits, with a uniformly distributed profile. A TRAP-RNA crystal of dimensions 60 µm × 80 µm × 20 µm (Figure 4.3) was initially orientated with the beam direction parallel to the smallest crystal dimension, such that the crystal was completely bathed in the beam throughout data collection.

Sixteen successive 180° data sets were collected over the same angular range from a TRAP-RNA crystal ($\phi = 40°$ to $\phi = 140°$ with $\Delta\phi = 1°$). The beam transmission was maintained at 28.9 % for each dataset except for $n = 11$,13 and 15, for which it was increased to 100 % to act as X-ray 'burns'. For the first nine data sets, the attenuated flux was reported to be $\sim 5 \times 10^{11}$ photons/s at 28.9 % transmission, following beam collimation, as determined using an on-line Si PIN diode (Owen *et al.*, 2009) prior to collection of the first dataset (J. McGeehan, private communications).

A beam refill took place immediately before dataset 10, and as such, the beam flux could not be assumed to be constant throughout data collection. The flux increase was estimated from a plot of relative intensity I_n/I_1 versus dose between datasets $n = 1$ and $n = 9$ (for which the flux was assumed constant). I_n/I_1 values were derived from the *AIMLESS* log file using Equation 3.2.1 as reported in Section 3.2.3 for the C.Esp1396I-DNA investigation. Based on the mean summed

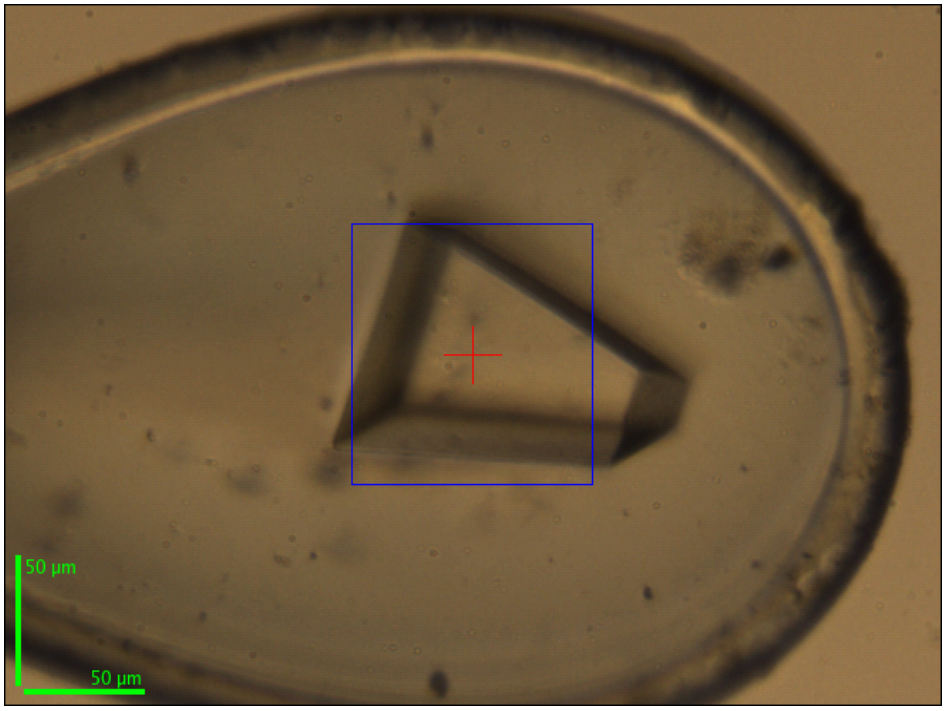


Figure 4.3: The TRAP-RNA crystal within a rayon cryo-loop on beamline. From the scale bars given, the x and y crystal dimensions were estimated to be 60 µm, 80 µm, 80 µm, 30 µm (travelling in a clockwise direction around the crystal from the left-most side). The z dimension is orientated directly into the page, perpendicular to the loop. (b) A model of the TRAP-RNA crystal was created using 3-dimensional modelling software package *BLENDER*. Care was taken to ensure the crystal was correctly orientated such that the z dimension was perpendicular to the cryo-loop.

intensity ratio $I_{10}/I_1 = 0.854$ at dataset 10 to that found by linear extrapolation from dataset 9 (estimated as 0.6 from Figure 4.4), a flux-scale factor increase of $0.854/0.6 = 1.42$ was estimated for dataset 10. As such, the dataset 10 flux was $\sim 7.12 \times 10^{11}$ photons/s, and this higher flux value was then used for all datasets $n > 9$. This flux value was further up-scaled by a factor of 100/29.8 for datasets $n = 11$, 13 and 15 to account for the 100 % transmission used. As shown in Figure 4.4, by scaling each I_n/I_1 value relative to the both transmission and incident flux attained at dataset $n = 1$, the resulting I_n/I_1 intensity decay with dose became smooth and monotonic as expected, suggesting that the assumptions stated above were suitable.

The data in Figure 4.4 indicated that, once the beam refill was accounted for, the $D_{1/2}$ half-dose limit was reached by dataset $n = 11$. It was anticipated that the onset of significant global radiation damage would disrupt the validity of site-specific damage observations, as reported in Owen *et al.* (2006). Consequently, datasets $n > 10$ have been omitted from all further site-specific radiation damage analysis presented in the current chapter.

Prof. J. McGeehan supplied all the TRAP-RNA raw diffraction images, but all of the subsequent

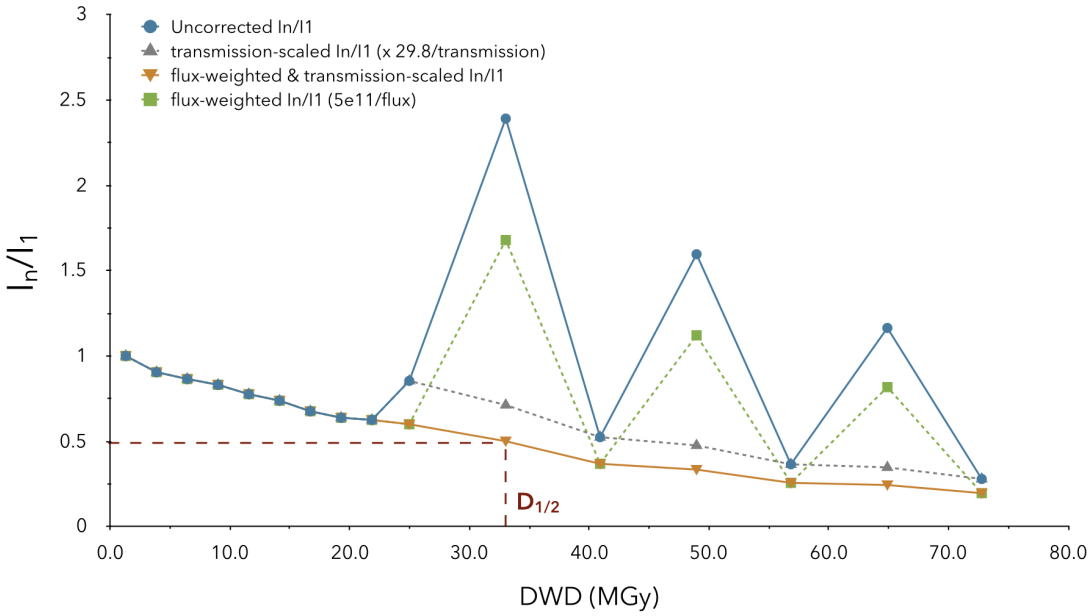


Figure 4.4: I_n/I_1 intensity decay with diffraction-weighted dose for the TRAP-RNA complex MX damage series. The experimentally measured I_n/I_1(dose) values are shown in blue. To account for variable the transmission $t(n)$ between datasets n, the I_n/I_1(dose) values are scaled by a factor of $^{29.8}/_{t(n)}$ (shown in grey) to allow the 100 % transmission used in datasets $n = 11$, 13 and 15 to be accounted for. To account for the variable flux $f(n)$ between datasets n, the I_n/I_1(dose) values are scaled by a factor of $^{5\times10^{11}}/_{f(n)}$ (shown in green) to allow the increasing flux for datasets $n > 9$ to be accounted for. The result of applying both a flux scale factor and transmission scale factor to the I_n/I_1(dose) values is shown in orange. The red dashed lines indicates the half-dose, $D_{1/2}$, at which the summed intensity, I_n, has reached 50 % of that of dataset 1.

data reduction and analysis for this specific damage investigation has been performed by the current author.

4.2.2 Dose calculation

RADDOSE-3D (Zeldin *et al.*, 2013b) was used to calculate the absorbed dose distribution after each dataset collected in the damage series. In order to specify the TRAP crystal geometry, a three-dimensional crystal was modelled with the open-source modelling software *BLENDER*, using an image of the crystal mounted on the beamline as a reference (Figure 4.3), and implemented into *RADDOSE-3D* using the *POLYHEDRON* crystal class (Figure 4.5). The crystal composition was calculated from the TRAP-RNA crystal structure previously deposited in the PDB with accession code *1gtf* (Hopcroft *et al.*, 2002), from which the unit cell dimensions, number of monomers, protein residues and RNA nucleotides, and solvent fraction could be established. To calculate the crystal absorption coefficients in *RADDOSE-3D*, the concentration (in mmol l^{-1}))

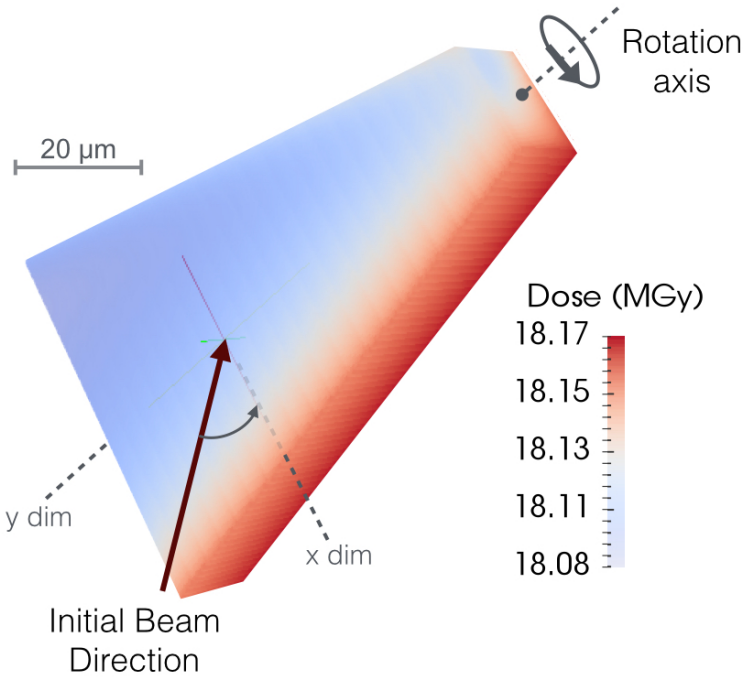


Figure 4.5: The dose distribution within the TRAP crystal calculated in *RADDOSE-3D* for the example case of the seventh dataset (DWD: 16.7 MGy). Other datasets exhibited qualitatively similar homogeneous dose distributions (data not shown), with lower than a 0.51 % variation predicted across the predicted across the crystal volume. The initial beam direction and rotation axis (concurrent with y-axis) are shown. The three-dimensional geometry of the TRAP-RNA crystal was originally created using modelling software package *BLENDER*, and the continuous dose gradient was rendered using the *ParaView* visualisation package by KitWare (*http://www.paraview.org/*).

of *heavy* elements (atomic mass larger than oxygen) occupying the solvent were included from the crystallisation conditions (i.e. potassium phosphate, potassium glutamate, KCl and $MgCl_2$ were included, see Section 4.2.1.2). The intensity profile of the beam was modelled as a uniform ('top-hat') distribution, and throughout the first nine collected data sets, the photon flux was estimated to be 5×10^{11} photons/s, but increased to 7.12×10^{11} photons/s following the beam refill at dataset 10. The resulting diffraction weighted dose (DWD) values (Zeldin *et al.*, 2013a) are reported in Table 4.1.

4.2.3 Data processing

Diffraction data were processed using a similar protocol to the C.Esp1396I-DNA complex described in Chapter 3. Spot finding and integration of diffraction images was performed independently for each dataset using *MOSFLM* (Leslie and Powell, 2007) and then unmerged intensity values were independently scaled and merged using the *POINTLESS*, *AIMLESS* and *CTRUNCATE*, as part of the data reduction pipeline in the CCP4i GUI (Winn *et al.*, 2011). The TRAP-

RNA macromolecule crystallized in space group C2, as determined by *POINTLESS*, and this is consistent with the previously deposited structure of identical composition (Hopcroft *et al.*, 2002). For each later dataset, the same 5 % set of randomly selected test reflections were extended from the first dataset in *AIMLESS* and used to calculate R_{free} during later refinement. For the first dataset, molecular replacement was then performed with the program *PHASER* (McCoy *et al.*, 2007), using a previously deposited TRAP structure bound to the same 53 bp RNA sequence (GAGUU-repeats) as in the current study (PDB: *1gtf*, resolution: 1.75 Å (Hopcroft *et al.*, 2002)) as a search model.

The TRAP-RNA model obtained from the first dataset was then refined using *REFMAC5* (Murshudov *et al.*, 2011), initially using 10 cycles of rigid-body refinement, followed by repeated rounds of restrained, TLS and isotropic *B*-factor refinement, coupled with manual inspection and addition/removal of solvent molecules in *COOT* (Emsley *et al.*, 2010). The final model derived from the first dataset was refined to 1.98 Å, with a dimeric assembly of the RNA-unbound and bound TRAP rings comprising the asymmetric unit. Whereas the repeating 53-mer GAGUU RNA sequence bound to TRAP physically consists of ten GAGUU repeats followed by one GAG triplet (i.e. it does not form a closed loop), consistent with previous crystal studies on the TRAP-RNA complex, the missing UU spacer after the eleventh GAG triplet was not detectable in the electron density of the crystal. Therefore, the first spacer U nucleotide was modelled at all 11 repeats around the TRAP ring. However there was insufficient $2mF_{\text{obs}} - DF_{\text{calc}}$ map electron density to model the second spacer U nucleotide, and it was thus omitted from the final refined structure at all 11 positions around the TRAP ring.

For the later datasets $n = 2, ..., 10$, coordinate models were generated by refining the first dataset coordinate model against the observed structure factor amplitudes $|\mathbf{F}_{\text{obs,n}}|$ with *phenix.refine* (Adams *et al.*, 2010), using only rigid body and isotropic *B*-factor refinement. A discussion into why independent restrained refinement of coordinate models at each higher dose was inappropriate has been presented in Section 2.5.3. In short, this was to avoid introducing dataset-specific bias into the corresponding higher dose electron density Fourier difference maps calculated as described below in Section 4.2.4. Rigid body refinement was performed to compensate for potential movement of the TRAP-RNA complex inside the crystal (due to a combination of predicted translation and rotation of the macromolecule, along with unit cell expansion caused by increasing

dose (Ravelli and McSweeney, 2000)). Isotropic *B*-factor refinement was included to account for the well recognised increasing relationship between atomic *B*-factor values and absorbed dose in a crystal (Weik *et al.*, 2000).

These higher dose coordinate models are required by the current implementation of the *RIDL* pipeline (introduced in Section 2.5), the results of which are presented later in Section 4.3 for the TRAP complex. As has been described previously in Section 2.5.5, the per-atom search radius employed by *RIDL* to sample each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map in the local region around each atom is *B*-factor-dependent, and thus the decision to include isotropic *B*-factor refinement directly affects these radii at each dose. However, as has previously been demonstrated in the testing of the *RIDL* pipeline in Section 2.5.9.3, the results of *RIDL* are relatively robust to the selected refinement strategy of higher dose datasets, and in particular the decision to include isotropic *B*-factor refinement.

A summary of the refinement statistics for each dataset can be found in Table 4.1. For each dataset, *MOLPROBITY4* (Chen *et al.*, 2010) structure validation was utilised to access the quality of the resulting macromolecular geometry, as presented in Table 4.2. Final refined models and corresponding structure factor amplitudes for each dataset were deposited in the PDB with accession codes: *5eeu*, *5eev*, *5eew*, *5eex*, *5eey*, *5eez*, *5ef0*, *5ef1*, *5ef2* and *5ef3*.

4.2.4 Generation of Fourier difference maps

The protocol presented in Section 2.5.4 was followed to calculate $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps between the increasingly damaged datasets and the initial *'low-dose'* dataset ($n = 1$). Quantitative analysis of radiation-induced electron density changes to individual atoms was performed using the *RIDL* pipeline, as presented in Section 2.5. *RIDL* permitted a systematic and objective assessment of damage across the large (~ 204 kDa) TRAP-RNA asymmetric unit, which was unachievable by visual inspection of $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps alone. In particular, the maximum density-loss metric, D_{loss}(atom), (with units of e/Å^3) was used to directly quantify electron density loss localised to specific atoms in the TRAP-RNA complex reproducibly throughout the unit cells with respect to the initial low-dose data set.

Table 4.1: Data processing and refinement statistics for the TRAP-RNA complex damage series. The current TRAP-RNA structure crystallised in space group C2 ($\alpha = \gamma = 90°$). Values in parentheses are for the highest-resolution shell (2.01 – 1.98 Å). The Wilson *B*-factor is estimated from the Wilson plot, as reported in each *AIMLESS* log file. For observed F_{obs} and calculated F_{calc} structure factors, $R_{\text{work}} = \frac{\sum_{hkl} |F_{\text{obs}} - F_{\text{calc}}|}{\sum_{hkl} F_{\text{obs}}}$ and R_{free} is the R_{work} formula calculated from a small (5 %) test set of randomly selected reflections that were excluded throughout refinement. Both R_{work} and R_{free} were calculated using *REFMAC*. The resolution range is 63.0 – 1.98 Å for all datasets. I_n/I_1 values were derived from the *AIMLESS* log files for each dataset, and calculated using Equation 3.2.1 as reported in Section 3.2.3 for the C.Esp1396I-DNA investigation. I_n/I_1 values have been reported as originally calculated (denoted '*observed*') and following correction to account for the differential beam flux at dataset 10 due to the beam refill between dataset 9 and 10 (denoted '*corrected*').

Dataset	1	2	3	4	5	6	7	8	9	10
PDB code	5eeu	5eev	5eew	5eex	5eey	5eez	5ef0	5ef1	5ef2	5ef3
DWD (MGy)	1.31	3.88	6.43	8.96	11.44	13.92	16.37	18.79	21.18	25.0
Data processing										
Cell dimensions										
a (Å)	140.90	141.00	141.04	141.04	141.12	141.14	141.16	141.19	141.21	141.23
b (Å)	110.89	110.98	111.02	111.05	111.08	111.10	111.13	111.17	111.16	111.16
c (Å)	137.81	137.93	137.99	138.05	138.09	138.15	138.18	138.21	138.28	138.29
β (°)	117.41	117.41	117.41	117.40	117.40	117.39	117.39	117.39	117.39	117.37
No. of observations	473080	471462	472091	472756	473225	476499	477213	477019	475519	472868
No. of unique reflections	129942	130202	130317	130454	130599	130738	130870	130920	130982	130840
Multiplicity	3.6 (3.5)	3.6 (3.4)	3.6 (3.4)	3.6 (3.4)	3.6 (3.4)	3.6 (3.4))	3.6 (3.4)	3.6 (3.4)	3.6 (3.4)	3.6 (3.4)
Completeness (%)	99.3 (98.2)	99.3 (98.2)	99.3 (98.1)	99.3 (98.1)	99.3 (98.0)	99.3 (98.1)	99.3 (97.9)	99.3 (97.9)	99.3 (97.9)	99.1 (97.6)
Mosaicity (°)	0.43	0.42	0.42	0.42	0.41	0.41	0.41	0.41	0.41	0.43
Wilson *B*-factor (Å^2)	26.1	27.3	28.2	29.3	30.8	32.8	34.7	36.3	38.5	39.6
R_{merge}	0.08 (0.70)	0.08 (0.78)	0.08 (0.86)	0.09 (1.02)	0.09 (1.13)	0.10 (1.31)	0.10 (1.50)	0.11 (1.68)	0.11 (2.07)	0.11 (1.98)
R_{meas}	0.11 (1.01)	0.11 (1.12)	0.11 (1.25)	0.12 (1.45)	0.13 (1.61)	0.13 (1.88)	0.14 (2.18)	0.15 (2.46)	0.15 (2.94)	0.15 (2.83)
$CC_{1/2}$	0.997 (0.824)	0.997 (0.772)	0.997 (0.731)	0.997 (0.610)	0.997 (0.542)	0.997 (0.456)	0.997 (0.363)	0.996 (0.300)	0.997 (0.267)	0.997 (0.227)
$\langle I/\sigma(I) \rangle$	9.6 (1.5)	9.4 (1.4)	9.3 (1.2)	8.8 (1.1)	8.4 (1.0)	8.1 (0.8)	7.4 (0.7)	7.0 (0.6)	7.0 (0.5)	7.2 (0.5)
I_n/I_1 (observed)	1.00	0.91	0.87	0.83	0.78	0.74	0.68	0.64	0.63	0.85
I_n/I_1 (corrected)	1.00	0.91	0.87	0.83	0.78	0.74	0.68	0.64	0.63	0.60
Refinement										
R_{work}	0.192	0.209	0.209	0.212	0.216	0.220	0.225	0.230	0.234	0.240
R_{free}	0.231	0.244	0.243	0.245	0.248	0.252	0.256	0.263	0.264	0.271
Mean isotropic *B*-factor (Å^2)	33.78	34.50	36.04	37.22	38.86	40.35	42.89	44.70	47.71	47.58

Table 4.2: Assessment of final macromolecular geometry for the refined coordinate model of the TRAP-RNA complex, corresponding to the first dataset, as reported by *MOLPROBITY4* (Chen *et al.*, 2010). Since for the coordinate models corresponding to the higher dose datasets, only rigid-body refinement was performed on the refined initial dataset coordinate model coupled with the higher dataset merged structure factor amplitudes, the reported geometry statistics were conserved for the higher dose datasets.

No. non-H atoms	13830
Protein atoms	12135
RNA	968
Water	743
Other	22× Trp ligands
RMSD bond length (Å)	0.024
RMSD bond angle (°)	2.276
Ramachandran analysis	
Favoured (%)	99.53
Outliers (%)	0.00
Allowed (%)	0.47
Rotamer outliers (%)	2.57
All atom clash-score	3.29

4.3 Results

4.3.1 Protein site-specific damage observations

Consistent with previous protein site-specific damage studies, the D_{loss} metric identified significant Glu and Asp side-chain decarboxylation events in TRAP (Figure 4.6 and 4.7). The main sequence of TRAP does not contain any Trp or Cys residues (and thus contains no disulphide bonds). Interestingly, the data in Figure 4.6 indicate that the substrate Trp ligands also exhibited disordering of the free terminal carboxyl groups at the highest doses ($\geq$ 21.9 MGy).

As shown in Figure 4.7, even for radiation-insensitive residues (e.g. Gly) the average D_{loss} increased with dose. This was an effect of global radiation damage, since as the dose increases the electron density associated with each refined atom becomes weaker as the atomic occupancy decreases. Only Glu & Asp residues and ordered solvent molecules exhibited a rate of D_{loss} increase that consistently exceeded the average increase in D_{loss} for the whole TRAP-RNA asymmetric

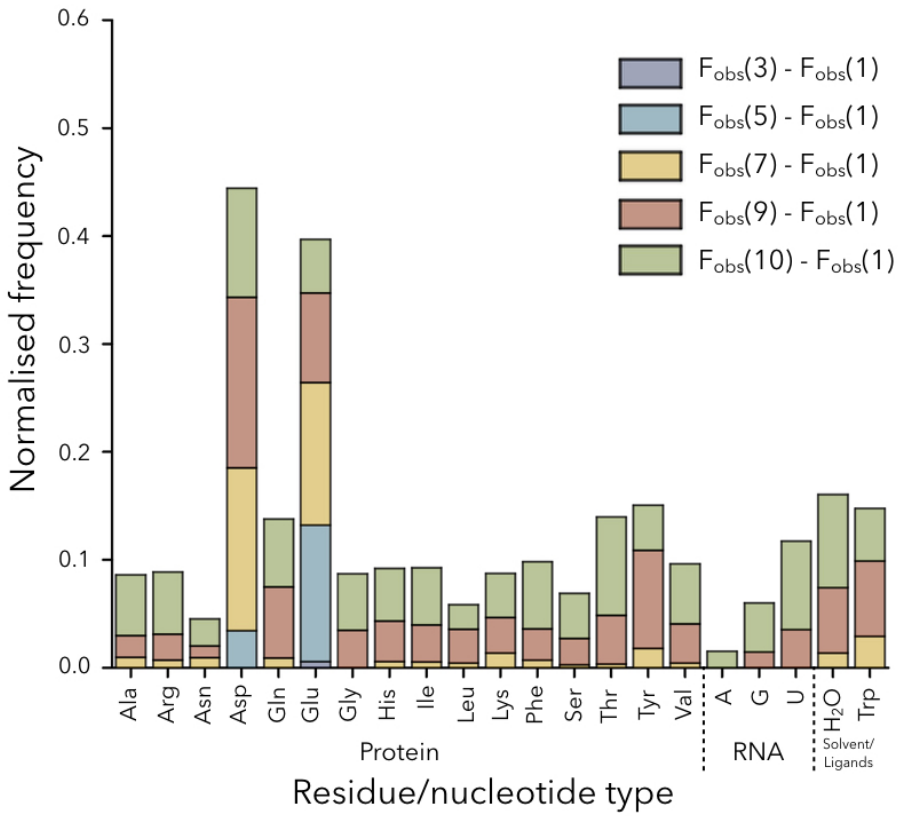


Figure 4.6: Top electron-density loss sites as indicated by the D_{loss}(atom) metric for the TRAP-RNA complex asymmetric unit, grouped by residue/nucleotide type for five increasing doses. Here, only sites above an arbitrarily chosen threshold of 4× the average D_{loss} calculated over all atoms in the TRAP-RNA structure for the first difference map ($F_{\text{obs},2} - F_{\text{obs},1}$) have been included, in order to isolate only the top damage sites throughout the asymmetric unit. Cumulative frequencies have been normalised to both the total number of non-hydrogen atoms per residue/nucleotide and the total number of that residue/nucleotide type present in the asymmetric unit.

unit at each dose (Figure 4.7, dashed line). Moreover, the rate of D_{loss} increase was marginally larger for Glu compared with Asp residues over the large dose range, consistent with density functional theory calculations that have suggested that CO_2 loss is more exothermic by around $8\,\text{kJ}\,\text{mol}^{-1}$ from oxidized Glu residues than from their Asp counterparts (I. Carmichael, private communications).

4.3.2 RNA is less susceptible to electron-density loss than protein

Visual inspection of Fourier difference maps indicated a clear lack of RNA electron-density degradation with increasing dose compared with the obvious protein damage manifestations (Figure 4.8). Only at the highest doses investigated ($>$ 20 MGy) could electron density loss be distinguished at the RNA phosphate and C−O bonds of the phosphodiester backbone, above the

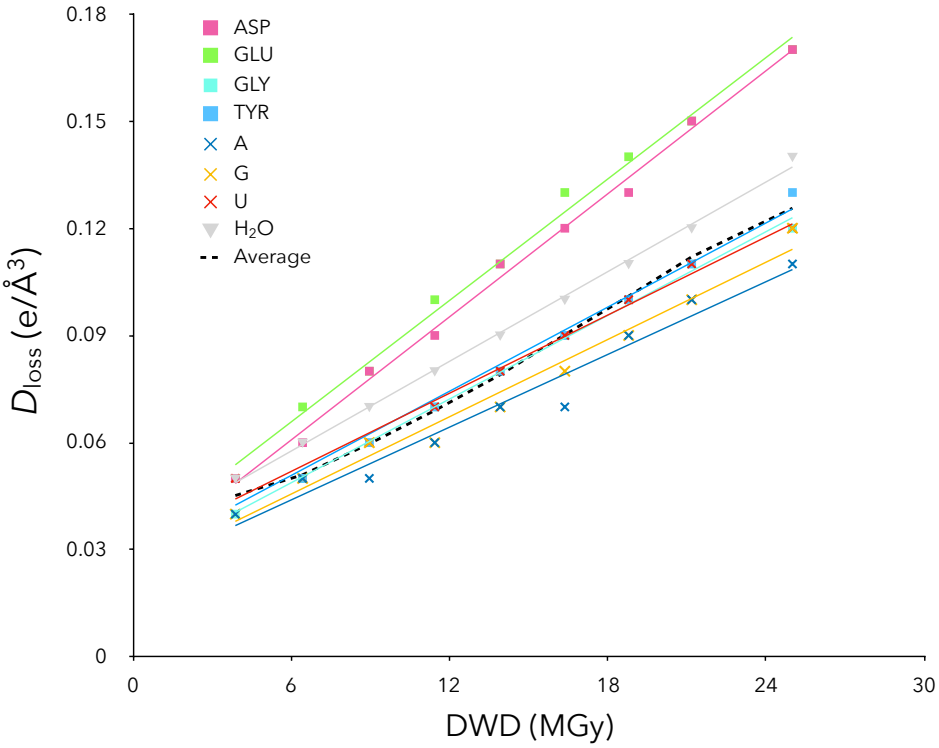


Figure 4.7: Average D_{loss} for each residue/nucleotide type with respect to diffraction-weighted dose (DWD). Only a subset of key TRAP residue and nucleotide types (and ordered solvent atoms) are included for clarity. For each residue/nucleotide, linear regression has been performed to provide a line of best fit against dose. The average D_{loss} calculated over all refined protein, RNA and ordered solvent atoms in the TRAP asymmetric unit is also shown at each dose (dashed black line).

background noise level of each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map*, but only at a few positions (e.g. U-119 in Figure 4.8d). The majority of the the 53 bp RNA backbone did not exhibit any clear electron density loss that could be attributed to radiation damage. Indeed, the median D_{loss} value was lower by a factor of > 2 for RNA P atoms than for Glu and Asp side-chain groups at 25.0 MGy (Figure 4.9), and furthermore was indistinguishable from that of Gly-C_α atoms, which are not believed to be radiation-sensitive at these doses in MX at 100 K.

Figure 4.10 indicates that the D_{loss} metric attained for each U4 nucleotide phosphate group was marginally higher than for the other refined nucleotide types (G1, A2 and G3). Since U4 is the only refined nucleotide not to exhibit significant base-protein interactions around TRAP (with a water-mediated hydrogen bond detected in only three of the eleven subunits and a single Arg-58 hydrogen bond suggested in a further four subunits), this increased U4 D_{loss} can be explained by its greater flexibility.

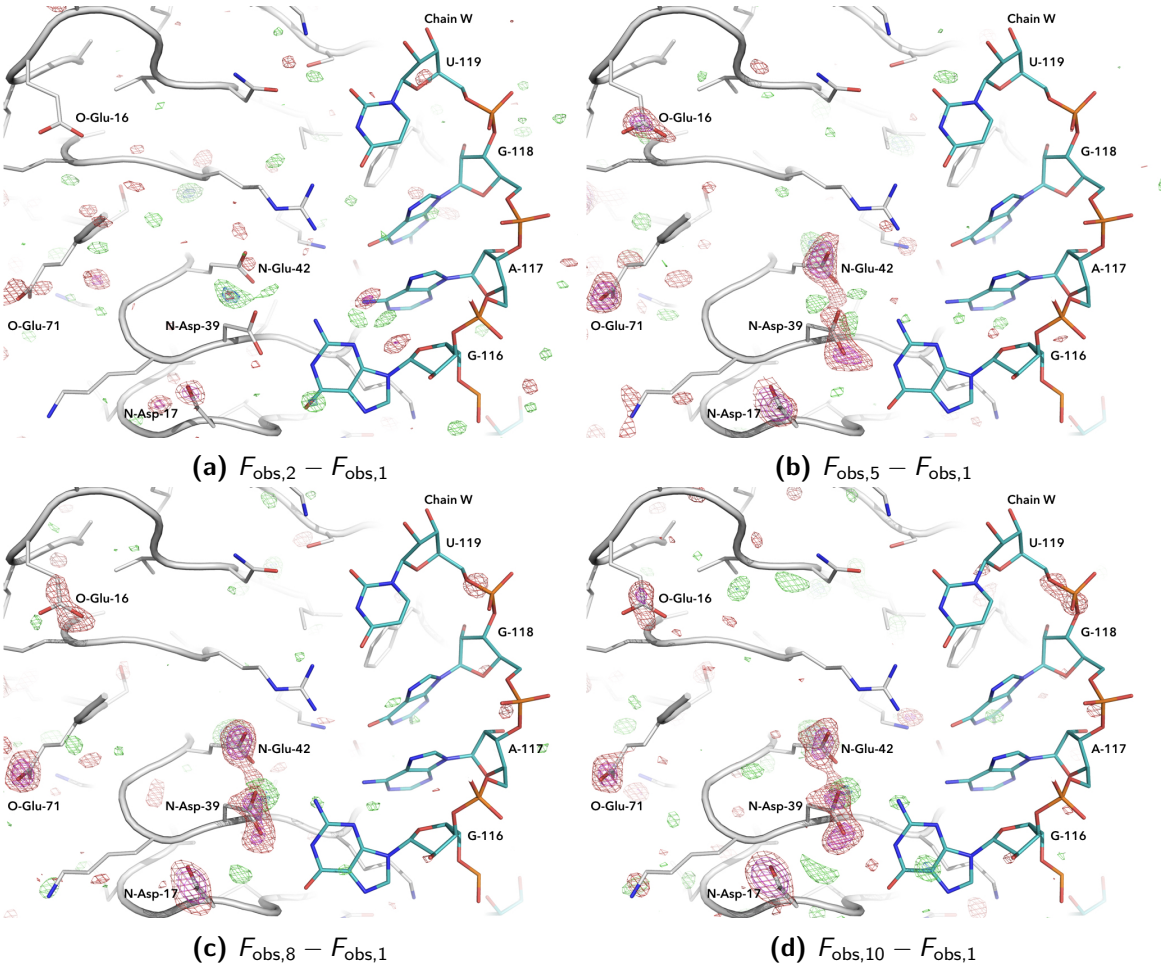
RNA backbone disordering thus appears to be the main radiation-induced effect in RNA, with the protein-base interactions maintained even at high doses (> 20 MGy). However, little evidence has been provided for significant RNA bond breakage events even at high dose values. It was thus concluded that RNA, similarly to DNA (as shown in Chapter 3), was highly radiation-insensitive relative to protein in crystalline nucleoproteins at 100 K.

4.3.3 Distribution of damage within the TRAP-RNA complex: investigation of Glu and Asp residues

Due to the existence of both a TRAP protein ring bound to RNA and another non-bound TRAP ring in the asymmetric unit, the TRAP-RNA damage datasets provided an invaluable system to investigate whether the act of RNA binding to a protein could induce altered specific damage dynamics, through controlled comparison to the non-bound TRAP ring.

For the large number of acidic residues per TRAP ring (4 Asp and 6 Glu residues per protein monomer, 220 in the asymmetric unit), a strong dependence of decarboxylation susceptibility on local environment was observed, as presented for the D_{loss} and density-weighted D^{-}_{mean} metrics in

*The background map noise level of each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map was deemed to be at approximately $|\sigma| < 3$, as judged by visual inspection over the asymmetric unit.



(a) $F_{\text{obs,2}} - F_{\text{obs,1}}$ **(b)** $F_{\text{obs,5}} - F_{\text{obs,1}}$

(c) $F_{\text{obs,8}} - F_{\text{obs,1}}$ **(d)** $F_{\text{obs,10}} - F_{\text{obs,1}}$

Figure 4.8: Radiation damage sites in the TRAP-RNA complex detectable in $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps, for $n = 2, 5, 8, 10$. The RNA-binding interface between TRAP chain N and the 53 bp RNA (chain W) is shown. Nucleotides G-116, A-117, G-118 comprise a single G1-A2-G3 repeat of the $(\text{GAGUU})_{10}$GAG RNA sequence. $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps are contoured at $\pm 3\sigma$ in green/red and $\pm 4\sigma$ in blue/magenta. Due to the 11-fold symmetry around the TRAP ring, the Glu & Asp decarboxylation events were also observed at other symmetry-related positions; the electron density loss exhibited at U-119 was less conserved around the RNA strand.

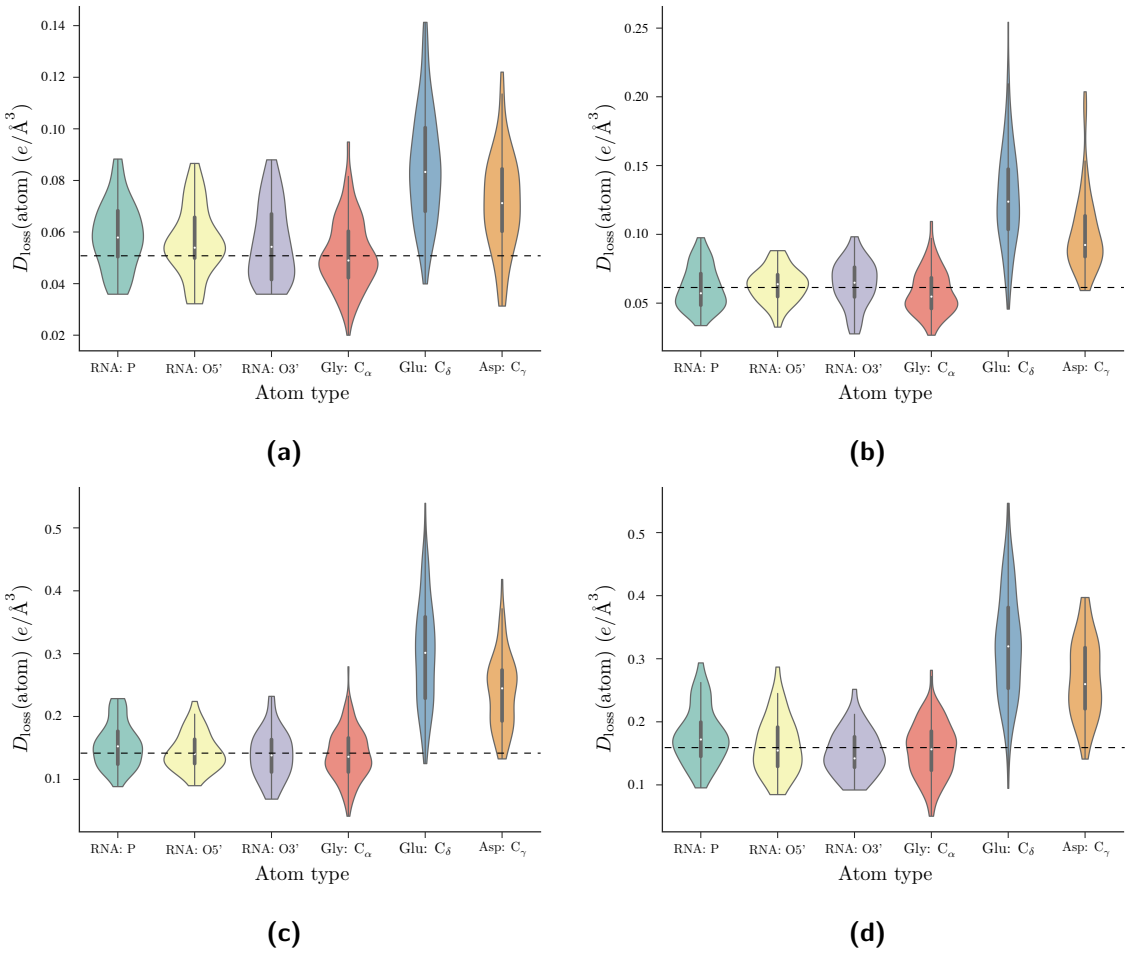
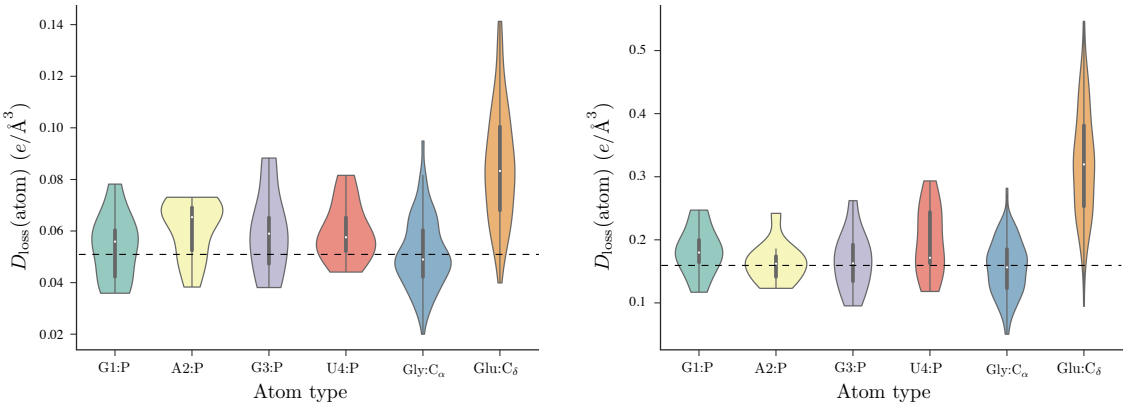


Figure 4.9: Violin plots illustrating the distribution D_{loss} metric calculated for all RNA P, O_5' and O_3' atoms, and all Glu-C_δ, Asp-C_γ and Gly-C_α atoms over 4 increasing doses (a) 3.88 MGy, (b) 6.45 MGy, (c) 21.9 MGy and (d) 25.0 MGy. The D_{loss} calculated for Glu-C_δ and Gly-C_α atoms are included for comparison in each plot. Each horizontal dashed line indicates the average D_{loss} calculated over all refined atoms in the TRAP complex, as a measure of the overall rate of density disordering due to global radiation damage effects. Violin plots have been calculated using the *Seaborn* package in *Python*. Each violin plot is comprised of a kernel density plot for a given atom type, indicating the underlying theoretical distribution that each set of metric values is predicted to follow. The violin plot is related to a box plot, however the x-axis width of the violin at a particular y-axis value is proportional to the probability that a randomly picked atom attains that y-axis value. For each violin, the median (white circle) and interquartile range (grey rectangle) are also provided. Each violin has been truncated to the minimum and maximum metric values attained for each atom.



(a) **(b)**

Figure 4.10: Violin plots illustrating the distribution D_{loss} metric calculated for all phosphorus atoms for each of 4 refined nucleotide types: G1, A2, G3 and U4 at (a) dataset $n = 2$ (3.88 MGy) and (b) at the highest investigated dose (25.0 MGy). The D_{loss} calculated for Glu-C_δ and Gly-C_α atoms are included for comparison in each plot. Each horizontal dashed line indicates the average D_{loss} calculated over all refined atoms in the TRAP complex, as a measure of the overall rate of density disordering due to global radiation damage effects. Violin plots have been calculated using the *Seaborn* package in *Python* in an identical procedure to that in Figure 4.9. For each violin, the median (white circle) and interquartile range (grey rectangle) are provided. Each violin has been truncated to the minimum and maximum metric values attained for each atom.

Figure 4.12 and 4.13 respectively (calculated using the *RIDL* pipeline, as detailed in Section 2.5). As has previously been discussed in Section 2.6, both these metrics provide useful, yet distinct, indicators of radiation-induced electron density loss to individual atoms. Whereas the D_{loss} metric provides a *worst case scenario* for an atom, by reporting the maximum density loss in the localised space around that atom, the density-weighted version of the D^{-}_{mean} metric is a weighted integration of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density around an atom. D^{-}_{mean} is therefore affected by the spatial distribution of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density local to an atom. Although the D^{-}_{mean} metric is therefore less sensitive to statistical outliers than D_{loss}, it is more dependent on the quality of built model. It is thus reasonable to present both metrics here for direct comparison.

Acidic residues in TRAP could be split into three disjoint categories (Figure 4.11): for acidic residues with no differing interactions between non-bound and bound TRAP (Asp-8, Glu-16, Asp-17 and Asp-29), similar damage was apparent between the two rings in the asymmetric unit, as expected. TRAP residues directly on the RNA-binding interfaces exhibited greater damage accumulation in non-bound TRAP (Glu-36, Asp-39 and Glu-42). Finally, for residues at the ring-ring interfaces (where crystal contacts were detected) bound TRAP exhibited enhanced damage accumulation (Glu-50, Glu-71 and Glu-73).

To test whether the differences in the extracted damage metric values where statistically significant

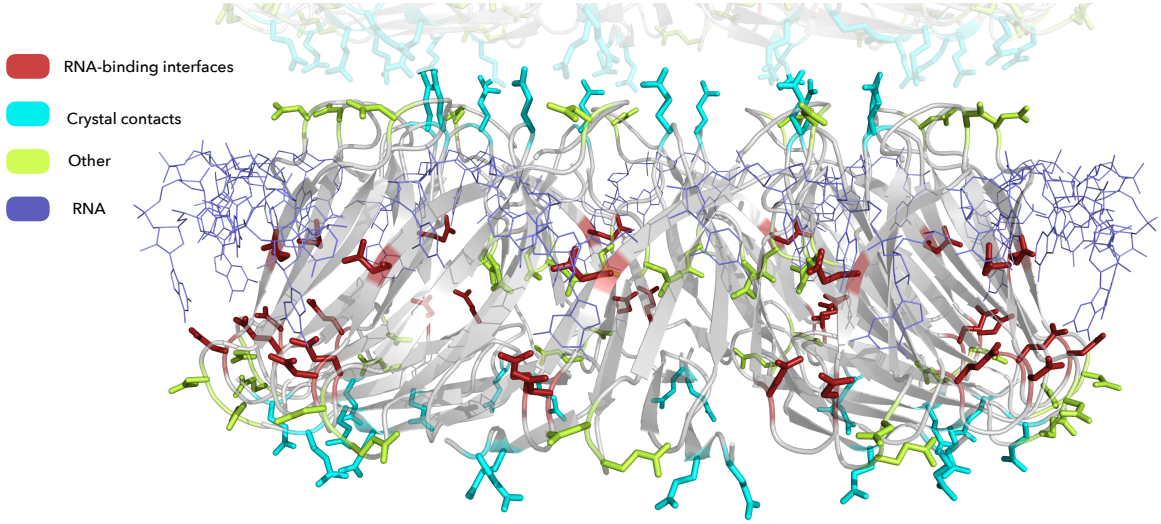


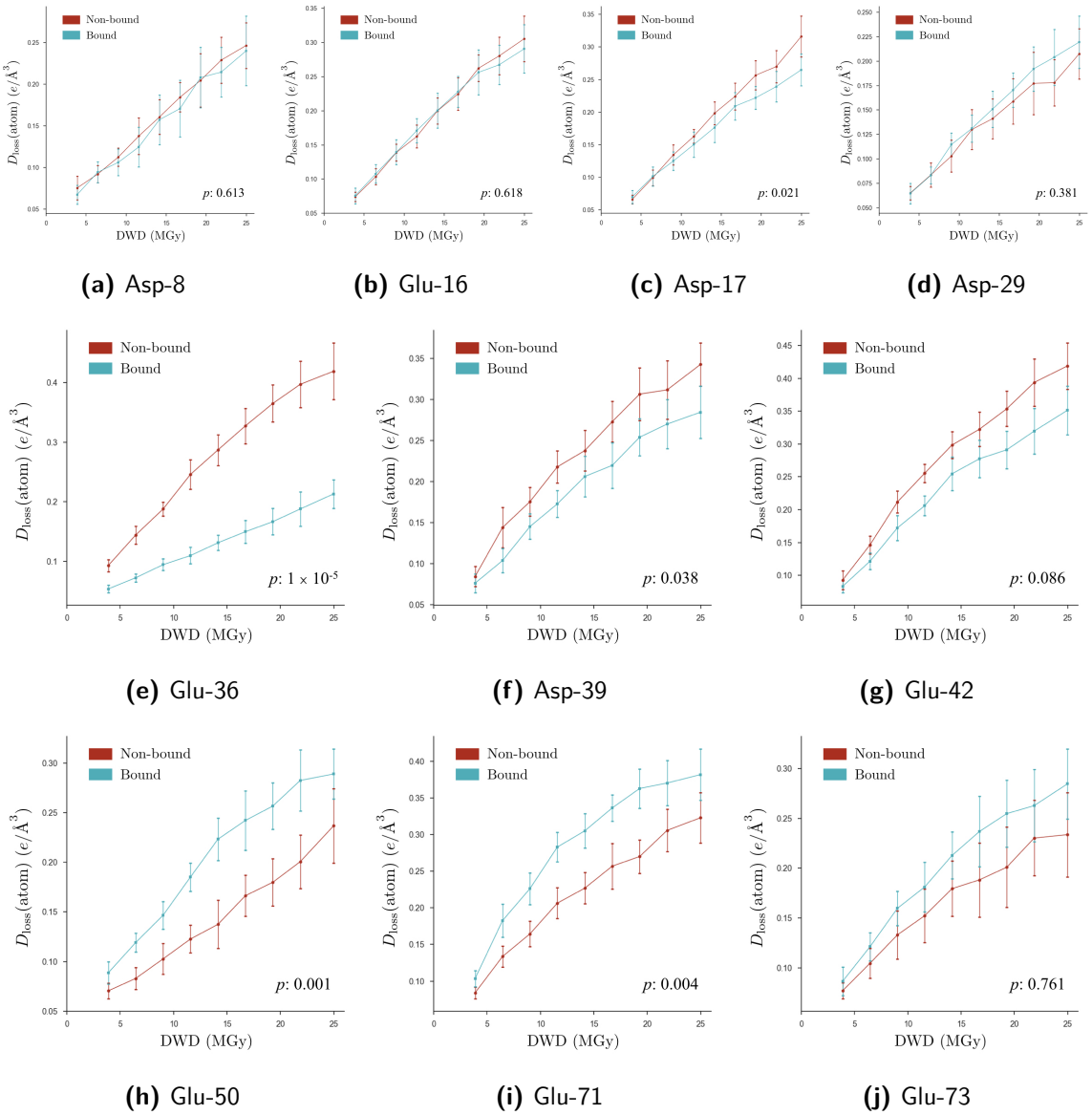
Figure 4.11: Schematic illustrating the distribution of Glu and Asp groups dispersed around the RNA-bound TRAP ring. Glu and Asp residues involved in RNA-binding interactions are coloured red. Residues suspected to be involved in inter-molecular crystal contacts (as determined by the CCP4 program *AREAIMOL*) are coloured cyan. All other Glu and Asp residues are coloured green. The single stranded 53 bp RNA sequence is shown as purple sticks.

in Figure 4.12 and 4.13, Hotelling's two-sample T^2 test (the multivariate counterpart of the Student's t-test) was used to test the null hypothesis, H_0, that the mean values of each radiation damage metric were equal for the bound and non-bound groups. A significance level of $\alpha = 0.05$ was used throughout this investigation to reject the null hypothesis H_0, and p-values have been presented in both Figure 4.12 and 4.13.

Section 4.3.4 below will present a detailed investigation into the damage dynamics in each RNA-binding interface. Section 4.3.5 will then probe the differential radiation-sensitivity of residues undergoing crystal contact interactions.

4.3.4 RNA protects residues at the RNA-binding interfaces

Since the Glu-36, Asp-39 and Glu-42 are positioned close to each other at each RNA-binding interface ($\leq 6-10$ Å distance between side-chain carboxyl oxygens of each residue), the differing interactions exhibited by these three residues appeared to have direct implications on their specific damage susceptibility (Figure 4.13).

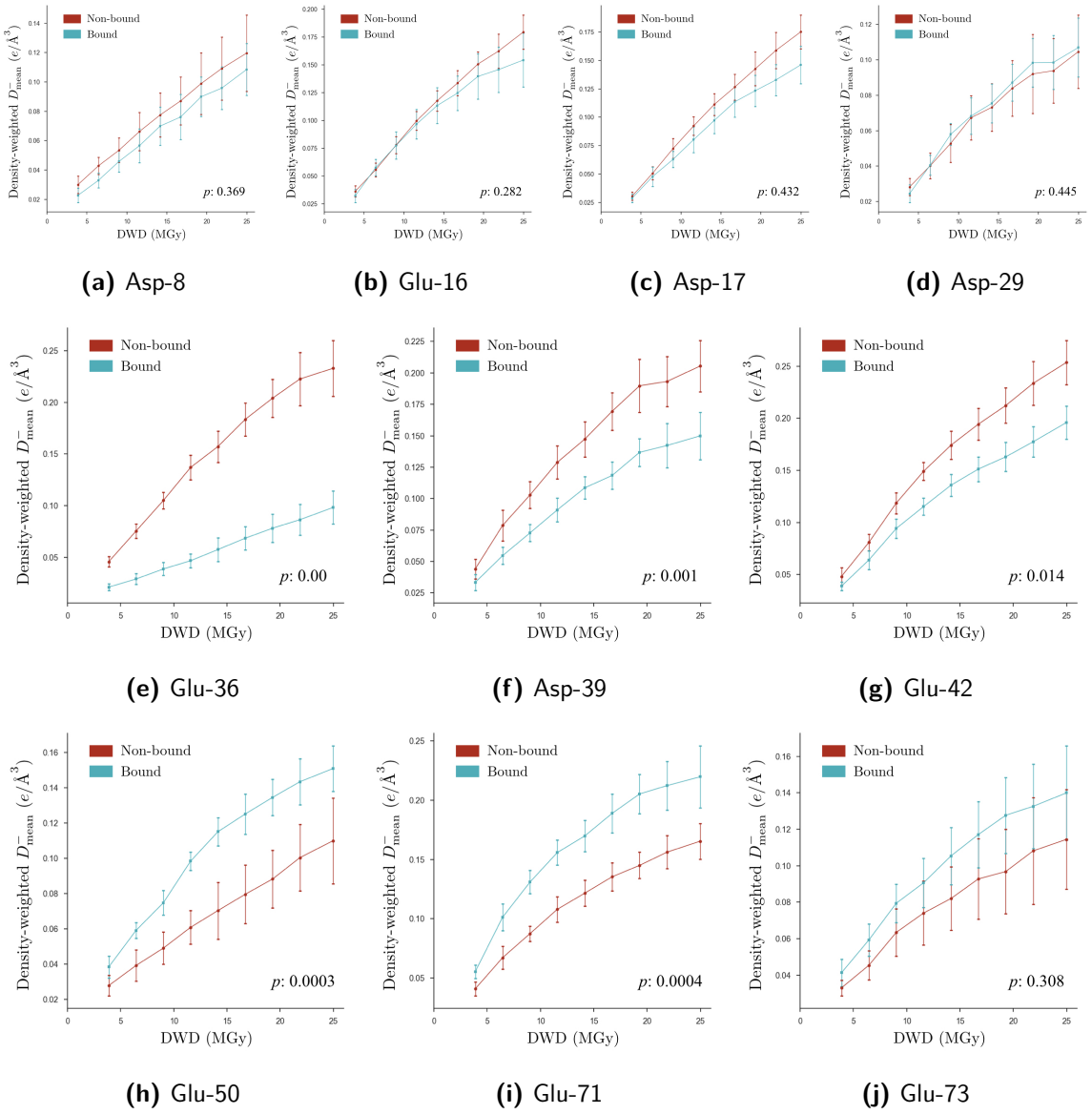


(a) Asp-8 **(b)** Glu-16 **(c)** Asp-17 **(d)** Asp-29

(e) Glu-36 **(f)** Asp-39 **(g)** Glu-42

(h) Glu-50 **(i)** Glu-71 **(j)** Glu-73

Figure 4.12: D_{loss} against dose for Glu and Asp carboxyl side-groups in TRAP. For each carboxyl group (e.g. Glu-36 of chain N) the D_{loss} metric value corresponding to the three atoms comprising the carboxyl group (e.g. C_δ, $O_{\epsilon 1}$, $O_{\epsilon 2}$) have been averaged to provide a single representative D_{loss} value for each group. Residues have then been split into bound and non-bound groupings, with such that each point representing the mean calculated over 11 equivalent instances of a residue (e.g. Glu-36) around each TRAP ring. Error bars indicate 95 % confidence intervals.



(a) Asp-8 **(b)** Glu-16 **(c)** Asp-17 **(d)** Asp-29

(e) Glu-36 **(f)** Asp-39 **(g)** Glu-42

(h) Glu-50 **(i)** Glu-71 **(j)** Glu-73

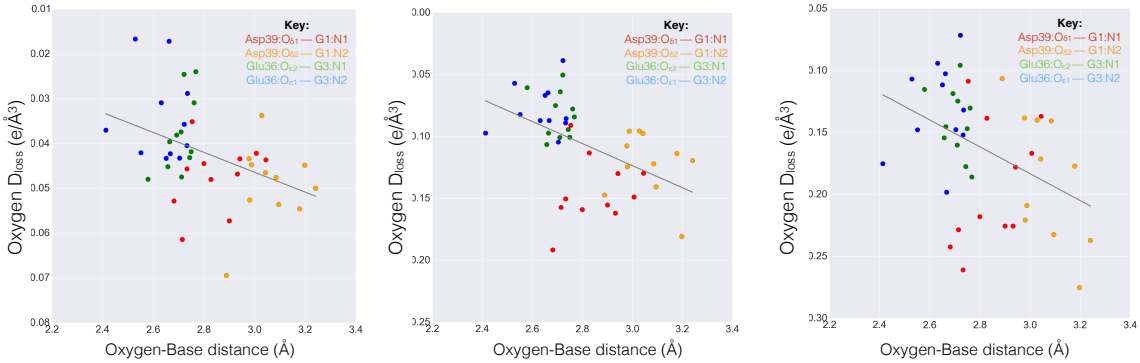
Figure 4.13: The density-weighted D^-_{mean} metric against dose for Glu and Asp carboxyl side-groups in TRAP. For each carboxyl group (e.g. Glu-36 of chain N) the D_{loss} metric value corresponding to the three atoms comprising the carboxyl group (e.g. C_δ, $O_{\epsilon 1}$, $O_{\epsilon 2}$) have been averaged to provide a single representative metric value for each group. Residues have then been split into bound and non-bound groupings, with such that each point representing the mean calculated over 11 equivalent instances of a residue (e.g. Glu-36) around each TRAP ring. Error bars indicate 95 % confidence intervals. Density-weighted D^-_{mean} metric values have been calculated by *RIDL* using the formulation previously described in Section 2.5.7.

4.3.4.1 Relative RNA-protection to Glu-36 and Asp-39

The reduction in D_{loss} upon RNA-binding was most significant for Glu-36 (Hotelling T^2 test p-value: 1×10^{-5} in Figure 4.12e, i.e. H_0 rejected). For each TRAP ring subunit, the Glu-36 side-chain carboxyl group accepts a pair of similar length hydrogen bonds from the two N atoms of the G3 RNA base, with mean interaction distances of 2.6 Å for $O_{\epsilon 1}$-N2 and 2.7 Å for $O_{\epsilon 2}$-N1. In the current analysis, Asp-39 in the TRAP-$(GAGUU)_{10}$GAG structure appears to exhibit two distinct hydrogen bonds to the G1 base in each of the eleven TRAP-RNA interfaces, as does Glu-36 to G3. However, the reduction in density disordering upon RNA binding was far less significant for Asp-39 than for Glu-36 (Figure 4.12f, p-value: 0.038). In order to elucidate the mechanism underlying the superior protection to Glu-36, it was crucial to compare the local interactions of Glu-36 and Asp-39.

The mean distances for the Asp-39-G1 interaction are consistently longer than for the Glu-36-G3 interaction, at 2.9 Å and 3.1 Å for $O_{\delta 1}$-N1 and $O_{\delta 2}$-N2. However, the accuracy of these bond length measurements is limited by the $\sim$ 2 Å resolution of the current data. It was thus unsurprising that no direct quantitative correlation could be established between hydrogen-bond length and D_{loss} (linear $R^2 < 0.23$ for all doses; Figure 4.14), and insufficient data have been presented here to conclude that the hydrogen bond strength alone dictates the relative radiation-sensitivities of Glu-36 and Asp-39.

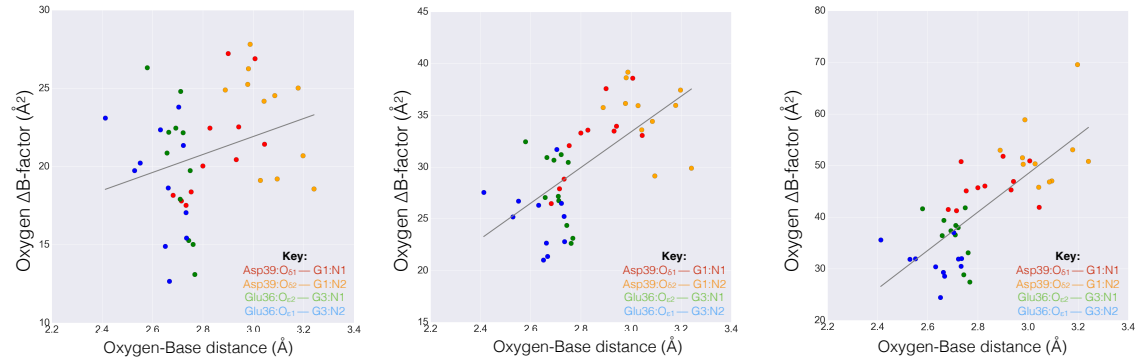
Nonetheless, previous studies have indicated that the Asp-39-G1 interaction is weak relative to the stronger Glu-36-G3 interaction, and could explain why only Glu-36 is significantly protected. In particular, whereas nucleoside analogue studies have demonstrated that guanine replacement with inosine at position G3 within each repeating GAG triplet resulted in $>$ 1000-fold decrease on TRAP binding affinity, inosine substitution at position G1 resulted in less than a two-fold decrease in binding affinity (Elliott *et al.*, 1999). Indeed, in other crystallographic studies on TRAP bound to differing RNA sequences: $(GAGAU)_{10}$GAG (PDB: *1c9s*, Antson *et al.* (1999)), $(GAGUUU)_{10}$GAG (PDB: *1utd*, Hopcroft *et al.* (2004)) and $(GAGCC)_{11}$G (PDB: *1gtn*, Hopcroft *et al.* (2002)), Asp-39 was documented to exhibit only a single distorted hydrogen bond with the NH_2 group of G1, and it has been suggested that the spacer nucleotides present between each GAG binding motif ultimately dictate the orientation of G1 (Figure 4.15). Furthermore, in the case of



(a) 3.88 MGy
R^2: 0.165

(b) 14.2 MGy
R^2: 0.229

(c) 25.0 MGy
R^2: 0.173



(d) 3.88 MGy
R^2: 0.0816

(e) 14.2 MGy
R^2: 0.416

(f) 25.0 MGy
R^2: 0.553

Figure 4.14: (a-c) D_{loss} and (d-f) ΔB-factor (change in atomic B-factor between dataset n and 1) metrics calculated for Asp-39 and Glu-36 side-chain carboxyl oxygen atoms, plotted against hydrogen bonding distance (Å) to the nearest G1 and G3 base respectively. Plots are given for three increasing doses between 3.88 MGy and 25.0 MGy. Linear regression R^2 values for each scatter plot are given. An increasing correlation with dose was observed between oxygen-base hydrogen bond length and ΔB-factor ($R^2 > 0.41$ for doses ≥ 11.6 MGy). No direct quantitative correlation could be established between hydrogen bond length and D_{loss} (linear $R^2 < 0.23$ for all doses).

the TRAP-$(UAGAU)_{10}$AU structure (PDB: *1utf*), Hopcroft *et al.* (2004)), where G1 is substituted for uracil (U1), nucleotide U1 was disordered at 19/22 instances in the asymmetric unit around two TRAP rings, without any loss in high RNA-binding affinity of TRAP ($K_d \simeq 1.1 \pm 0.4$ nM).

4.3.4.2 A proposed mechanism for RNA-protection of Glu-36

In the light of the results in the above section, for the current TRAP structure it is noted that the average solvent-accessible area[†] of the Glu-36 side-chain $O_{\epsilon 1}$ and $O_{\epsilon 2}$ atoms is reduced from 15 Å^2 to 0 Å^2 when bound to RNA. It is thus suggested that with negligible solvent accessibility Glu-36 decarboxylation would be inhibited, with steric hindrance preventing each radicalised Glu-36 CO_2 group from achieving the planar conformation required for complete dissociation from TRAP. This would extend the duration of time for which any migrating free electrons could refill an electron loss centre at the Glu-36 side-chain (the proposed precursor for Glu decarboxylation) and return Glu-36 to a non-ionised state.

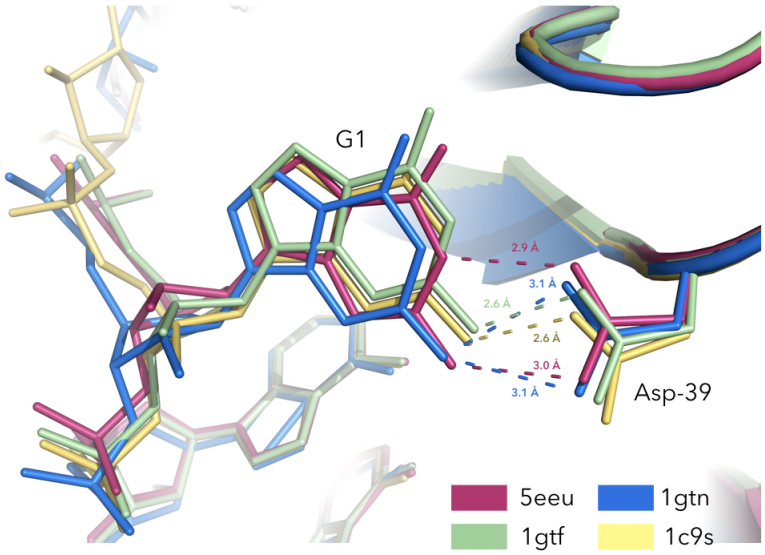
Upon RNA binding, the Asp-39 side-chain carboxyl group solvent-accessible area changes from 75 Å^2 to 35 Å^2, still permitting a high CO_2-formation rate.

4.3.4.3 The role of RNA binding on other acidic residues

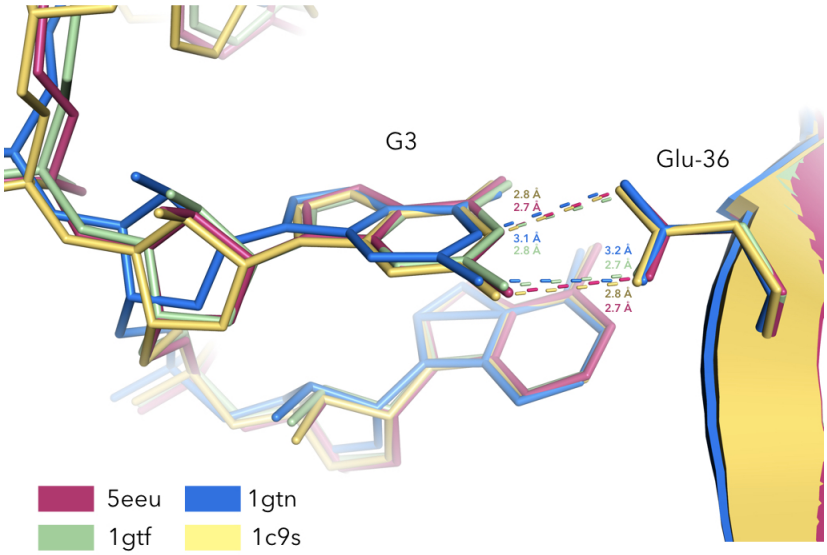
The radiation-sensitivity of Glu-42 was also reduced in the presence of RNA (Figure 4.12g and 4.13g), however Glu-42 does not bind directly to any nucleobase. The $O_{\epsilon 1}$ oxygen of Glu-42 forms a water-mediated hydrogen bond to basic subunit A2 at each TRAP RNA-binding interface, whilst the other ($O_{\epsilon 2}$) is involved in a salt-bridge interaction with Arg-58 (Hopcroft *et al.*, 2002; Antson *et al.*, 1999). It is further noted that the salt-bridge interaction is present in both the RNA-bound and non-bound TRAP rings, however the water interacting with $O_{\epsilon 1}$ oxygen is only present in RNA-bound TRAP.

A significant difference was observed between the D_{loss} dose-dynamics for the non-bound/bound Glu-42 $O_{\epsilon 1}$ atoms (Figure 4.16c, p-value: 0.007)) but not for the Glu-42 $O_{\epsilon 2}$ atoms (Figure 4.16d, p-value: 0.239). The presence of the salt-bridge interaction to Arg-58 in both TRAP rings thus

[†]Solvent accessibility has been calculated using the program *AREAIMOL* as part of the CCP4 software suite, whilst accounting for symmetry related copies of TRAP in adjacent asymmetric units of the crystal.



(a)



(b)

Figure 4.15: (a) Rendering in *PYMOL* to illustrate (a) the variation in hydrogen bonding interactions between TRAP residue Asp-39 and the G1 RNA nucleobase for TRAP bound to different RNA sequences. (b) The conservation of hydrogen bonding interactions between Glu-36 and nucleobase G3 for different RNA sequences has been shown for comparison. A representative binding pocket (protein chain P in the current TRAP-RNA structure) has been illustrated. Structures have been retrieved from the PDB with deposition codes: *1gtf* (in green, RNA sequence: $(GAGUU)_{10}GAG$), *1c9s* (in yellow, RNA sequence: $(GAGAU)_{10}GAG$) and *1gtn* (in green, RNA sequence: $(GAGCC)_{11}G$). The TRAP-RNA structure solved in the current investigation is shown is magenta for comparison (PDB code: *5eeu*, RNA sequence: $(GAGUU)_{10}GAG$). Hydrogen bonds under 3.2 Å between the G1 base and Asp-39 have been determined in *PYMOL* for each structure. Sequence and structural alignment has been performed for all structures, relative to the *5eeu* TRAP-RNA structure solved in the current investigation.

appeared to dictate the overall rate of decarboxylation of the full Glu-42 CO_2 group. Indeed, salt-bridge interactions have previously been suggested to reduce the glutamate decarboxylation rate for the large ($\sim$ 62.4 kDa) myrosinase protein structure (Burmeister, 2000).

It would intuitively be expected that the behaviour of D_{loss} for both oxygens in Glu-42 should be the same, however since D_{loss} takes into account both disordering of individual atoms as well as full decarboxylation, this is not necessary. It is suggested that the significant difference in D_{loss} between the non-bound/bound Glu-42 $O_{\epsilon 1}$ atoms may arise due to the absence of any hydrogen-bonded water to the $O_{\epsilon 1}$ atoms in non-bound TRAP, such that each $O_{\epsilon 1}$ atom may then be more susceptible to ionisation, or free to spatially disorder whist preserving the Glu-42 $O_{\epsilon 2}$ salt bridge interaction to Arg-58.

The RNA-stabilisation effect on acidic residues is observed only at short ranges and is restricted to within the RNA-binding interfaces around TRAP. For example, Asp-17 is located $\sim$ 6.8 Å from each G1 base, but orientated outside of each RNA-binding interface. Whereas the D_{loss} dose-dynamics between RNA-bound and non-bound TRAP were statistically dissimilar (Figure 4.12c, p-value: 0.021), this was not true for other metrics quantifying radiation-induced electron density loss (e.g. for the density-weighted D^{-}_{mean} metric in Figure 4.13c, p-value: 0.432), and it was unclear whether any RNA-mediated effect was present.

4.3.4.4 RNA-stabilisation of non-acidic residues

The RNA-stabilising effect was not restricted to radiation-sensitive acidic residues. With increasing dose, the D_{loss} associated with the Phe-32 side-chain was significantly reduced upon RNA binding (Figure 4.16e, p-value: 0.0014). The aromatic ring of Phe-32 stacks against the G3 base in the eleven TRAP RNA-binding interfaces (Antson *et al.*, 1999), and it is evident such stacking hinders radiation-induced conformation disordering of Phe-32. Similarly, the extended aliphatic side-chain of Lys-37 stacks against the nearby G1 base, making a series of nonpolar contacts in each RNA-binding interface. The D_{loss} for Lys-37 side-chain atoms was also reduced when held in this conformation (Figure 4.16f, p-value: 0.0243).

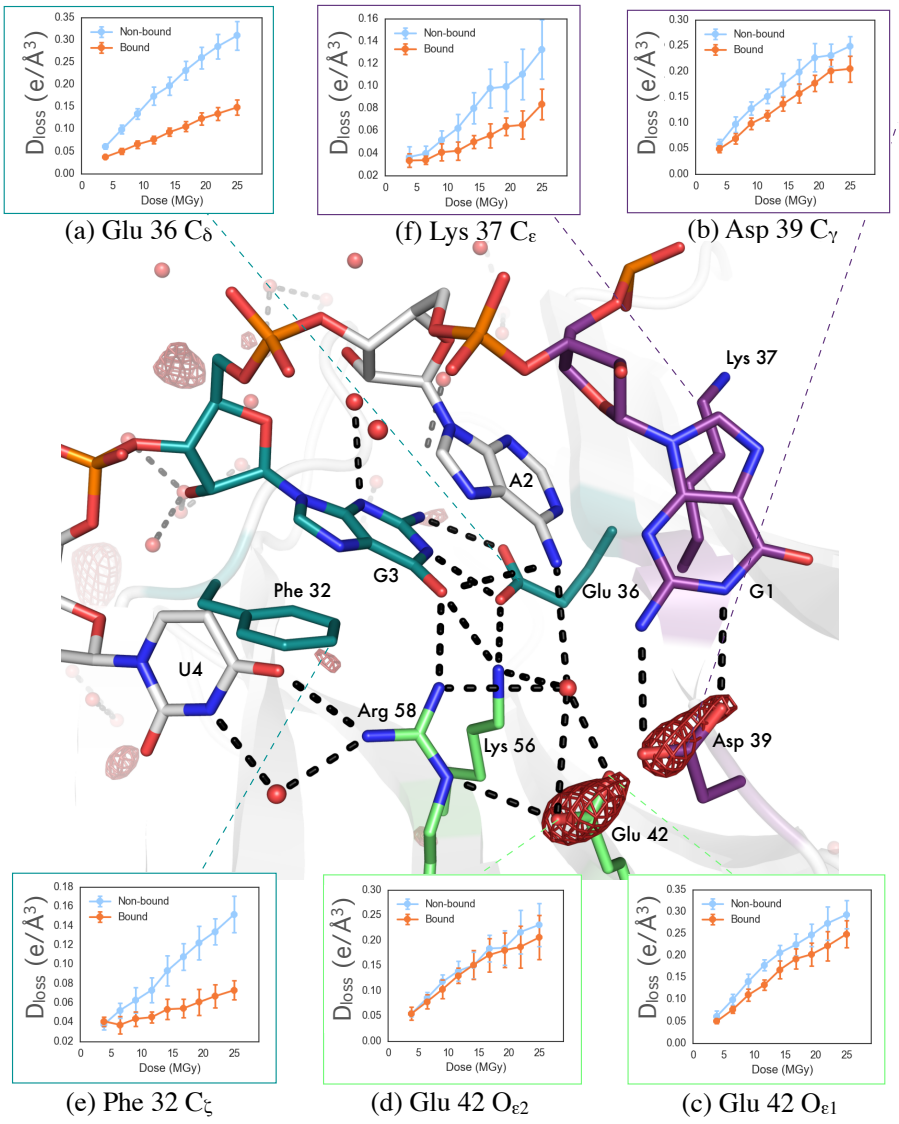


Figure 4.16: D_{loss} metric against dose for (a) Glu-36 C_{δ},(b) Asp-39 C_{γ},(c) Glu-42 $O_{\epsilon 1}$,(d) Glu-42 $O_{\epsilon 2}$,(e) Phe-32 C_{ζ} and (f) Lys-37 C_{ϵ} atoms. 95 % confidence intervals are included for each set of eleven equivalent atoms grouped as bound/non-bound. The RNA-binding interface interactions are shown for TRAP chain N, with the $F_{\text{obs},7} - F_{\text{obs},1}$ Fourier difference map (dose: 16.7 MGy) overlaid and contoured at a $\pm 4\sigma$ level.

4.3.5 The role of crystal contacts

The CCP4 program *NCONT* indicated that a total of 344 non-water inter-TRAP ring contacts ($\leq$ 4.5 Å) were present between the bound and non-bound TRAP rings inside the asymmetric unit, and 675 non-water crystal contacts ($\leq$ 4.5 Å) were present between adjacent asymmetric units. Of the acidic residues, Glu-50, Glu-71 and Glu-73 repeatedly exhibited such interactions. The presence of RNA was seen to significantly enhance the rate of decarboxylation for these 3 residue types (Figure 4.12h-j). The aim of this section is to elucidate whether the RNA plays any role in this enhancement.

Due to the C2 space group symmetry and crystal packing, for each of these residues the exact crystal contacts are not preserved between bound and non-bound TRAP or even between monomers in one TRAP ring (Figure 4.17). As such, it was challenging to detect any clear dependence between types of crystal contacts and damage. For example, on the RNA-unbound TRAP ring, due to the orientation of the TRAP rings packed within the crystal, Glu-50 exhibits varying interactions, with a possible Glu-50-Arg-31 salt bridge interaction present at only several chains (A,C and E). However, in RNA-bound TRAP, Glu-50 exhibits a hydrogen bond interaction with $-$OH group of Tyr-62 at all eleven positions around the ring (Figure 4.18).

As has been anecdotally suggested in previous literature (Burmeister, 2000) (and will be quantitatively demonstrated in Chapter 6), Tyr residues undergoing such Glu/Asp interactions are themselves more susceptible to radiation-induced disordering. The data in Figure 4.19 indicate that this is true for Tyr-62 in TRAP, since Tyr-62 in non-bound TRAP interacts with directly with the CO_2 group of Glu-50 (Figure 4.18a), whereas in RNA-bound TRAP no such interaction occurs (Figure 4.18b). In turn, it appears that the interaction with Tyr-62 may also increase the radiation-sensitivity of the Glu-50 side-group.

For both non-bound and RNA-bound TRAP, the Glu-71 and Glu-73 residues occupy TRAP-TRAP stacked binding interfaces, where the dominant interactions between the rings appear to be to mediated by ordered solvent molecules, and no striking differences in interactions could be distinguished. As illustrated in Figure 4.18, the $2mF_{obs} - DF_{calc}$ map electron density was visibly better defined for both Glu-71 and Glu-73 around the bound TRAP ring (chains L-V). This observation was reinforced by the real-space Z-score metric computed by *EDSTATS* (as part of

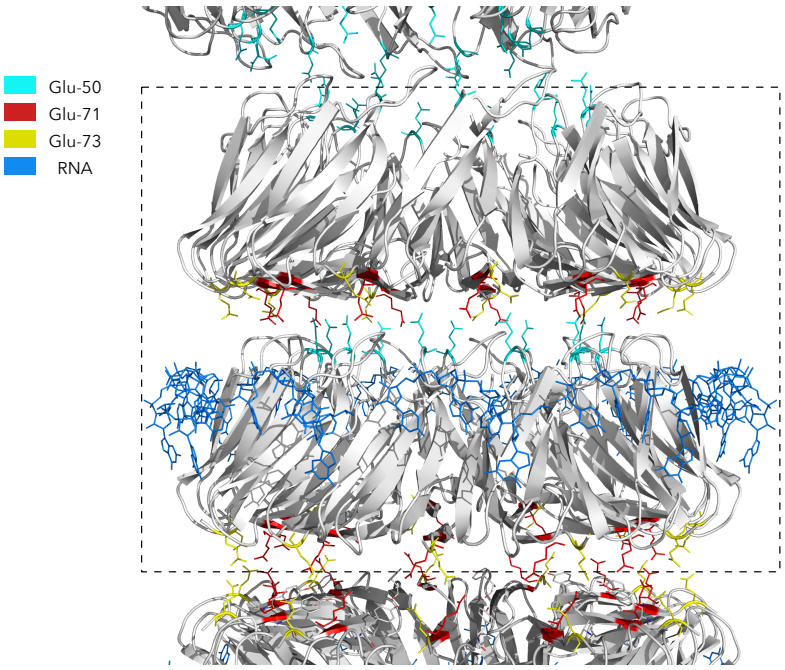
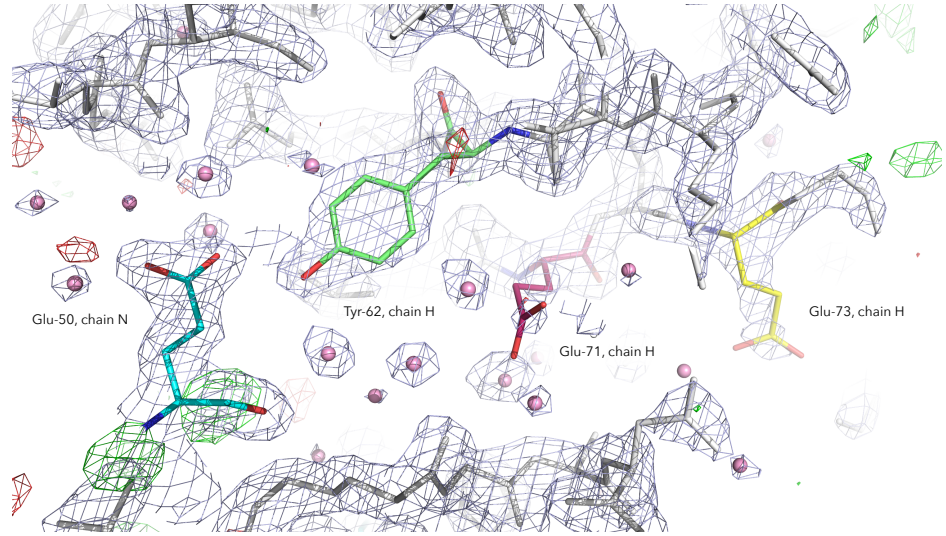


Figure 4.17: The crystal packing arrangement for the TRAP-RNA complex in space group C2. The TRAP-RNA asymmetric unit is highlighted by the dashed black box. Acidic residues exhibiting crystal contacts between TRAP rings, either within the same or to adjacent asymmetric units are highlighted.

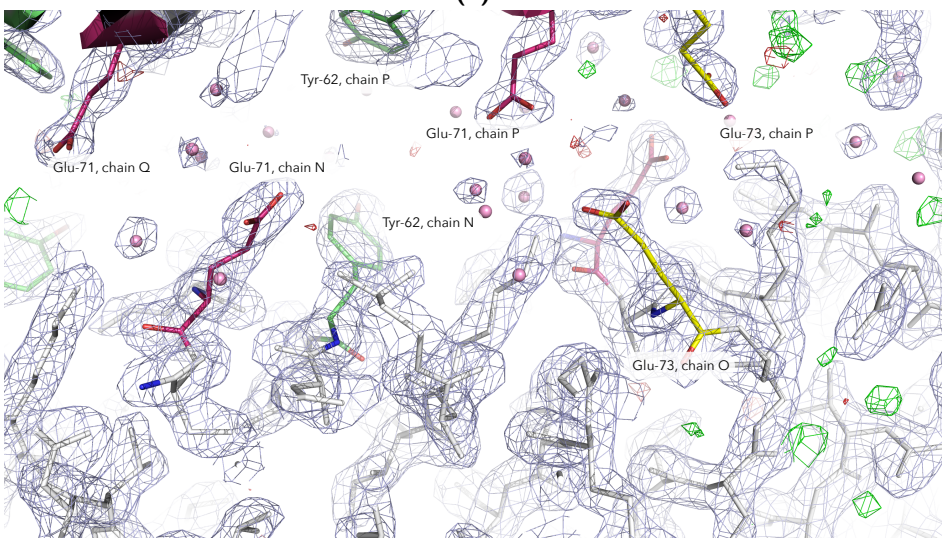
the *CCP4* software suite) as a measure of the quality of electron density at each residue/nucleotide in TRAP, as shown in Figure 4.20a. Furthermore, the isotropic atomic B-factor values attained by carboxyl groups of Glu-71 and Glu-73 in RNA-bound TRAP were also significantly lower than their non-bound counterparts (Figure 4.20b).

It is thus suggested that Glu-71 and Glu-73 on non-bound TRAP were initially more disordered, and consequently, a further change in electron density due to further disordering or decarboxylation at higher doses would be less evident in the generated $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps from which the D_{loss} metric is derived. For each protein monomer comprising TRAP, Glu-71 and Glu-73 are amongst the last few residues of the amino acid primary sequence that could be reliably built into $2\text{m}F_{\text{obs}} - \text{D}F_{\text{calc}}$ map electron density. It is therefore justifiable that these residues may then be partially disordered, however it is still unclear why the $2\text{m}F_{\text{obs}} - \text{D}F_{\text{calc}}$ map electron density for these residues is much better defined in RNA-bound TRAP.

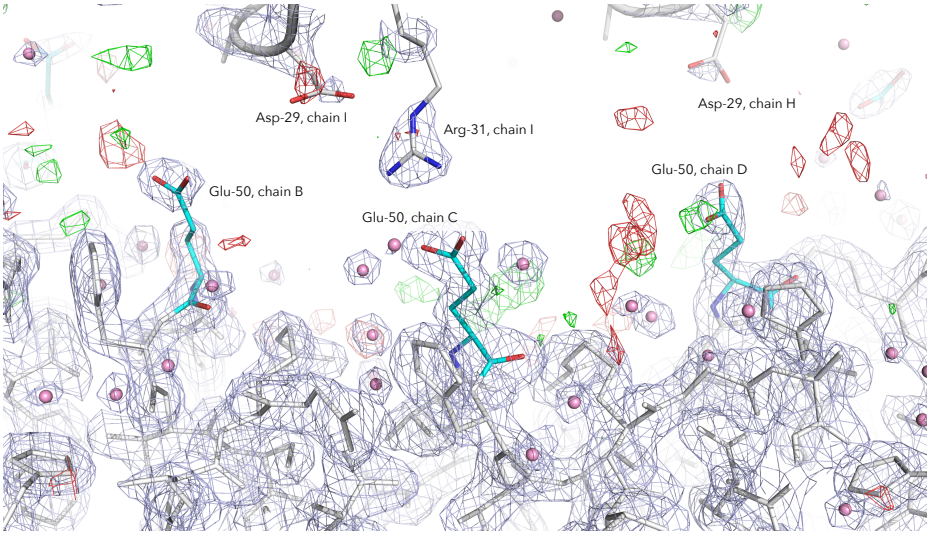
Ultimately, whereas the exact nature behind the differing damage dynamics displayed in Figure 4.12h-j by the crystal-contact forming residues Glu-50, Glu-71 and Glu-73 is unclear, there appears to be insufficient evidence to suggest that the RNA component itself plays any role in this heterogeneity. In RNA-bound TRAP, each Glu-50 residue is positioned an average of > 20 Å from



(a)

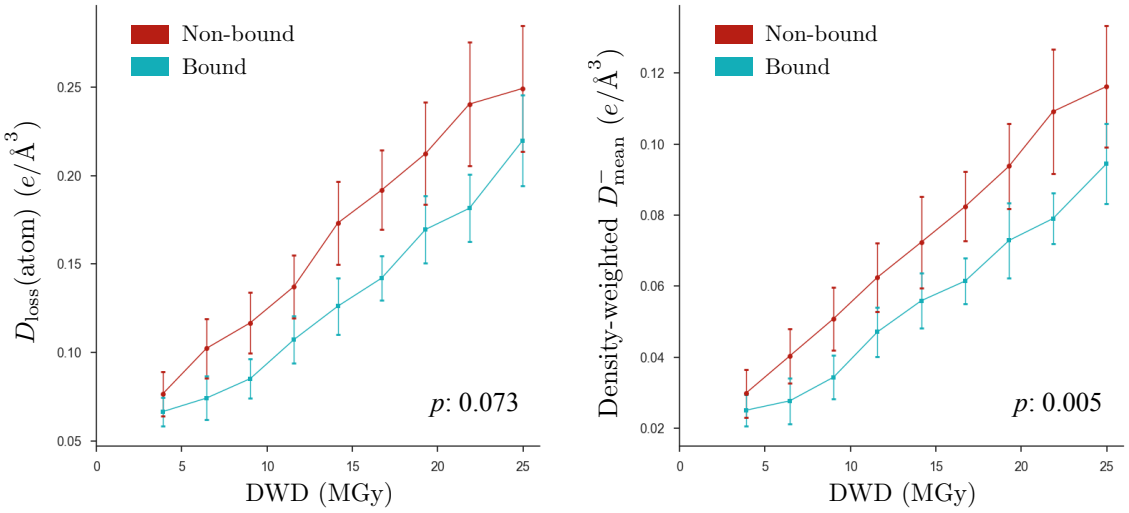


(b)



(c)

Figure 4.18: Renderings of crystal contacts exhibited by Glu-50, Glu-71 and Glu-73 residues at each ring-ring boundary. All $2mF_{obs} - DF_{calc}$ maps (in light blue) are contoured at $\pm 1.5\sigma$ and all $mF_{obs} - DF_{calc}$ maps (in green/red) are contoured at $\pm 3\sigma$.



(a) **(b)**

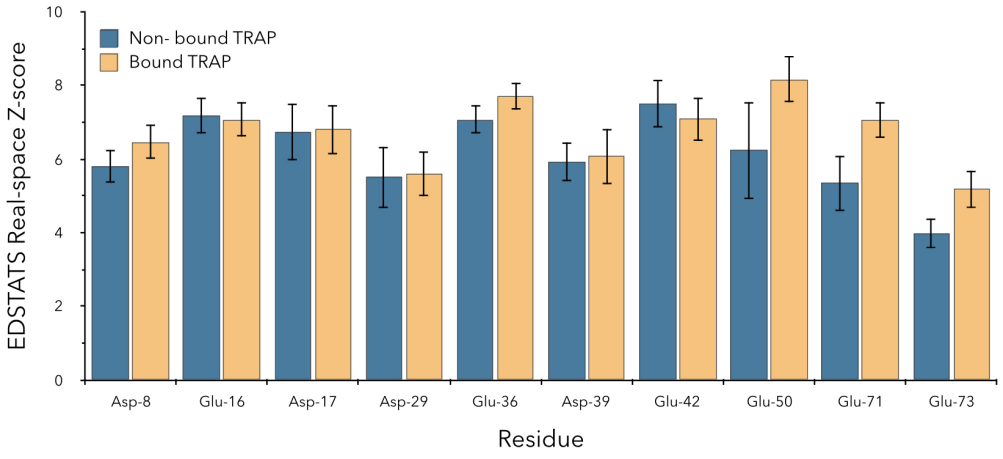
Figure 4.19: The (a) D_{loss} and (b) density-weighted D^{-}_{mean} metrics against dose for Tyr-62 O_{η} atoms in TRAP. Residues have been split into bound and non-bound groupings, with such that each point representing the mean calculated over 11 equivalent instances of a Tyr-62 O_{η} atom around each TRAP ring. Error bars indicate 95 % confidence intervals.

the nearest RNA backbone, with Glu-71 and Glu-73 positioned even further. It is not suspected that TRAP undergoes a full structural reorientation upon RNA-binding, and thus it must be concluded that it is the differing crystal contact interactions, not RNA-binding, which contributes to the damage rates of Glu-50, Glu-71 and Glu-73.

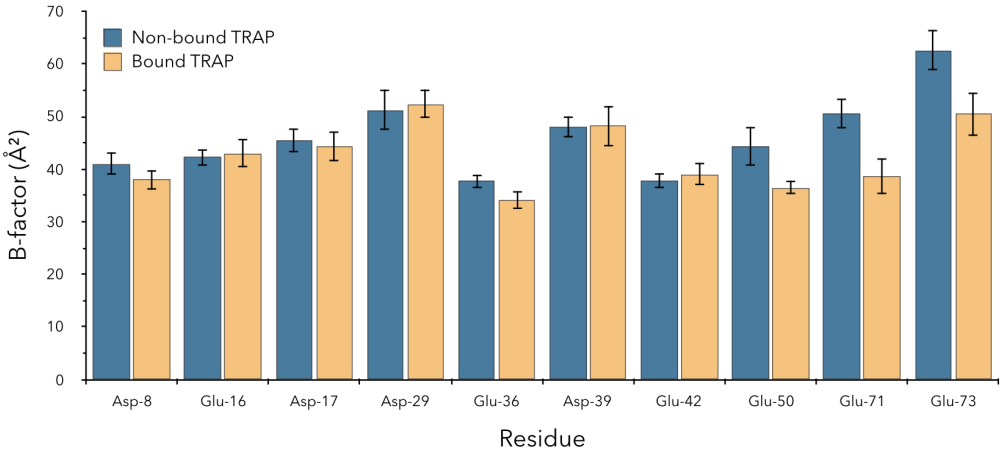
4.4 Discussion

In the current chapter, MX radiation-induced specific structural changes in the large TRAP-RNA assembly over a large dose range (1.3 – 25.0 MGy) have been analysed. Using the high-throughput quantitative approach permitted by the *RIDL* pipeline (Section 2.5), the bound RNA component was found to be substantially more radiation-resistant than the protein, even at the highest doses investigated (25.0 MGy). This observation was in strong concurrence with the investigation that had been reported in Chapter 3 for both the C.Esp1396I and C.Csp231I -DNA complexes.

Consistent with Chapter 3, only at high doses (above 20 MGy) was any distinguishable $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density present that could indicate radiation-induced electron density loss around the P, O3‘ and O5‘ atoms of the RNA backbone. No significant $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density



(a)



(b)

Figure 4.20: (a) Real-space Z-score metric values and (b) isotropic B-factor values attained by each Glu and Asp group around TRAP for the initial refined coordinate model corresponding to a dose of 1.31 MGy. The Z-score metric has been computed by *EDSTATS* as part of the CCP4 program suite, to assess the local quality of $2mF_{obs}-DF_{calc}$ map electron density for each refined residue in TRAP. Residues have been divided by residue number and into non-bound/bound groups. Each bar represents the average metric value calculated over all carboxyl atoms of a given residue type (e.g. all C_{δ}, $O_{\epsilon 1}$ and $O_{\epsilon 2}$ atoms for Glu-36 in RNA-bound TRAP). For each bar, whiskers represent the 95 % confidence interval over the eleven copies of each residue type.

was localised to RNA ribose and basic subunits, and RNA backbone disordering thus appeared to be the main radiation-induced effect in RNA. In contrast to the high prevalence of protein decarboxylation events across each TRAP ring, these RNA effects were comparably small, and there was insufficient evidence to suggest any bond rupture events were present at the RNA backbone. In contrast to the double helical structure of DNA in the C.Esp1396I and C.Csp231I systems, the 53 bp RNA bound to TRAP is non-rigid, and contains flexible 'spacer' nucleotides (U4 and U5) at regular intervals around TRAP that make no direct interactions to the TRAP ring. The absence of RNA radiation damage in the current system suggests that the relatively rigid helical structure of DNA in C.Esp1396I is not alone responsible for the radiation-insensitivity of dsDNA reported in Chapter 3.

Since RNA-binding in the current TRAP-$(GAGUU)_{10}$GAG complex is mediated predominantly through base-protein interactions, the biological integrity of the RNA complex in the crystal was therefore dictated by the rate at which protein decarboxylation occurred. The absence of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density at the RNA basic subunits suggested that the protein-base interactions were maintained even at the highest reported doses, and that the complex remained intact. Moreover, RNA interacting with TRAP was shown to offer protection against radiation-induced structural changes, with the most significant protection provided to amino acids that were most functionally important for RNA-binding. Both Glu-36 and Asp-39 bind directly to RNA, each through two hydrogen bonds to guanine bases (G3 and G1, respectively); however, compared with Asp-39, Glu-36 was strikingly less decarboxylated when bound to RNA. Previous mutagenesis and nucleoside analogue studies (Elliott *et al.*, 2001), have reported that the G1 nucleotide does not bind to TRAP as strongly as do A2 and G3, and plays minimal role in the high RNA-binding affinity of TRAP ($K_d \simeq 1.1 \pm 0.4$ nM). Consistent with those observations, in the current analysis, the Asp-39-G1 hydrogen bond lengths were significantly longer than for the Glu-36-G3 interaction around the bound TRAP ring. It has thus been suggested here that the differing strength of these bonds plays a role in reducing the rate of Glu-36 decarboxylation. By examining the change in solvent accessibility in the presence of RNA, it was deduced that dissociation of the $-CO_2$ group of Glu-36 was sterically hindered by the presence of the G3 base. For Asp-39, G1 did not block the escape route for dissociating CO_2, resulting in a less significant reduction in decarboxylation.

Aqueous RT studies on protein bound to DNA (e.g. DNA glycosylase Fpg and its abasic DNA

target site, as studied by Gillard *et al.* (2004)) has reported a lower radiation-sensitivity for protein in its DNA-bound form, although at much lower doses ($\sim$ 1 kGy). In aqueous studies, it has been suggested that DNA physically shields key protein-DNA interactions sites from hydroxyl mediated-damage, ultimately extending the total life-dose of the complex under irradiation. In the current MX study at 100 K, due to the immobilisation of $OH^{\bullet}$ radicals, the main damaging species are believed to be migrating LEEs and holes produced directly in the protein-RNA components or in closely associated solvent. The findings reported in the current chapter suggest that biologically relevant nucleoprotein complexes also exhibit prolonged life-doses under the effect of LEE-induced structural changes, involving direct physical protection of key RNA-binding residues. It is anticipated that if such observations persisted to the cellular environment, such reduced radiation-sensitivity would ensure that interacting protein remained bound long enough to the RNA to complete its function, even whilst exposed to ionising radiation. Furthermore, in RNA non-bound TRAP, the acidic residues in the unoccupied RNA-binding interfaces (Asp-39, Glu-36, Glu-42) were notably amongst the most susceptible residues in the asymmetric unit. When exposed to X-rays, these residues are preferentially damaged by X-rays and subsequently reduce the affinity with which TRAP binds to RNA. In the cellular environment, this mechanism could reduce the risk that radiation-damaged proteins might bind to RNA and function incorrectly.

Apart from these RNA-binding interfaces, the presence of RNA appeared to enhance the rate of decarboxylation for residues Glu-50, Glu-71 and Glu-73, all of which are involved in crystal contacts between TRAP rings. Due to the large distance between these residues and the RNA component (> 20 Å), it was not suspected that RNA binding had directly altered the radiation-sensitivity of these residues. Instead, the differing crystal contacts exhibited by these residues (for Glu-50), and also initial disordering of these residues for non-bound TRAP in the lowest dose coordinate model (1.31 MGy), was believed to contribute to the differential damage between the two TRAP rings. Interestingly, since the electron density for Glu-50, Glu-71 and Glu-73 was significantly more ordered in the lowest dose coordinate model for RNA-bound TRAP (Figure 4.20), the overall integrity of the irradiated crystal (and thus diffractive power of the crystal) appears to be most dependent on the radiation-sensitivity of these residues in the bound TRAP ring.

CHAPTER 5

Characterisation of site-specific radiation damage in DNA

5.1 Introduction

It was previously demonstrated in Chapter 3 and 4 that DNA and RNA, respectively, are both radiation-insensitive when bound directly to protein components in crystals irradiated by X-rays at 100 K. Whereas characteristic protein site-specific damage was clearly detectable in these investigated systems at doses of the order of $1-5$ MGy, distinguishable electron density loss in RNA and DNA only became present at high dose values (typically exceeding 20 MGy), and restricted only at the phosphate backbone of nucleic acids. It was hypothesised that bound protein may offer a protective effect to nucleic acids by scavenging the major damaging species at 100 K, migrating electrons and holes, away from the DNA/RNA. Indeed, in the cellular environment, it would be advantageous for expendable protein to protect DNA, due the high price that radiation-induced mutations and scissions in DNA have on the eventual fate of the cell. Studies of interacting protein-nucleic acids in solution at RT have observed this effect, with bound protein protecting DNA from direct attack by diffusing hydroxyl radicals (Štísová *et al.*, 2006); in turn protein can also be protected by DNA to preserve the functionally active form of a protein-DNA complex (Gillard *et al.*, 2004).

However in the case of MX experiments at 100 K, without a damage series on nucleic acid in isolation without protein, it could not be determined whether damage to DNA was dimished in the presence of protein, or whether DNA is intrisically hardier to the effects of ionising electrons and holes. To elucidate this, the current chapter presents an investigation on site-specific radiation damage effects for three independent DNA-containing crystals, each of variable base length, crystal symmetry and crystallisation buffer conditions. Whereas the conclusions derived from the previous studies on nucleoprotein complex crystals were often limited by the diffracting resolution ($\geq$ 2 Å), the higher resolution of the DNA crystals presented in this chapter ($\sim$ 1.4 Å) should permit a more precise insight into the exact covalent bonds undergoing electron density loss at the phosphate backbone.

Three DNA crystals have been chosen for the current analysis:

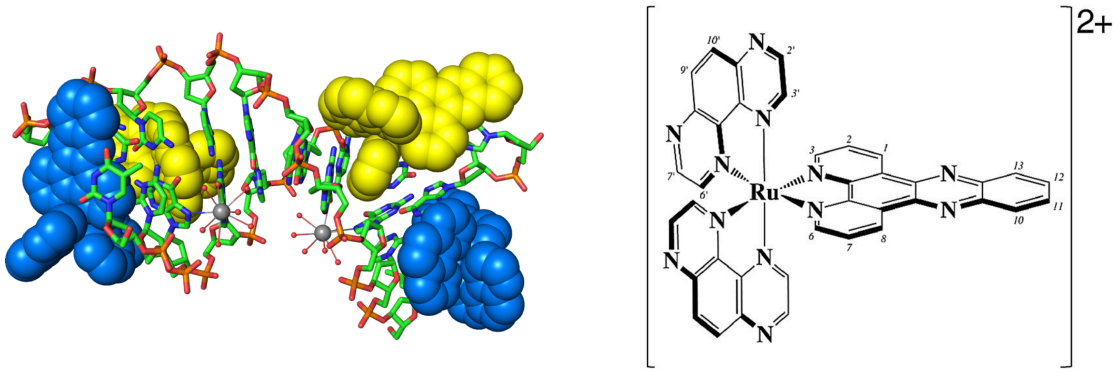
5.1.0.1 The TCGGCGCCGA-Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ complex

A TCGGCGCCGA double-stranded base sequence bound to an intercalating ruthenium polypyridyl complex, Δ-[Ru(1,4,5,8-tetraazaphenanthrene)$_2$(dipyridophenazine)]$^{2+}$ (Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$), is the first reported instance of kinking to duplex DNA caused by semiintercalation of a small molecule (Hall *et al.*, 2011). Previous structural studies have focussed on exploiting the Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ ligand for photosensitised DNA-directed therapies, in which targeted photo-induced DNA damage can be induced by one-electron photo-oxidation of guanine near to the ligand intercalation site (Hall *et al.*, 2015), resulting in the eventual formation of irreversible adducts between the TAP ligand in Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ and the amino group of guanine (Marcélis *et al.*, 2013).

For the purposes of the current investigation, the Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$-TCGGCGCCGA complex (PDB accession code: *3qrn*) was chosen since well diffracting ($\sim$ 1.1 Å) crystals were readily available and provided a suitable candidate as a ligand-bound DNA complex (Figure 5.1). The space group is P4$_3$2$_1$2 and each asymmetric unit consists of one single strand of DNA decamer, bound to one Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ complex and one ordered Ba^{2+} cation with direct coordination to the G3 and G4 bases. It will be determined whether Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ intercalation, or proximity of either the Ru or Ba atoms, plays a role in the susceptibility of each nucleotide to phosphate backbone damage.

5.1.0.2 The CGCAAATTTGCG duplex

The CGCAAATTTGCG DNA duplex (PDB: *1d65*) was chosen for the current investigation as a representative B-form DNA sequence in the absence of any bound ligands (Figure 5.2a). The space group is P2$_1$2$_1$2$_1$ and each asymmetric unit consists of one double stranded DNA 12-mer. The original structural study on CGCAAATTTGCG aimed to evaluate intrinsic bending in the AT-rich intermediate region of this dodecanucleotide duplex, and revealed the presence of a propeller twist between the central A−T base pair of up to -20° to permit such bending (Edwards *et al.*, 1992). For the C.Esp1396I-DNA complex described in Chapter 3, it was suggested that minor group compression (induced by C-protein binding) at the TATA recognition sequences of the 35 bp DNA duplex might be responsible for preferential damage to these sites. The current study



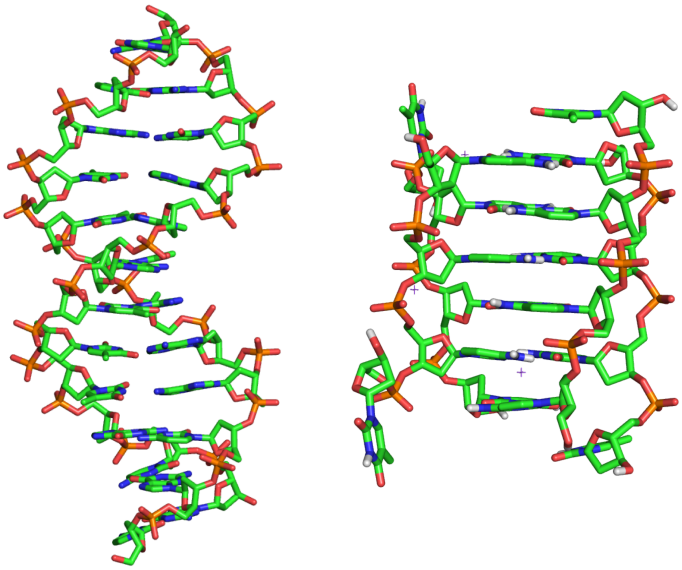
(a) (b)

Figure 5.1: (a) A schematic illustrating the semiintercalating and intercalating interactions of the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand with the double-stranded decamer DNA sequence (four copies shown in blue and yellow spheres). The asymmetric unit comprises of one single stranded TCGGCGCCGA sequence and one Δ-$[Ru(TAP)_2(dppz)]^{2+}$, however symmetrically related copies are shown for completeness. The TAP ligand of Δ-$[Ru(TAP)_2(dppz)]^{2+}$ semiintercalates at the minor groove between G3 and G4 (in yellow), with intercalation of the dppz component also occurring from the minor groove at the six membered purine rings of G9 and A10 (in blue). Bound Ba^{2+} cations are shown as grey spheres. Coordinates correspond to the *3qrn* model deposited in the PDB (Hall *et al.*, 2011). (b) The chemical composition of the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand.

aims to determine whether such a relationship persists in an independent DNA duplex.

5.1.0.3 Intercalated four-stranded CCCT

This DNA system consists of two parallel-stranded CCCT duplexes, each of which is held compactly together by cytosine-protonated cytosine ($C-CH^+$) base pairing (PDB: *191d* (Kang *et al.*, 1994)). In contrast to the two previous DNA systems, this system presents an example in which non Watson-Crick base pairs are present. The two CCCT duplexes intercalate to form a four-stranded molecule under acidic conditions, with each having opposite strand orientation (Figure 5.2b). The space group is $C222_1$ and each asymmetric unit consists of one CCCT tetrameric assembly. The geometry of the molecule is strikingly different from the B-form double helix in the previous two crystals, with the four-stranded molecule adopting a flat strip-like appearance, called an I-motif. It has been suggested that such cytosine-rich I-motifs can act as pH-controlled switches in regulatory regions of the human genome such as centromeres, telomeres and oncogene promoter regions, since at neutral or basic pH values, deprotonation of the cytosine involved in the $C-CH^+$ base pairing causes the structure to unfold into single oligonucleotide strands (Assi *et al.*, 2016; Day *et al.*, 2014). Due to its pH-dependence on folding, I-motifs have also attracted much attention in the field of nanotechnology; of particular interest, I-motifs have been exploited



(a) (b)

Figure 5.2: A schematic illustrating (a) the CGCAAATTTGCG DNA duplex and (b) the intercalated four-stranded CCCT DNA I-motif structure.

as photo-sensitive conformational switches in which a light-induced hydroxide ion emitter can be used to deprotonate the $C-CH^+$ base pairings (Liu *et al.*, 2007). For the current MX experiment, it is believed that radical species such as hydroxide ions should be immobile at 100 K, however it is unknown whether the cytosine protonation state could nonetheless be affected.

5.2 Materials and methods

5.2.1 Crystallisation

All oligonucleotide crystallisation of the three chosen DNA systems was performed by Dr J. Hall, University of Reading, UK. Crystallisation plates containing pre-formed crystals were then supplied to the current author. Since it is not the author's own work, only a brief summary has been presented in the current section, containing specific information on crystallisation buffer composition for accurate estimation of the X-ray absorption cross section when calculating per-dataset doses in Section 5.2.4 below.

5.2.1.1 The TCGGCGCCGA-Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ complex

In a similar protocol to that described in Hall *et al.* (2011), crystals were grown by sitting drop vapour diffusion, with the drop consisting of 1 µl 2 mM single stranded oligonucleotide and 1 µl 4 mM Δ-[Ru(TAP)$_2$(dppz)]·Cl, mixed with 6 µl of a solution containing 12 mM spermine tetra-HCl, 40 mM Na-cacodylate pH 7, 80 mM KCl, 20 mM $BaCl_2$ and 10 % (v/v) 2-methyl-2,4-pentane diol (MPD). The drop was equilibrated against a reservoir of 500 µl 35 % (v/v) MPD at 18 °C. Orange-brown roughly cubic crystals grew after a few days to a maximum diameter of 200 µm.

5.2.1.2 The CGCAAATTTGCG duplex

Crystals were grown by sitting drop vapour diffusion by mixing a 1:1 solution of 400 µm single stranded oligonucleotide and a buffer containing 80 mM NaCl, 12 mM KCl, 20 mM $MgCl_2$, 40 mM Na-cacodylate pH 7, 35 % (v/v) MPD and 12 mM spermine tetra-HCl. Drops were equilibrated against a reservoir of the buffer solution at 18 °C. A mixture of long spiky crystals (< 500 µm length) and shorter rod crystals ($50 - 150$ µm diameter) grew in each drop, however only the short rod variants were deemed suitable to be fished.

5.2.1.3 Intercalated four-stranded CCCT

Crystals were again grown by sitting drop vapour diffusion by mixing a 1:1 solution of 5 mM single stranded CCCT with a solution containing 40 mM LiCl, 20 mM $MgCl_2$, 40 mM sodium cacodylate trihydrate pH 5.5, 30 % (v/v) MPD and 20 mM $[Co(NH_3)_6]Cl_3$. Again, drops were equilibrated against a reservoir of the buffer solution at 18 °C. Large cubic rod-shaped crystals grew with average diameters between $100 - 400$ µm after several days.

5.2.2 X-ray data collection

All of the work detailed in the following sections was performed by the current author.

Crystals were mounted in cryoloops and flash-cooled in liquid nitrogen prior to X-ray data collection. For all three crystal type, no cryoprotectant was required (J. Hall, private communications),

due to the sufficiently high MPD content in each crystal, which is known to be a viable cryoprotectant (Garman and Schneider, 1997). X-ray diffraction data were collected in the same beam time session as for the DpCel7B protein crystal presented previously in Section 2.3.2.

Diffraction data were collected from a single crystal of each DNA type at 100 K on beamline I02 at the Diamond Light Source, using an incident wavelength of 0.980 Å (12.7 keV). A Pilatus 6M detector was positioned at a distance of 244.9 mm from the crystal throughout data collection, such that a maximum diffraction resolution of 1.4 Å could be achieved in the largest inscribed circle on the detector. The beam passed through a 200 µm diameter aperture and was slitted to $60 \times 120\,\mu m^2$ (vertical $\times$ horizontal), with an approximate Gaussian profile (vertical $\times$ horizontal full-width half-maximum (FWHM) of $17.7 \times 104.8\,\mu m^2$). As discussed in Section 2.3.2, the flux at the sample position was determined using a pre-calibrated 500 µm-thick silicon PIN photodiode (Owen *et al.*, 2009) to be $\sim 1.77 \times 10^{12}$ ph/s at 100 % beam transmission.

For each crystal type, the crystal was initially orientated with respect to the incident beam as shown in Figure 5.3. For each crystal, a 2000° continuous sweep of data was collected, consisting of 9999 frames (with an angular resolution of $\Delta\phi = 0.2°$) of 0.04 s exposure time each. For the CGCAAATTTGCG and CCCT tetramer crystals, the beam transmission was held at 25 % throughout. For the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal, the transmission was instead held at 50 %.

5.2.3 Diffraction data processing

For the CCCT tetramer and CGCAAATTTGCG duplex, the 2000° sweep of diffraction images was divided into eleven disjoint 180° datasets. Conversely, for the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal, images were divided into twenty-two disjoint 90° datasets. Despite the space group being $C222_1$ and $P2_12_12_1$ for the CCCT tetramer and CGCAAATTTGCG duplex respectively, it was found that 180° of data per dataset were required to ensure an overall completeness above 90 %.

Each wedge of images was independently integrated with *DIALS* (Waterman *et al.*, 2016), followed by scaling, outlier rejection and conversion to structure factor amplitudes using the data reduction pipeline in the CCP4i GUI (comprising of *POINTLESS*, *AIMLESS* and *CTRUNCATE* (Evans and

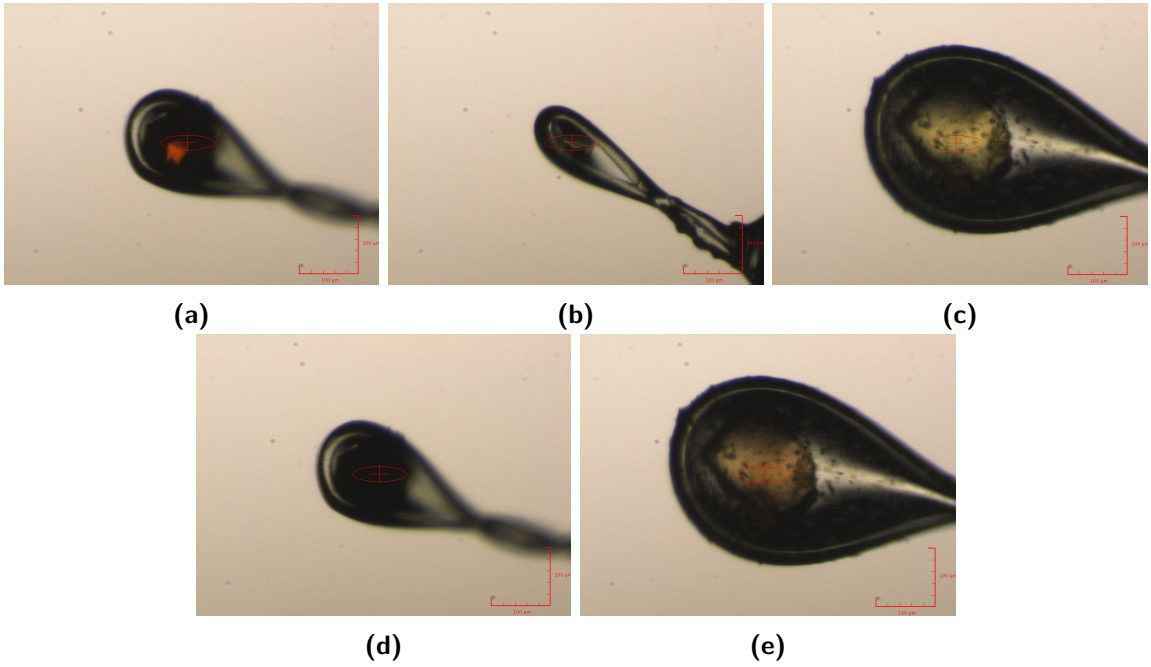


Figure 5.3: The initial orientation of the (a) TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$, (b) CGCAAATTTGCG duplex and (c) CCCT tetramer crystals mounted in nylon loops at Diamond Light Source beamline I02. In each case, the incident X-ray beam direction is directly into the page. Following 2000° of X-ray exposure, a crystal colour change was observed for both the (d) TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ and (e) CCCT tetramer crystals. Unfortunately, a post exposure image was not recorded for the CGCAAATTTGCG duplex DNA.

Murshudov, 2013)). For the first dataset for each DNA crystal, molecular replacement was performed using *PHASER* (McCoy *et al.*, 2007) to obtain an initial set of phases. Molecular replacement search model coordinates were retrieved from the PDB with accession codes *3qrn* (1.1 Å), *1d65* (2.2 Å) and *191d* (1.4 Å), each having 100 % base sequence identity with the TCGGCGCCGA, CGCAAATTTGCG and CCCT DNA sequences, respectively. For the case of the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal, it was opted to include both the ordered Ba^{2+} cation and the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand HETATM information when performing molecular replacement (omission of these led to failure of molecular replacement).

For each DNA crystal, the resulting initial model was then refined using *REFMAC* (Murshudov *et al.*, 2011), first using 10 cycles of rigid body refinement, followed by repeated rounds of 10 cycles of restrained and TLS refinement, coupled with manual inspection in *COOT* to update water positions (Emsley *et al.*, 2010). For the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ structure, anisotropic *B*-factor refinement was used for all atoms in the coordinate model, selected based on the Hamilton *R* ratio test (Merritt, 2012). Isotropic *B*-factor refinement was used for the other crystal types.

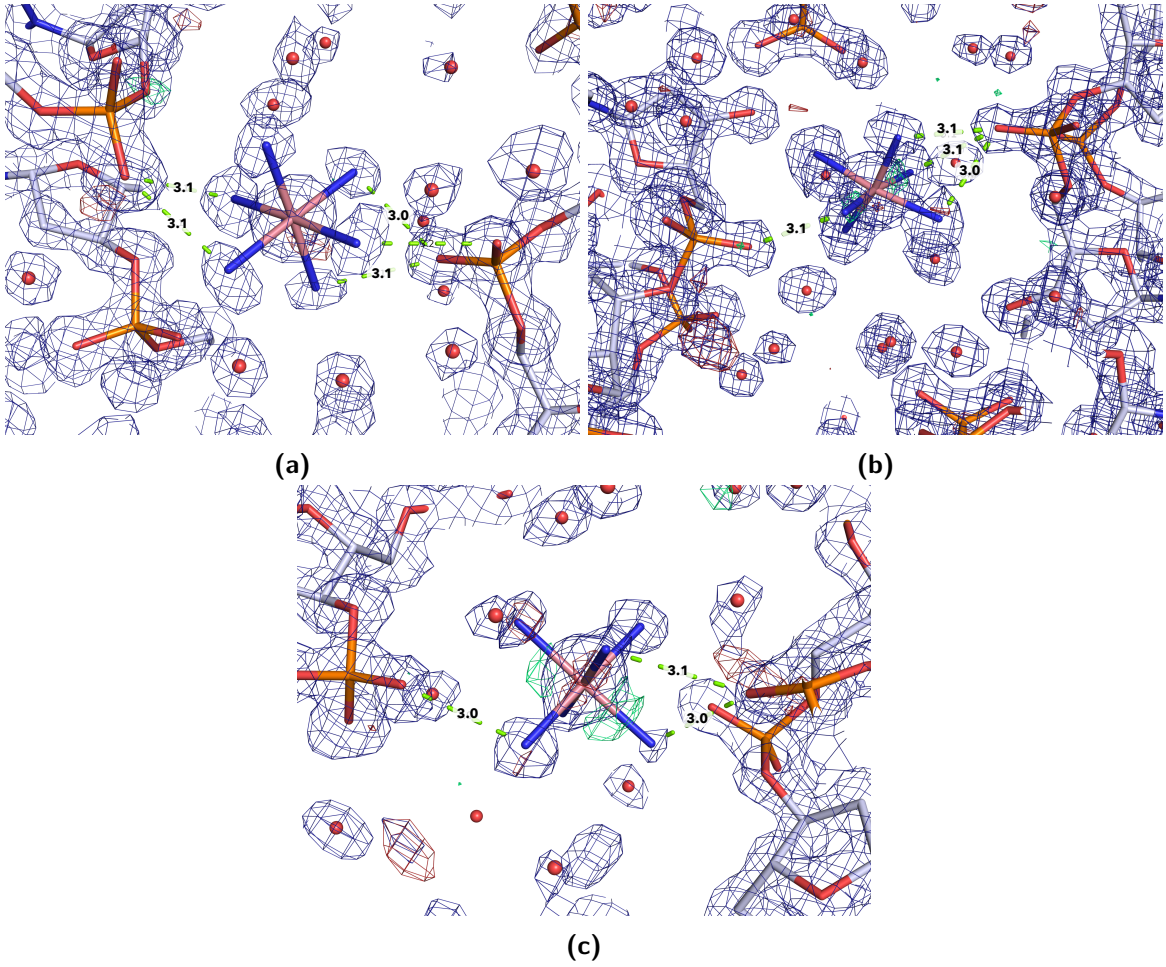


Figure 5.4: (a) $2mF_{obs}-DF_{calc}$ map density consistent with an ordered cobalt hexammine molecule was distinguished in the solvent channel for the CCCT tetrameric DNA crystal. (b) and (c) indicate the presence of two possible additional sites of cobalt hexammine that were more poorly defined in the $2mF_{obs}-DF_{calc}$ density map. Images have been generated using *PyMol*, with all $2mF_{obs}-DF_{calc}$ map density contoured at 1σ (in blue) and $mF_{obs}-DF_{calc}$ map density contoured at $\pm 3\sigma$ (in green/red).

For the case of the CCCT tetramer model, $2mF_{obs}-DF_{calc}$ map density consistent with an ordered cobalt hexammine molecule was distinguished in the solvent channel neighbouring the DNA (Figure 5.4a). As reported in Section 5.2.1, 20 mM $[Co(NH_3)_6]Cl_3$ had been included in the crystallisation buffer. Two other solvent regions of $2mF_{obs}-DF_{calc}$ map density were also interpreted as cobalt hexammine sites, however were less defined (Figure 5.4b-c). Other ionic species present in the buffer (Na^+ and Mg^{2+}) were first trialled at these sites, but the inclusion of Cobalt hexammine produced a better fit to density.

For the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand bound to TCGGCGCCGA, *ACEDRG* (Long *et al.*, 2017) was used to derive bond lengths and angles from the Crystallography Open Database (COD) and generate a ligand description that could be used by both *REFMAC* and *phenix.refine*. It

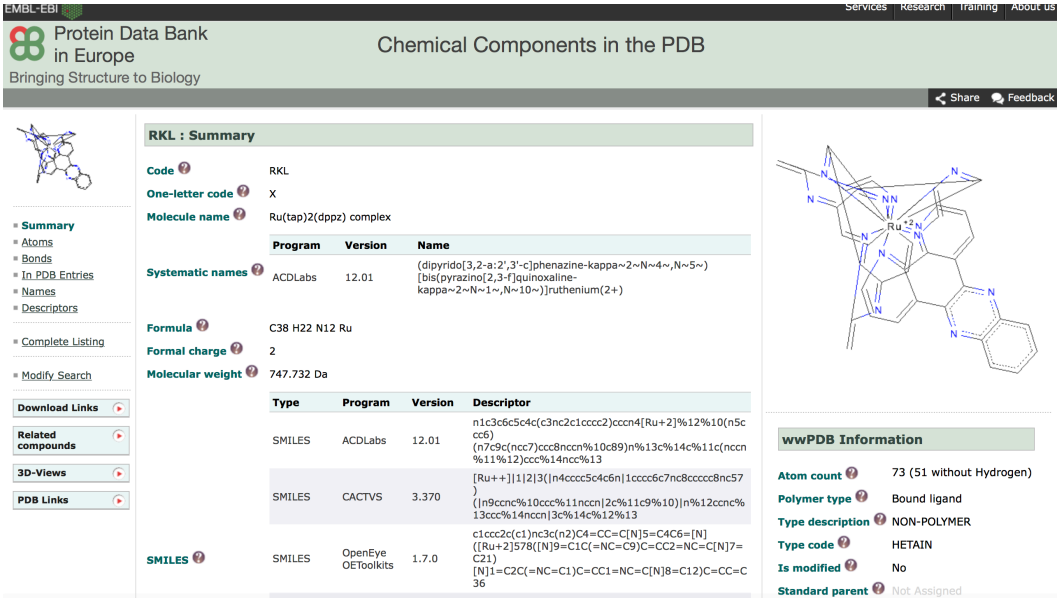


Figure 5.5: A screenshot of the Protein Data Bank in Europe (PDBe) ligand summary page for entry *rkl*, indicating the unusually distorted ligand geometry for the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand (*http://www.ebi.ac.uk/pdbe-srv/pdbechem/chemicalCompound/show/RKL*)

was noted that both *LIBCHECK* (Vagin *et al.*, 2004) and *JLigand* (Lebedev *et al.*, 2012) could not successfully generate sufficient ligand restraints and led to distorted aromatic ring geometry around the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand. Indeed a similar issue could also be observed at the corresponding ligand description page in the PDBe, as shown in Figure 5.5.

Final rounds of refinement were performed using *phenix.refine* in the *PHENIX* program suite (Adams *et al.*, 2010), with reference to the *MolProbity4* (Chen *et al.*, 2010) structure validation integrated into the *PHENIX* interface to manually assess the fit of individual nucleotides and ordered solvent atoms to electron density. The final coordinate model for the CCCT tetramer was refined to 2.0 Å, whereas the other two crystal types were refined to 1.4 Å.

For each DNA crystal type, coordinate models for each higher dose dataset were generated using rigid body refinement and atomic *B*-factor refinement in *phenix.refine*, with the refined coordinate model for the first dataset and the corresponding higher dose dataset structure factor amplitudes, $|\mathbf{F}_{obs,n}|$. This is an identical refinement strategy as has been previously performed for the C.Esp1396I-DNA, C.Csp231I-DNA and TRAP-RNA damage series described in Chapters 3 and 4, and is a standard protocol for refining multi-dataset protein damage series (Southworth-Davies *et al.*, 2007). See Section 2.3.3 for a detailed discussion on this protocol.

Table 5.1: Data processing and refinement statistics for the TCGGCGCCGA-Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ DNA data series. Values in parentheses are for the highest-resolution shells (fixed at 1.42 − 1.4 Å throughout). R_{free} has been calculated using a small (5 %) test set of randomly selected reflections, with the same test set maintained between datasets. Unit cell dimensions of $\alpha = \beta = \gamma = 90^\circ$ are conserved throughout. The DNA crystallised in space group P4$_3$2$_1$2. Table continues on next page.

Dataset	1	2	3	4	5	6	7	8	9	10	11
Dose (MGy)	4.29	8.90	16.8	21.4	29.5	34.0	41.9	46.6	54.6	59.2	67.0
Resolution range (Å)	42.34 - 1.40	39.58 - 1.40	42.41 - 1.40	39.57 - 1.40	42.34 - 1.40	39.51 - 1.40	42.40 - 1.40	39.56 - 1.40	42.34 - 1.40	39.50 - 1.40	42.38 - 1.40
Cell dimensions											
a (Å)	42.34	42.41	42.41	42.41	42.34	42.34	42.40	42.41	42.34	42.34	42.38
b (Å)	42.34	42.41	42.41	42.41	42.34	42.34	42.40	42.41	42.34	42.34	42.38
c (Å)	39.69	39.58	39.67	39.57	39.66	39.51	39.64	39.56	39.65	39.50	39.62
No. Observations	41406 (2048)	41801 (2041)	41670 (2039)	41701 (2122)	41984 (2080)	41920 (2041)	42071 (2016)	41992 (2056)	42262 (2243)	42087 (2035)	42252 (2047)
Unique reflections	7497 (369)	7523 (358)	7497 (366)	7519 (377)	7499 (370)	7495 (356)	7497 (359)	7527 (362)	7500 (395)	7502 (356)	7495 (364)
R_{merge}	0.05 (0.07)	0.04 (0.07)	0.05 (0.08)	0.04 (0.09)	0.05 (0.08)	0.04 (0.10)	0.05 (0.11)	0.04 (0.13)	0.04 (0.10)	0.04 (0.12)	0.05 (0.15)
R_{meas}	0.06 (0.08)	0.05 (0.09)	0.06 (0.10)	0.05 (0.11)	0.05 (0.09)	0.06 (0.13)	0.06 (0.13)	0.05 (0.15)	0.05 (0.12)	0.05 (0.15)	0.06 (0.18)
Completeness (%)	99.6 (99.9)	99.6 (99.8)	99.4 (100.0)	99.6 (100.0)	99.6 (100.0)	99.7 (99.8)	99.5 (100.0)	99.7 (100.0)	99.7 (100.0)	99.8 (99.8)	99.6 (100.0)
CC$_{1/2}$ (%)	0.993 (0.994)	0.994 (0.992)	0.993 (0.991)	0.994 (0.989)	0.993 (0.991)	0.995 (0.981)	0.994 (0.983)	0.994 (0.981)	0.995 (0.990)	0.996 (0.977)	0.995 (0.973)
$\langle I/\sigma(I)\rangle$	22.8 (13.3)	24.8 (15.0)	22.1 (12.2)	19.9 (8.8)	23.7 (12.3)	22.9 (10.4)	20.7 (9.6)	18.1 (5.7)	23.1 (9.3)	21.8 (7.4)	18.8 (5.8)
Multiplicity	5.5 (5.6)	5.6 (5.7)	5.6 (5.6)	5.5 (5.6)	5.6 (5.6)	5.6 (5.7)	5.6 (5.6)	5.6 (5.7)	5.6 (5.7)	5.6 (5.7)	5.6 (5.6)
Wilson *B*-factor (Å^2)	7.4	7.6	7.7	8.9	10.3	10.5	11.1	11.5	12.2	12.5	13.2
Resolution estimate from CC$_{1/2}$ (Å)	1.15	1.14	1.22	1.15	1.14	1.14	1.27	1.28	1.27	1.30	1.37
Resolution estimate from $\langle I/\sigma(I)\rangle$ (Å)	1.22	1.22	1.47	1.42	1.38	1.38	1.59	1.55	1.50	1.54	1.70
Refinement											
R_{work} (%)	0.108	0.116	0.120	0.123	0.131	0.127	0.143	0.149	0.154	0.150	0.168
R_{free} (%)	0.134	0.138	0.137	0.140	0.153	0.149	0.165	0.172	0.175	0.169	0.183
Mean isotropic *B*-factor (Å^2)	15.4	16.6	17.2	17.8	21.8	21.5	23.0	24.5	25.9	25.6	26.1

Table 5.1: (cont.) Data processing and refinement statistics for the TCGGCGCCGA-Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ DNA data series. Values in parentheses are for the highest-resolution shells (fixed at 1.42 − 1.4 Å throughout). R_{free} has been calculated using a small (5 %) test set of randomly selected reflections, with the same test set maintained between datasets. Unit cell dimensions of $\alpha = \beta = \gamma = 90^\circ$ are conserved throughout. The DNA crystallised in space group P4$_3$2$_1$2.

Dataset	12	13	14	15	16	17	18	19	20	21	22
Dose (MGy)	71.7	79.8	84.3	92.1	96.8	104.9	109.4	117.2	122.0	130.1	134.6
Resolution range (Å)	39.56 - 1.40	42.35 - 1.40	39.52 - 1.40	42.36 - 1.40	39.57 - 1.40	42.34 - 1.40	39.56 - 1.40	42.35 - 1.40	39.58 - 1.40	42.33 - 1.40	39.56 - 1.40
Cell dimensions											
a (Å)	42.41	42.35	42.35	42.36	42.41	42.34	42.38	42.35	42.40	42.33	42.37
b (Å)	42.41	42.35	42.35	42.36	42.41	42.34	42.38	42.35	42.40	42.33	42.37
c (Å)	39.56	39.66	39.52	39.60	39.57	39.67	39.56	39.60	39.58	39.67	39.56
No. Observations	42094 (2177)	42391 (2075)	42201 (2031)	42335 (2075)	42144 (2038)	42482 (2082)	42365 (2069)	42392 (2068)	42151 (2039)	42446 (2187)	42424 (2073)
Unique reflections	7535 (388)	7500 (366)	7513 (356)	7489 (367)	7545 (362)	7501 (367)	7530 (363)	7489 (364)	7549 (362)	7501 (384)	7528 (363)
R_{merge}	0.04 (0.19)	0.04 (0.16)	0.04 (0.20)	0.04 (0.25)	0.04 (0.38)	0.04 (0.31)	0.04 (0.38)	0.04 (0.50)	0.04 (0.74)	0.04 (0.59)	0.04 (0.82)
R_{meas}	0.05 (0.23)	0.04 (0.20)	0.05 (0.24)	0.05 (0.30)	0.05 (0.46)	0.04 (0.37)	0.04 (0.45)	0.05 (0.60)	0.05 (0.90)	0.04 (0.71)	0.04 (0.99)
Completeness (%)	99.8 (100.0)	99.6 (100.0)	99.8 (99.8)	99.6 (100.0)	99.8 (100.0)	99.6 (100.0)	99.8 (99.8)	99.6 (100.0)	99.9 (100.0)	99.7 (100.0)	99.9 (99.8)
CC$_{1/2}$ (%)	0.996 (0.966)	0.998 (0.975)	0.995 (0.966)	0.992 (0.927)	0.995 (0.911)	0.997 (0.937)	0.997 (0.906)	0.996 (0.868)	0.993 (0.739)	0.998 (0.786)	0.997 (0.722)
$\langle I/\sigma(I)\rangle$	17.1 (4.2)	22.0 (6.2)	21.1 (4.9)	17.9 (3.6)	15.5 (2.6)	19.4 (3.3)	18.6 (3.1)	16.8 (2.2)	14.9 (1.7)	17.7 (2.0)	16.6 (1.7)
Multiplicity	5.6 (5.6)	5.7 (5.7)	5.6 (5.7)	5.7 (5.7)	5.6 (5.6)	5.7 (5.7)	5.6 (5.7)	5.7 (5.7)	5.6 (5.6)	5.7 (5.7)	5.6 (5.7)
Wilson *B*-factor (Å^2)	13.5	13.9	14.5	15.6	15.9	16.2	17.6	19.3	19.0	19.5	21.2
Resolution estimate from CC$_{1/2}$ (Å)	1.39	1.39	1.42	1.46	1.51	1.50	1.53	1.55	1.61	1.63	1.62
Resolution estimate from $\langle I/\sigma(I)\rangle$ (Å)	1.65	1.64	1.64	1.77	1.85	1.85	1.81	1.93	1.94	1.93	1.93
Refinement											
R_{work} (%)	0.174	0.176	0.173	0.185	0.194	0.193	0.193	0.204	0.210	0.209	0.212
R_{free} (%)	0.193	0.202	0.190	0.207	0.217	0.214	0.211	0.238	0.238	0.237	0.241
Mean isotropic *B*-factor (Å^2)	28.6	29.5	29.2	32.5	32.1	32.6	34.4	37.5	35.8	37.4	40.9

Table 5.2: Data processing and refinement statistics for the CGCAAATTTGCG duplex DNA data series. Values in parentheses are for the highest-resolution shells (fixed at 2.05 − 2.00 Å throughout). R_{free} has been calculated using a small (5 %) test set of randomly selected reflections, with the same test set maintained between datasets. Unit cell dimensions of $\alpha = \beta = \gamma = 90^{\circ}$ are conserved throughout. The DNA crystallised in space group $P2_12_12_1$.

Dataset	1	2	3	4	5	6	7	8	9	10	11
Dose (MGy)	3.1	9.2	15.3	21.4	27.4	33.5	39.6	45.7	51.7	57.8	63.9
Resolution range (Å)	40.63 - 2.00	65.59 - 2.00	40.56 - 2.00	65.58 - 2.00	40.66 - 2.00	65.61 - 2.00	65.65 - 2.00	40.59 - 2.00	40.70 - 2.00	65.91 - 2.00	34.71 - 2.00
Cell dimensions											
a (Å)	24.79	24.91	24.79	24.89	24.85	24.90	24.87	24.95	24.92	24.90	24.98
b (Å)	40.63	40.72	40.56	40.58	40.66	40.51	40.57	40.59	40.70	40.56	40.81
c (Å)	65.43	65.59	65.42	65.58	65.65	65.61	65.65	65.86	65.82	65.91	65.98
No. Observations	26393 (1826)	26913 (1853)	26368 (1857)	26909 (1824)	26659 (1807)	26917 (1870)	26750 (1831)	27093 (1846)	27049 (1893)	27046 (1851)	27211 (1847)
Unique reflections	4849 (352)	4896 (356)	4841 (356)	4874 (347)	4879 (345)	4870 (353)	4873 (348)	4904 (350)	4915 (358)	4900 (350)	4942 (351)
R_{merge}	0.10 (0.45)	0.08 (0.48)	0.12 (0.69)	0.10 (0.89)	0.12 (1.76)	0.11 (1.72)	0.13 (4.00)	0.12 (3.60)	0.17 (8.79)	0.12 (14.67)	0.20 (26.95)
R_{meas}	0.13 (0.55)	0.10 (0.59)	0.15 (0.84)	0.12 (1.09)	0.15 (2.14)	0.13 (2.09)	0.16 (4.86)	0.14 (4.38)	0.20 (10.67)	0.14 (17.89)	0.24 (32.79)
Completeness (%)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)
$CC_{1/2}$ (%)	0.988 (0.920)	0.990 (0.918)	0.985 (0.844)	0.981 (0.813)	0.991 (0.586)	0.990 (0.655)	0.990 (0.331)	0.990 (0.183)	0.980 (0.184)	0.987 (0.103)	0.983 (-0.026)
$\langle I/\sigma(I)\rangle$	8.4 (3.1)	9.1 (2.4)	7.2 (2.2)	7.5 (1.7)	6.6 (1.4)	6.9 (1.2)	6.2 (1.0)	5.8 (0.7)	3.7 (0.8)	5.4 (0.2)	3.8 (0.3)
Multiplicity	5.4 (5.2)	5.5 (5.2)	5.4 (5.2)	5.5 (5.3)	5.5 (5.2)	5.5 (5.3)	5.5 (5.3)	5.5 (5.3)	5.5 (5.3)	5.5 (5.3)	5.5 (5.3)
Wilson *B*-factor ($Å^2$)	46.8	56.9	39.2	57.7	36.8	28.5	56.8	62.0	58.2	66.0	60.3
Resolution estimate from $CC_{1/2}$ (Å)	2.16	2.11	2.33	2.33	2.45	2.56	2.53	2.61	2.64	2.70	2.76
Resolution estimate from $\langle I/\sigma(I)\rangle$ (Å)	2.62	2.63	2.61	2.63	2.62	2.63	2.79	2.80	2.80	2.95	3.00
Refinement											
R_{work} (%)	0.227	0.238	0.239	0.237	0.244	0.250	0.249	0.268	0.284	0.272	0.296
R_{free} (%)	0.244	0.260	0.275	0.257	0.273	0.264	0.269	0.284	0.294	0.287	0.288
Mean isotropic *B*-factor ($Å^2$)	42.2	48.0	32.9	53.7	47.3	56.2	52.8	53.9	50.9	63.9	52.7

Table 5.3: Data processing and refinement statistics for the CCCT tetramer DNA data series. Values in parentheses are for the highest-resolution shells (fixed at 1.42 – 1.40 Å throughout). R_{free} has been calculated using a small (5 %) test set of randomly selected reflections, with the same test set maintained between datasets. Unit cell dimensions of $\alpha = \beta = \gamma = 90^{\circ}$ are conserved throughout. The DNA crystallised in space group $C222_1$

Dataset	1	2	3	4	5	6	7	8	9	10	11
Dose (MGy)	2.3	6.8	11.2	15.6	20.1	24.5	29.0	33.4	37.9	42.3	46.7
Resolution range (Å)	49.61 - 1.40	24.85 - 1.40	24.86 - 1.40	24.87 - 1.40	24.85 - 1.40	24.86 - 1.40	24.85 - 1.40	23.74 - 1.40	24.84 - 1.40	23.75 - 1.40	24.84 - 1.40
Cell dimensions											
a (Å)	28.14	28.19	28.21	28.22	28.21	28.21	28.20	28.20	28.19	28.20	28.19
b (Å)	43.95	43.98	44.00	44.02	44.01	44.02	44.01	44.02	44.02	44.05	44.02
c (Å)	49.61	49.69	49.72	49.73	49.71	49.71	49.70	49.70	49.68	49.70	49.68
No. Observations	36622 (1859)	36568 (1833)	36701 (1868)	36915 (1799)	36894 (1791)	37065 (1892)	37018 (1786)	37130 (1786)	37132 (1771)	37247 (1851)	37188 (1768)
Unique reflections	6340 (307)	6350 (305)	6360 (313)	6361 (302)	6355 (301)	6360 (317)	6349 (301)	6353 (302)	6349 (300)	6362 (313)	6356 (300)
R_{merge}	0.06 (0.05)	0.05 (0.04)	0.04 (0.04)	0.04 (0.05)	0.04 (0.06)	0.04 (0.06)	0.05 (0.08)	0.05 (0.09)	0.06 (0.10)	0.05 (0.11)	0.06 (0.13)
R_{meas}	0.08 (0.06)	0.06 (0.05)	0.05 (0.05)	0.05 (0.06)	0.05 (0.07)	0.05 (0.08)	0.06 (0.10)	0.06 (0.11)	0.07 (0.12)	0.07 (0.13)	0.07 (0.15)
Completeness (%)	99.9 (100.0)	99.8 (100.0)	99.8 (99.7)	99.8 (100.0)	99.7 (100.0)	99.7 (100.0)	99.7 (100.0)	99.7 (100.0)	99.7 (100.0)	99.8 (100.0)	99.8 (100.0)
$CC_{1/2}$ (%)	0.987 (0.996)	0.988 (0.998)	0.987 (0.999)	0.990 (0.998)	0.988 (0.998)	0.989 (0.996)	0.987 (0.989)	0.986 (0.990)	0.982 (0.986)	0.987 (0.986)	0.990 (0.982)
$\langle I/\sigma(I) \rangle$	21.0 (19.3)	30.9 (27.2)	31.3 (26.1)	29.3 (23.0)	27.5 (20.6)	25.9 (18.2)	22.3 (14.8)	21.1 (12.8)	18.1 (11.3)	19.6 (11.2)	16.3 (9.4)
Multiplicity	5.8 (6.1)	5.8 (6.0)	5.8 (6.0)	5.8 (6.0)	5.8 (6.0)	5.8 (6.0)	5.8 (5.9)	5.8 (5.9)	5.8 (5.9)	5.9 (5.9)	5.9 (5.9)
Wilson *B*-factor ($Å^2$)	0.2	1.0	2.1	2.6	3.2	4.1	4.1	4.7	4.8	5.6	4.9
Refinement											
R_{work} (%)	0.157	0.160	0.162	0.161	0.163	0.163	0.168	0.168	0.172	0.173	0.175
R_{free} (%)	0.186	0.188	0.181	0.189	0.191	0.189	0.203	0.201	0.214	0.208	0.215
Mean isotropic *B*-factor ($Å^2$)	6.3	7.1	8.2	9.9	11.0	12.6	13.0	14.5	15.4	16.6	17.4

5.2.4 Dose calculation

For each crystal damage series, diffraction weighted dose (DWD) values (Zeldin *et al.*, 2013a) were calculated using *RADDOSE-3D* (Zeldin *et al.*, 2013b) for each dataset. DWD values have been reported for each crystal in Tables 5.1, 5.2 and 5.3.

Crystals were modelled as approximately cuboid shaped with dimensions of $100 \times 100 \times 100\,\mu m^3$, $40 \times 50 \times 40\,\mu m^3$ and $150 \times 100 \times 75\,\mu m^3$ for the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$, CG-CAAATTTGCG duplex and CCCT tetramer, respectively. These dimensions were based on static images of each crystal mounted at the beamline taken at 90° intervals, examples of which have been shown in Figure 5.3.

Information on the crystal composition (the unit cell dimensions, number of nucleotides per monomer, number of monomers per unit cell, and crystal solvent content) were manually specified for each DNA crystal, with reference to each coordinate model and space group for each crystal. A Gaussian-shaped beam was modelled with parameters as detailed in Section 5.2.2.

Account had to be taken of the contribution of heavy elements in the crystallisation buffer for the calculation of the crystal absorption and attenuation coefficients in *RADDOSE-3D*. These were specified in units of $mmol\,l^{-1}$, using the crystallisation conditions for each crystal presented previously in Section 5.2.1. However, in order to accurately estimate the dose for the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal, it was crucial to obtain an accurate value for the content of both Ru and Ba^{2+} cations, due to high X-ray cross section of these elements. It was determined that the addition of a single Ba atom per unit cell would result in a 23.5 % increase in DWD for a single 90° wedge of data for the current experimental set up.

Due to the large size of Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand involved in DNA intercalation (747.73 Da, spanning 14.1 Å), and its role in forming the crystal contact interactions in the TCGGCGCCGA crystal (see Figure 5.9 in Section 5.3 later), a ratio of one Ru atom per single stranded DNA has been assumed (Figure 5.1b). Only one ordered Ba^{2+} ion has been identified per single-stranded DNA monomer (Hall *et al.*, 2011), however this does not discount the possibility of non-ordered barium in the crystal solvent channels. To determine the stoichiometric ratio of barium to DNA phosphorus atoms, a micro-particle-induced X-ray emission (microPIXE) experiment was

performed on a single TCGGCGCCGA-Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ crystal, the full details of which are presented in Section 5.2.5. An average P : Ba atom ratio of 9 : 6.06 was determined, indicating the presence of ~ 6 Ba^{2+} cations per single stranded DNA.

5.2.5 MicroPIXE experiment

MicroPIXE has previously been used as a tool for determining the precise stoichiometric ratio between metal ions and known sulphur content (frequency of Met and Cys residues) of a number of proteins (Garman, 1999; Garman and Grime, 2005), and in particular has been employed for accurate measurement of the iron content of holoferritin necessary for the experimental determination of the MX $D_{1/2}$ dose limit for proteins at 100 K (Owen *et al.*, 2006). The technique offers a spatial resolution $1 - 10\,\mu m$ that is suitable for protein crystal samples and a quoted detection limit of $1 - 10\,\mu g\,g^{-1}$ (Garman and Grime, 2005). A major benefit of using microPIXE for metal detection involves the low error of $10 - 20\,\%$ due to the internal calibration to the known number of S or P atoms in a biological sample. For the Ba^{2+}-containing TCGGCGCCGA-Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ crystal, the technique was therefore ideal to measure the stoichiometric ratio of barium to phosphorus atoms of the DNA backbone.

MicroPIXE data collection was performed at the Ion Beam Centre, University of Surrey, UK, using a separate crystal to that irradiated in Section 5.2.2. This crystal had been collected from the same crystallisation tray as the irradiated crystal, under identical crystallisation conditions as specified in Section 5.2.1. Care was taken such that the selected crystal was also of similar morphology and dimensions as the X-ray irradiated crystal.

MicroPIXE involves inducing characteristic X-ray emission of the atoms in a sample by proton bombardment with a focused ($< 2\,\mu m$) 2.5 MeV proton beam. Detection of the characteristic X-ray emission spectra of each element in a sample allows the relative concentration of each element (in units of parts per million, ppm) to be determined (Figure 5.7a). In addition, simultaneous measurement of the energy spectrum of Rutherford Back Scattered (RBS) protons (Figure 5.7b), resulting from protons being elastically scattering back from the nucleus of atoms at near 180°, permits the determination of the sample thickness and bulk composition (Grime, 1996). This is important, since C, N, O and H atoms emit X-rays of too low energy to be recorded in the

PIXE spectrum, but account must be taken of them to correct for the differential absorption of emitted X-rays at different energies before exiting the sample. The RBS spectrum is fitted to determine these sample parameters, and this information is then incorporated into the PIXE spectrum fitting algorithm as part of the *GUPIX* software package (Campbell *et al.*, 2010). For the current analysis, fitting of both the RBS and PIXE spectra were handled by the *OMDAQ-3* software interface (*http://www.microbeams.co.uk*).

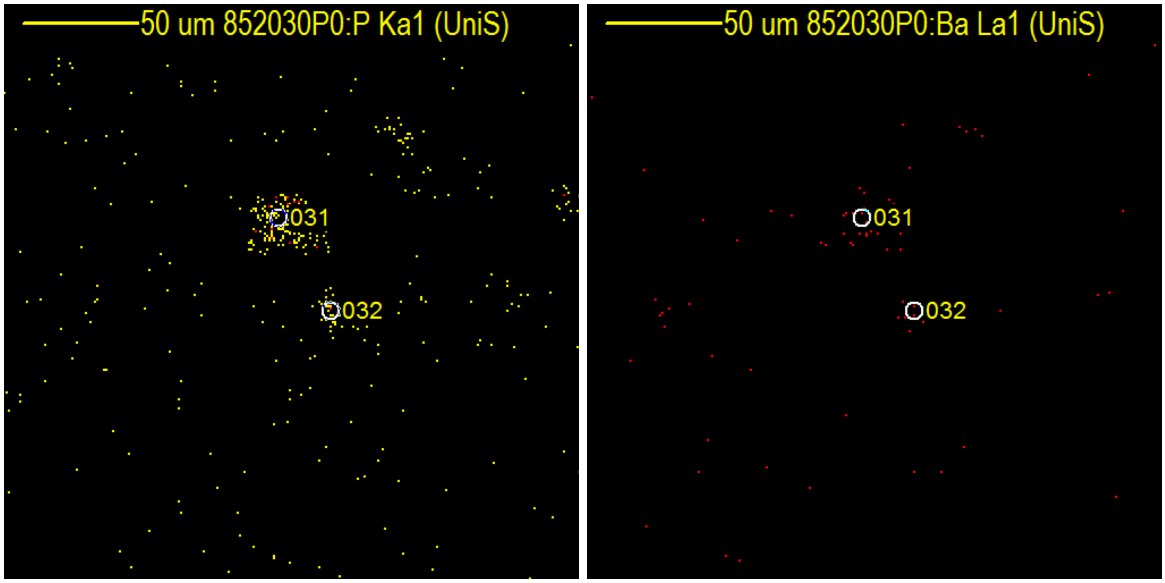
Internal calibration of elemental concentrations was performed using the known number of phosphorus atoms in the DNA backbone to extract an explicit ratio of phosphorus to barium atoms. This calibration could only be achieved due to the absence of phosphorus-containing compounds in the crystallisation buffer (Section 5.2.1). The absolute number of barium atoms per asymmetric unit was calculated using the formula:

$$\frac{[\mathrm{Ba}]}{\mathrm{Mr(Ba)}} \times \frac{\mathrm{Mr(P)}}{[\mathrm{P}]} \times 9 \tag{5.2.1}$$

where [X] and Mr(X) are the concentration (in ppm) and molecular weight (in Da) of element X. The factor of 9 corresponds to the number of phosphorus atoms per asymmetric unit. Note that despite TCGGCGCCGA being a 10 base sequence, the 5' thymine does not possess its phosphate group in the crystal.

Two PIXE spectra were collected at positions on a two fragmented shards from a single crystal (031 and 032 as shown in Figure 5.6), separated by $\sim 40\,\mu m$; Table 5.4 reports the concentrations (ppm) of each identified element. Using Equation 5.2.1, this corresponded to 6.58 and 5.53 Ba atoms for every 9 P atoms in the crystal asymmetric unit at the two sample positions. Consequently, an average value of 6.06 Ba atoms per single stranded DNA* has been used in the dose calculation in Section 5.2.4 above. Cl, K, As and Ru atoms were also detected above the microPIXE limit of detection (Garman and Grime, 2005) at an average of 6.61, 3.42, 2.26 and 2.53 atoms per 9 phosphorus atoms, all of which would be expected from the crystallisation buffer composition in Section 5.2.1. Trace iron (0.08 atoms per 9 P atoms) was also detectable. For Ru, only the L-shell characteristic X-ray line values from Table 5.4 have been used, since the K line fit had a high error. This was presumably due to the high energy of this K line ($\sim 22.0\,keV$)

*A non-integer number of Ba atoms is permitted per DNA strand in the dose calculation, since this is an overall average over all asymmetric units in the crystal. The number of Ba atoms for each individual asymmetric unit can only take integer values, but the average is not required to.



(a) **(b)**

Figure 5.6: Elemental areal concentration $250 \times 250\,\mu m^2$ microPIXE maps for (a) P (K edge) and (b) Ba (L edge). Maps were obtained by raster scanning the $\sim 2\,\mu$m-diameter proton beam along the shown x and y dimensions across the TCGGCGCCGA crystal. The positions of the two points of PIXE spectra collection (as reported in Table 5.4) are shown (labelled 031 and 032 in each map).

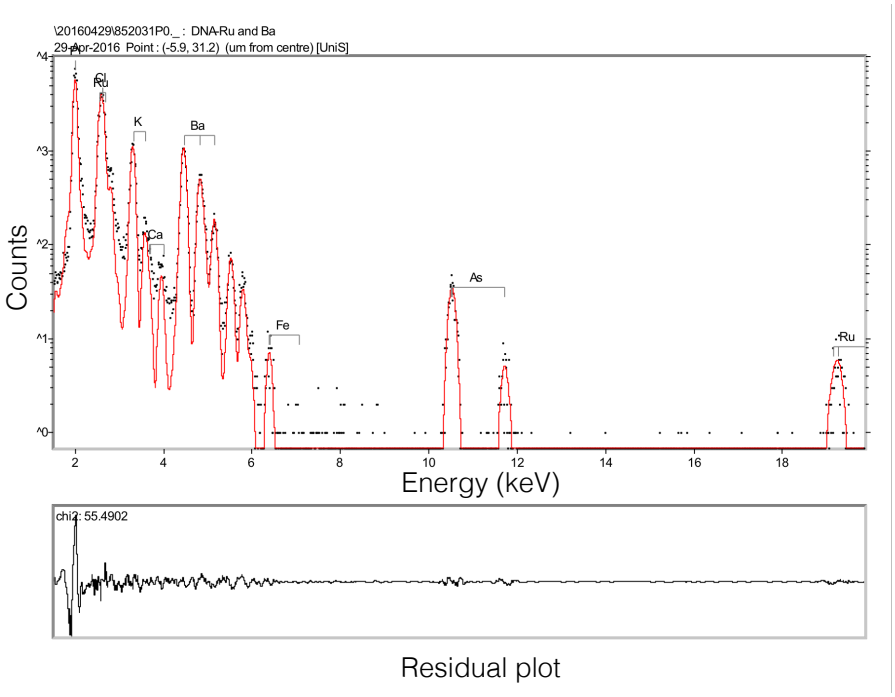
being outside the calibrated energy range for the detector's response.

The value of 2.53 Ru atoms per DNA strand differs from the theoretical value of 1 as discussed in Section 5.2.4 previously. The data in Figure 5.7b indicate that the PIXE peak corresponding to the K edge of Cl and L_{III} edge of Ru concided (at 2.8224 keV and 2.8379 keV); consequently the ratio of Ru to P atoms quoted above was not deemed to be accurate, and a further experiment in the absence of Cl buffer atoms would be required to improve the accuracy (G. Grime, private communications). The experimentally-determined stoichiometric ratio for Ru has thus been disregarded in the current work.

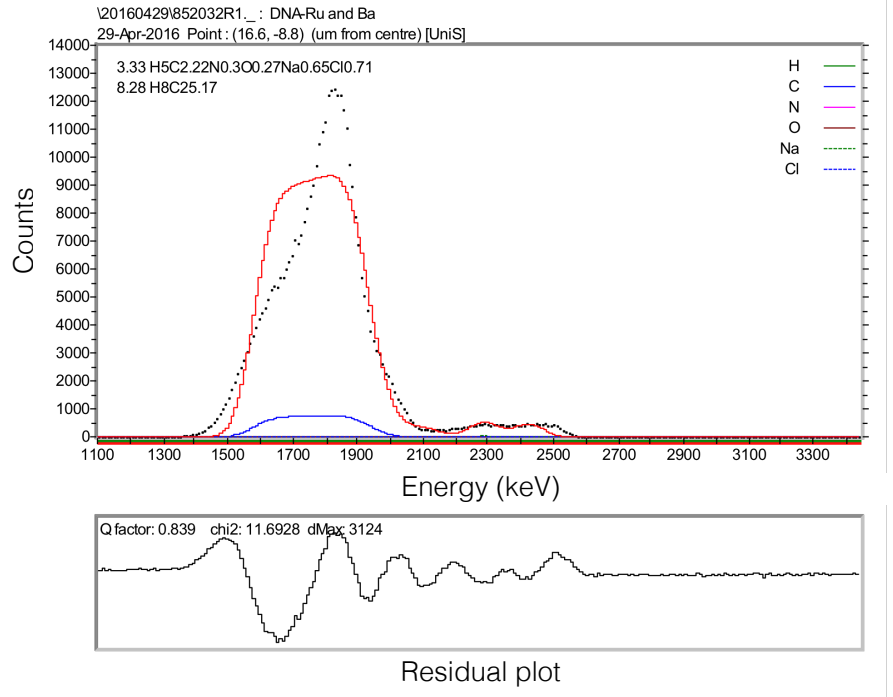
5.3 Radiation damage analysis

5.3.1 Global radiation damage effects with dose

This section aims to investigate the robustness of each DNA crystal type to global radiation damage effects. For each crystal, the decay in average intensity, I_n, relative to the initial diffraction dataset, I_1, could be monitored as a function of dose (Figure 5.8). Here, the average intensity



(a)



(b)

Figure 5.7: a) Fitted PIXE spectrum (in red) for one single point spectrum on the TCGGCGCCGA-Δ-$[Ru(TAP)_2(dppz)]^{2+}$ crystal allows the simultaneous measurement of all elements heavier than neon. A logarithmic y-axis scale has been used for clarity. (b) Corresponding RBS spectrum (black points) and fit (in red) for one point spectrum on the crystal. The depth of the main peak indicates the crystal sample thickness and the position of the high energy edge of peaks indicate the species of atoms present in the sample. H, C, N, O, Na and Cl atoms have been used in the RBS fitting to comprise the bulk of the crystal composition.

Table 5.4: Concentrations (parts per million, ppm) of each identified element derived from the PIXE spectra taken at two TCGGCGCCGA-Δ-[Ru(TAP)$_2$(dppz)]$^{2+}$ crystal sample positions collected over $\sim$ 40 µm apart. For each element, the percent error in fitting has been given, as reported directly from the *OMDAQ-3* software interface (*http://www.microbeams.co.uk*) used for data processing. Other elements common to biological samples (e.g. S and Br) were not identified above the microPIXE limit of detection, and have not been included in the table.

		Point 1		Point 2	
Element	X-ray line	Concentration (ppm)	Fit error (%)	Concentration (ppm)	Fit error (%)
P	K	10948.9	0.68	26151.0	0.57
Cl	K	8607.0	1.13	23427.9	0.77
K	K	4757.9	1.35	13713.1	0.89
Fe	K	184.7	17.42	395.4	14.04
As	K	6766.4	6.39	15633.0	4.27
Ru	K	25830.5	15.45	42102.3	13.98
Ru	L	11099.8	1.92	21394.0	1.65
Ba	L	35482.8	1.00	71291.8	0.79

at each dataset has been extracted from the *AIMLESS* log file, in a protocol previously defined in Section 3.2.3 using Equation 3.2.1. In Figure 5.8, care was taken to only sample the same region of reciprocal space for successive datasets, in order to monitor the intensity decay of a fixed set of reflections and constant crystal volume. To achieve this, for a given crystal, only images corresponding to a fixed 90° rotation range ($\phi = 0°$ to $\phi = 90°$) were included throughout. To motivate the need for this criterion, when considering the first 180° sweep for the CGCAAATTTGCG duplex crystal, the average intensities for images between $\phi = 90 - 180°$ was a factor of 1.45 higher than for the first $\phi = 0 - 90°$, owing to the exposure of fresh crystal volume.

Exponential decay functions of the form $y = A\exp(-k \cdot x)$ were fitted to each set of data points in Figure 5.8 using a least squares fitting procedure, leading to decay constant k values of 0.0141, 0.0194 and 0.133 MGy^{-1} for the TCGGCGCCGA duplex, CGCAAATTTGCG duplex and CCCT tetramer sequences, respectively. As indicated in Figure 5.8, the TCGGCGCCGA and CCCT DNA crystals are particularly radiation-insensitive, and exceed the experimentally-determined half-dose limit, $D_{1/2}$, which has been proposed as an upper bound for general protein crystals at 100 K (Owen *et al.*, 2006)[†]. The CGCAAATTTGCG crystal, although less radiation-hardy, still coincides

[†]As has been remarked previously in this thesis, for most protein crystals, the decay in mean diffraction intensity I_n/I_1 with respect to dose at 100 K is more rapid than in the upper bound reported by Owen *et al.* (2006) (e.g.

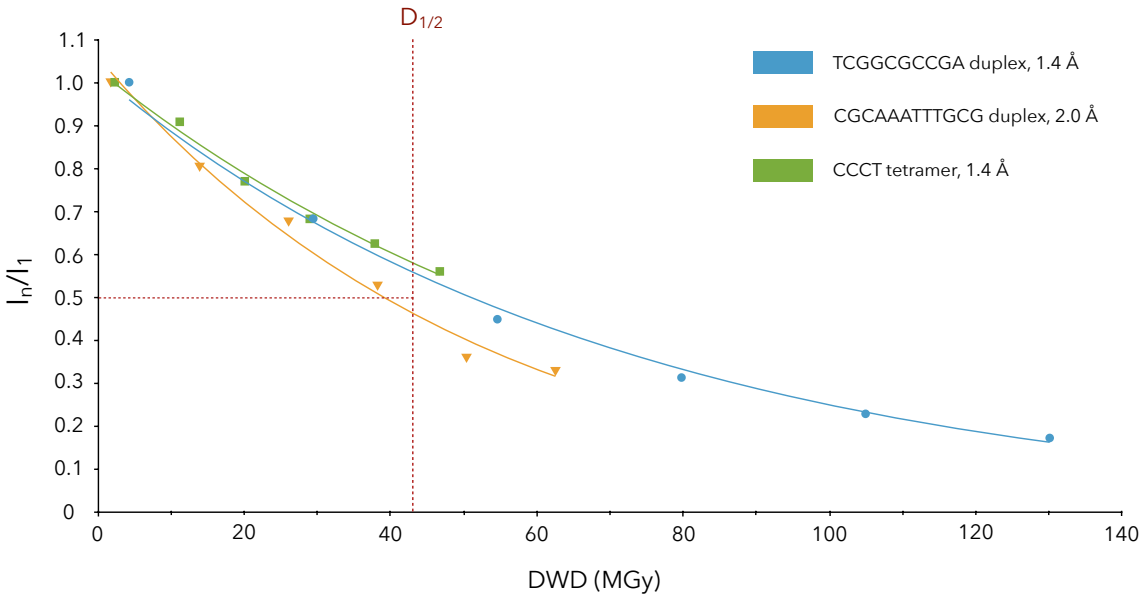


Figure 5.8: The decay in average intensity, I_n, relative to the initial diffraction dataset, I_1, as a function of diffraction-weighted dose. For each crystal, data points have been fitted to exponential decay functions. The experimentally-determined half-dose limit, $D_{1/2} = 43$ MGy, for protein crystals at 100 K has also been shown (Owen *et al.*, 2006).

with the $D_{1/2}$ limit for proteins. Due to the Gaussian beam profile used, it is noted that each crystal was inhomogeneously irradiated throughout data collection. Indeed, for the last 90° wedge of data on the TCGGCGCCGA crystal, a maximum dose value of 503.9 MGy was calculated by *RADDOSE-3D*, and the $D_{1/2}$ limit was exceeded in under 450° of rotation for the maximum dose metric.

It must be noted that for the TCGGCGCCGA duplex and CCCT tetramer crystal types, whereas the diffraction resolution was truncated to 1.4 Å based on the distance of the crystal from detector on the I02 beamline at DLS, both crystals were predicted to diffract to even higher resolution ($\sim$ 1.1 Å, as indicated by high resolution reflections detected at the detector corners). It therefore cannot be ruled out that the radiation-insensitivity of these crystals in Figure 5.8 was an artefact of the exclusion of these high resolution reflections. However, it is noted that the resolution estimate used for point group estimation by *POINTLESS* predicted a reduction in resolution from 1.15/1.22 Å to 1.62/1.93 Å, based on the criterion of either $CC_{1/2} > 0.6$ or $\langle I/\sigma(I)\rangle > 6$, between collection of the first and last 90° dataset on the TCGGCGCCGA duplex crystal (Table 5.1). Howells *et al.* (2009) have predicted a resolution-dependent dose limit of $D_{1/2} = 10 \times s$ MGy

De la Mora *et al.* (2011); Allan *et al.* (2013); Zeldin *et al.* (2013a)), and thus 43 MGy is strictly an upper bound for protein crystals at 100 K.

for proteins (for a diffraction spot of resolution s Å), however the reduction in resolution for the TCGGCGCCGA crystal falls significantly below this criterion. This provides additional evidence that the DNA crystals are robust to radiation damage relative to protein at 100 K.

5.3.1.1 The role of crystal packing

For proteins, crystal packing and intermolecular crystal contacts have been suggested to play a direct role in radiation damage susceptibility (Murray *et al.*, 2005), with loss of short range crystal ordering responsible for the reduction in diffracting power (Nave *et al.*, 2016). The robustness of the DNA crystals reported in the previous section suggests that the isomorphous packing in the crystals is preserved at high dose values. In this section, intermolecular DNA interactions will be identified that may be, in part, responsible.

The crystal packing interactions for each DNA crystal are illustrated in Figure 5.9. As shown in Figure 5.9a, the TCGGCGCCGA duplex exhibits multiple π-stacking interactions between nucleobases and the multiple aromatic rings of the intercalated Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand. Base stacking is known to play a dominant role in the thermal stability of the double helix in double stranded DNA (Yakovchuk *et al.*, 2006), and also between aromatic amino acid side chains in proteins (McGaughey *et al.*, 1998). It is therefore proposed that such stacking is also responsible for the observed radiation-insensitivity of the TCGGCGCCGA duplex crystal shown in Figure 5.8 by preserving the isomorphism of the crystal lattice.

In Figure 5.9b, it can be seen that the CCCT tetramer also exists intermolecular base stacking, in addition to a protonated thymine-thymine base pairing (involving two hydrogen bonds) between adjacent tetrameric molecules (Kang *et al.*, 1994). As a result, stacks of neighbouring CCCT tetramers form long strands in the crystal lattice. In addition, ordered colbalt hexammine molecules have been identified in the solvent channels running parallel to each DNA strand, which are suggested to also form stabilising interactions between phosphate backbone subunits on adjacent strands. It is suggested that a combination of these factors results in the rigidity of the CCCT lattice under X-ray irradiation.

In contrast, for the CGCAAATTTGCG duplex in Figure 5.9c, whereas adjacent duplexes also stack to form long strands (that extend the double helix topology between adjacent copies), only

base-ribose stacking was permitted between duplexes. These interactions have been theoretically calculated to be weaker (Sponer *et al.*, 1997), and could explain the increased radiation-sensitivity of the CGCAAATTTGCG duplex shown in Figure 5.8.

5.3.2 Site-specific damage observations

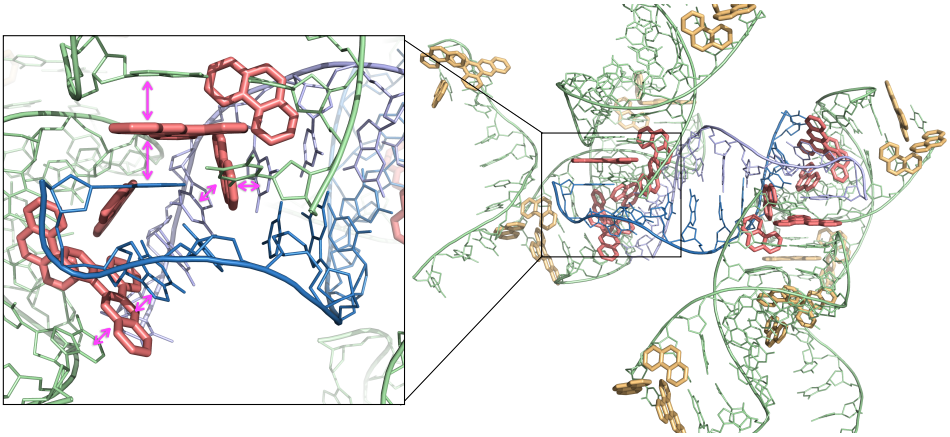
For each crystal, radiation-induced changes to localised electron density at atoms were detected by generating $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps, relative to each initial dataset. $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps were calculated using the protocol described previously in Section 2.4.1. It is important to note that site-specific damage analysis using $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps has only been performed here for higher dose datasets $n > 1$ generated by exposing an identical region of crystal volume as for dataset 1. For the CGCAAATTTGCG and CCCT sequences this corresponds to datasets $n =$ 3, 5, 7, 9 and 11, however for the TCGGCGCCGA duplex, only datasets $n =$ 5, 9, 13, 17 and 21 have been analysed.

Objective quantification of electron density loss sites has been performed using the program *RIDL*, as presented previously in Section 2.5. Figure 5.10, 5.11 and 5.12 report representative examples of $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps calculated over each DNA crystal type, in addition to corresponding analysis with *RIDL*. It was observed that other intermediate datasets exhibited similar trends in $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density and these have been not been included.

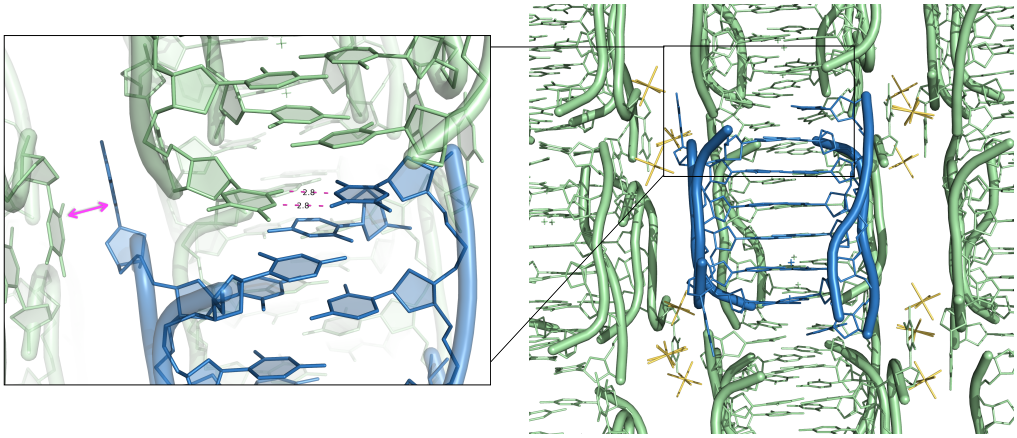
5.3.2.1 Sites of DNA electron density loss

As indicated in Figure 5.10b, 5.11b and 5.12b, for all three crystals, electron density loss was visually observed to accumulate in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps in the vicinity of backbone phosphate groups with increasing dose. This observation was subjective, and it was noted that considerable unexplainable $F_{\text{obs,n}} - F_{\text{obs,1}}$ map noise was also present in each map at the chosen σ map contouring level for manual inspection (fixed at $\pm 3\sigma$ in Figure 5.10, and $\pm 4\sigma$ in 5.11 and 5.12). This was particularly evident in the $F_{\text{obs,5}} - F_{\text{obs,1}}$ map for TCGGCGCCGA shown in Figure 5.10a.

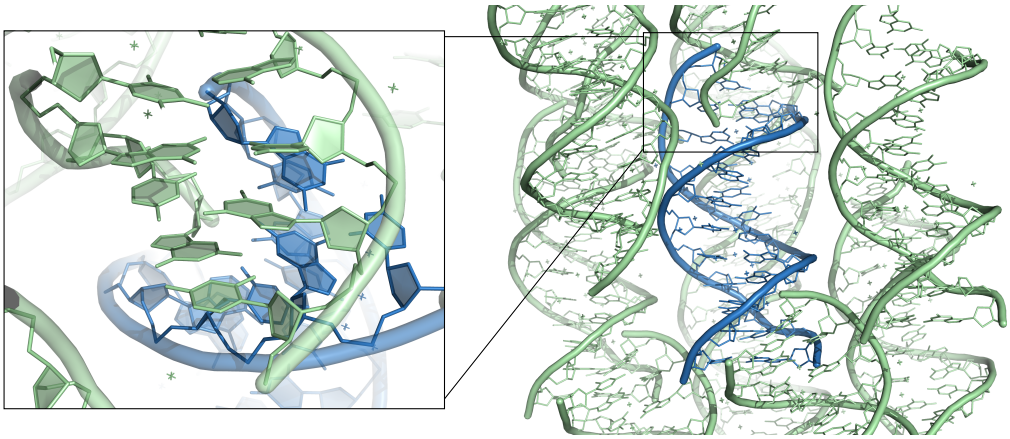
However, as the data in Figure 5.10c-d, 5.11c-d and 5.12c-d illustrate, atoms comprising the phosphate backbone consistently exhibited greater electron density loss across the large dose range



(a) TCGGCGCCGA duplex



(b) CCCT tetramer



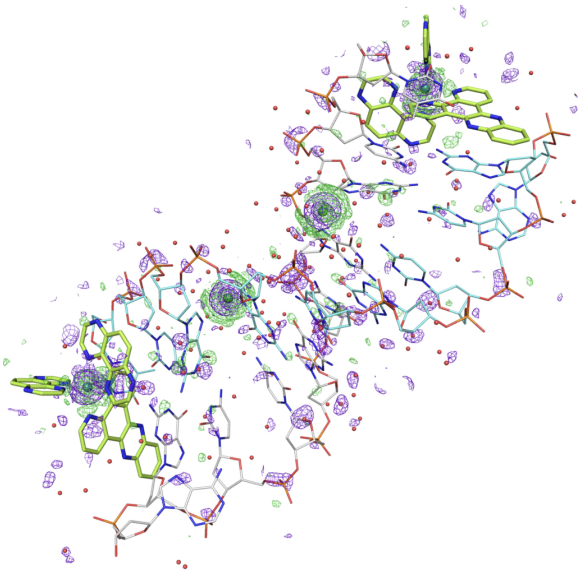
(c) CGCAAATTTGCG duplex

Figure 5.9: Crystal packing interactions for the (a) TCGGCGCCGA duplex, (b) CCCT tetramer and (c) CGCAAATTTGCG duplex. In (a) the single stranded TCGGCGCCGA is coloured dark blue, and bound to its symmetrically related complementary strand (in light blue). Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligands have been shown as sticks, with those interacting directly with the central DNA duplex shown in pink. Stacking interactions between the Δ-$[Ru(TAP)_2(dppz)]^{2+}$ ligand and DNA bases have been indicated with magenta arrows. In (b) the CCCT tetrameric complex is shown in blue, with symmetrically related copies in green. Symmetrically related copies interact through a protonated thymine-thymine base interaction (shown with magenta dashes), and a base-base stacking interaction (shown with magenta arrows). Ordered cobalt hexammine molecules have been detected in the solvent channels between parallel DNA strands, and are suggested to provide a stabilising bridge between phosphate backbone subunits on adjacent strands. In (c) the CGCAAATTTGCG duplex occupying the asymmetric unit is shown in blue with all symmetrically related copies shown in green. No direct base stacking interactions are observed between the duplex and its adjacent copy.

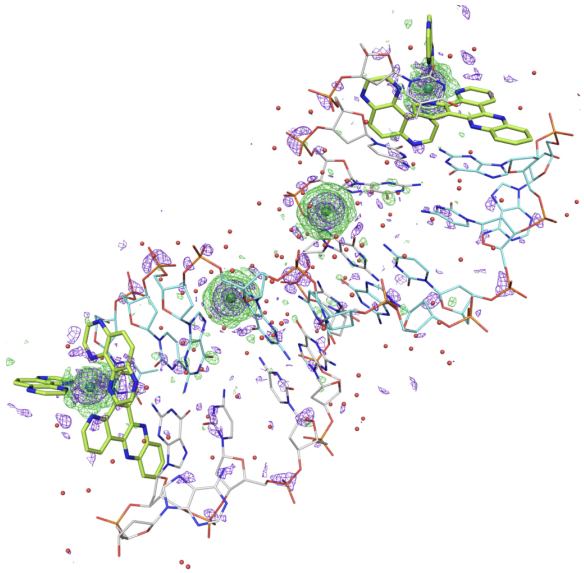
for each crystal. Here the D_{loss} metric has been presented as a quantifier for radiation-induced electron density loss, reporting the most negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map value in the localised region around each refined atom (Section 2.5.6). As indicated in Figure 5.10e-f, 5.11e-f and 5.12e-f, the distributions of D_{loss}(atom) values for each nucleobase type could not be distinguished. Since the dominant sites of radiation-induced density loss were located only at the phosphate backbone, this was anticipated.

In general, due to limited $F_{\text{obs,n}} - F_{\text{obs,1}}$ map quality, insufficient evidence was available to attribute phosphate electron density loss sites to any specific bond cleavage event. Figure 5.13a presents an example of a particularly dubious region of negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density for the CGCAAATTTGCG duplex crystal. Around phosphates T8, T9 and T20 on two complimentary strands, $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density accumulates with increasing dose, however is poorly aligned with the phosphate group. The map quality could be limited by the 2.0 Å resolution limit of the CGCAAATTTGCG crystal, or the quality of model building ($R_{\text{work}}/R_{\text{free}}$: 0.23/0.24). Indeed negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks are better aligned with phosphate units in the CCCT tetramer (sites C6, C7, C11 and T12 in Figure 5.13b), for which the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map had been calculated with data to 1.4 Å. It is also suggested that the poor $F_{\text{obs,n}} - F_{\text{obs,1}}$ map quality for the CGCAAATTTGCG crystal could be due to increased non-isomorphism between datasets: the CGCAAATTTGCG crystal underwent a larger unit cell expansion with dose (with approximately a 2 % volume increase between the first and last dataset, see Table 5.2), compared to that of the CCCT crystal for which the cell volume instead fluctuated by only $-0.38 - 0.03\,\%$ between datasets.

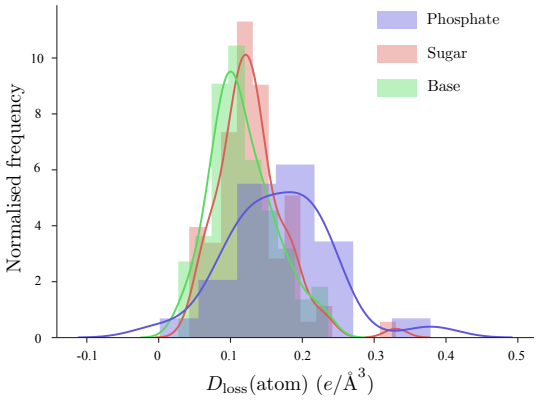
For the CCCT tetramer and TCGGCGCCGA duplex at 1.4 Å, $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks coincide with the P atom of each phosphate group, as opposed to either the 5' or 3' C–O phosphate-ribose bonds. This suggests that no specific bond is ruptured repeatedly across unit cells of the crystal, but instead, the phosphate group undergoes increased thermal motion as a direct result of X-ray energy deposition.



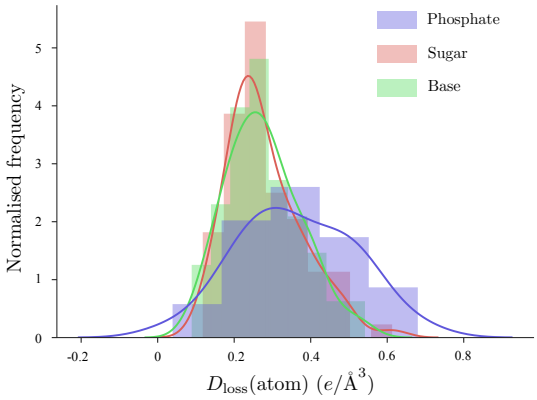
(a) $F_{\text{obs,5}} - F_{\text{obs,1}}$



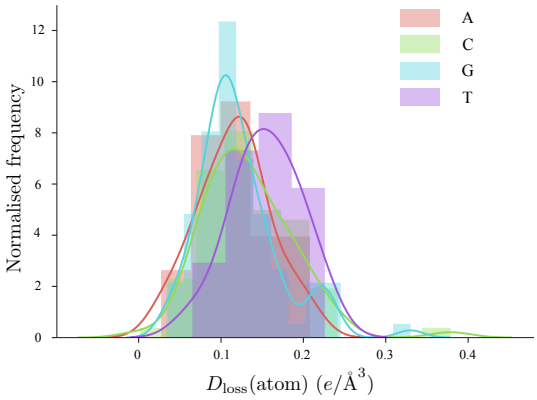
(b) $F_{\text{obs,13}} - F_{\text{obs,1}}$



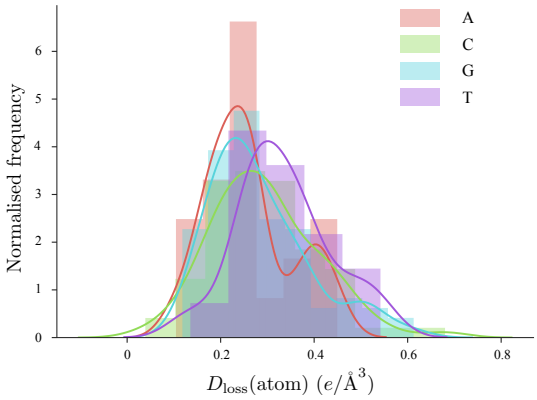
(c)



(d)



(e)



(f)

Figure 5.10: In (a) and (b), $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps have been calculated for the TCGGCGCCGA duplex crystal, for datasets (a) $n = 5$ and (b) $n = 13$. All difference maps are contoured at $\pm 3\sigma$ in green/purple. Ordered Ba^{2+} ions and Ru are shown as spheres. In (c) and (d) the distribution of D_{loss}(atom) metric values over all refined DNA atoms have been grouped by DNA subunit type, with D_{loss} values derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps for (c) $n = 5$ and (d) $n = 13$. In (e) and (f), similarly, the distribution of D_{loss} values have been grouped by nucleobase type for $n = 5$ and $n = 13$. D_{loss} metric values have been calculated by the program *RIDL*, as presented previously in Section 2.5.

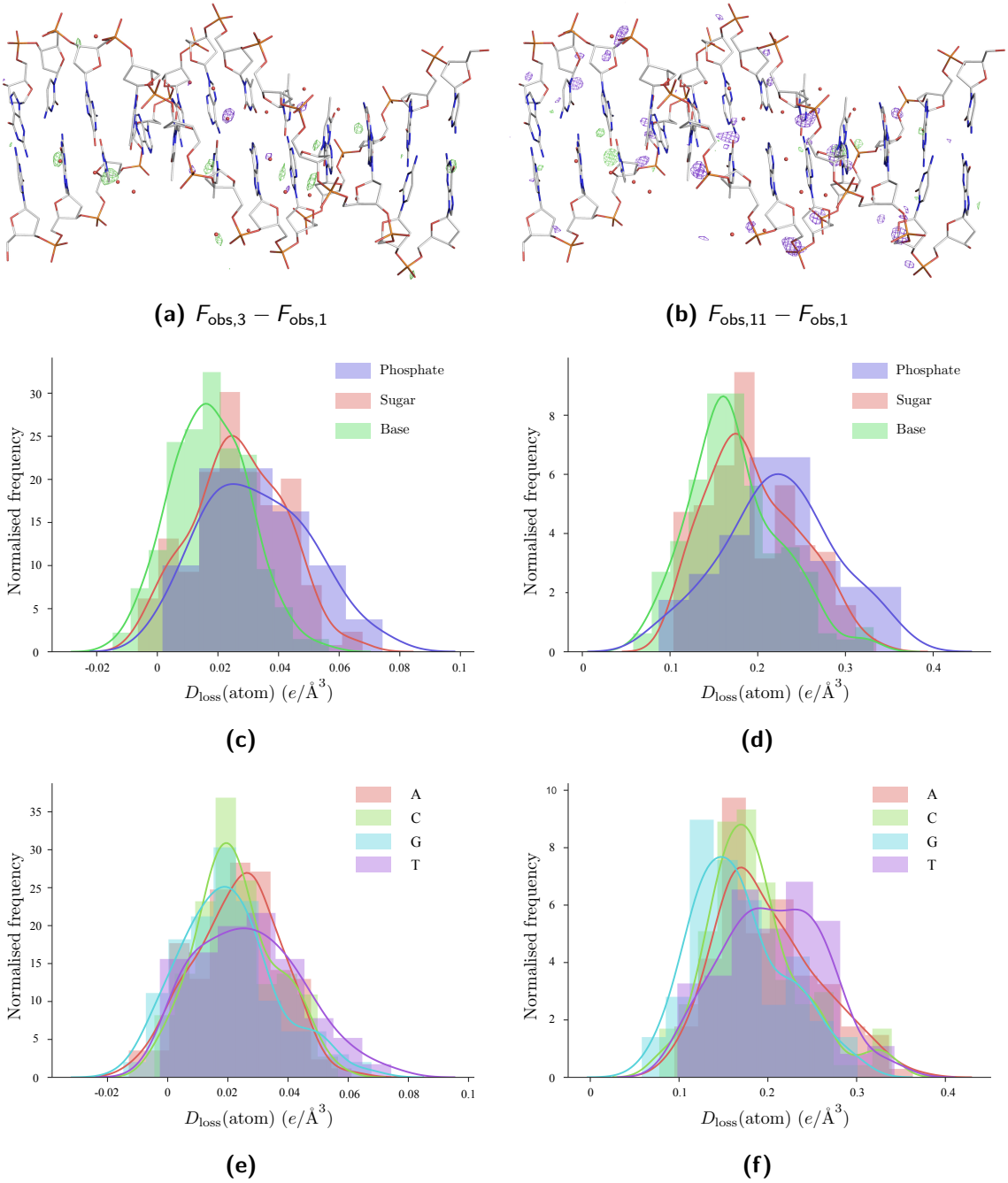


Figure 5.11: In (a) and (b), $F_{obs,n} - F_{obs,1}$ Fourier difference maps have been calculated for the CGCAAATTTGCG duplex crystal, for datasets (a) $n = 5$ and (b) $n = 13$. All difference maps are contoured at $\pm 4\sigma$ in green/purple. In (c) and (d) the distribution of D_{loss}(atom) metric values over all refined DNA atoms have been grouped by DNA subunit type, with D_{loss} values derived from $F_{obs,n} - F_{obs,1}$ maps for (c) $n = 5$ and (d) $n = 13$. In (e) and (f), similarly, the distribution of D_{loss} values have been grouped by base type for $n = 5$ and $n = 13$. D_{loss} metric values have been calculated by the program *RIDL*, as presented previously in Section 2.5.

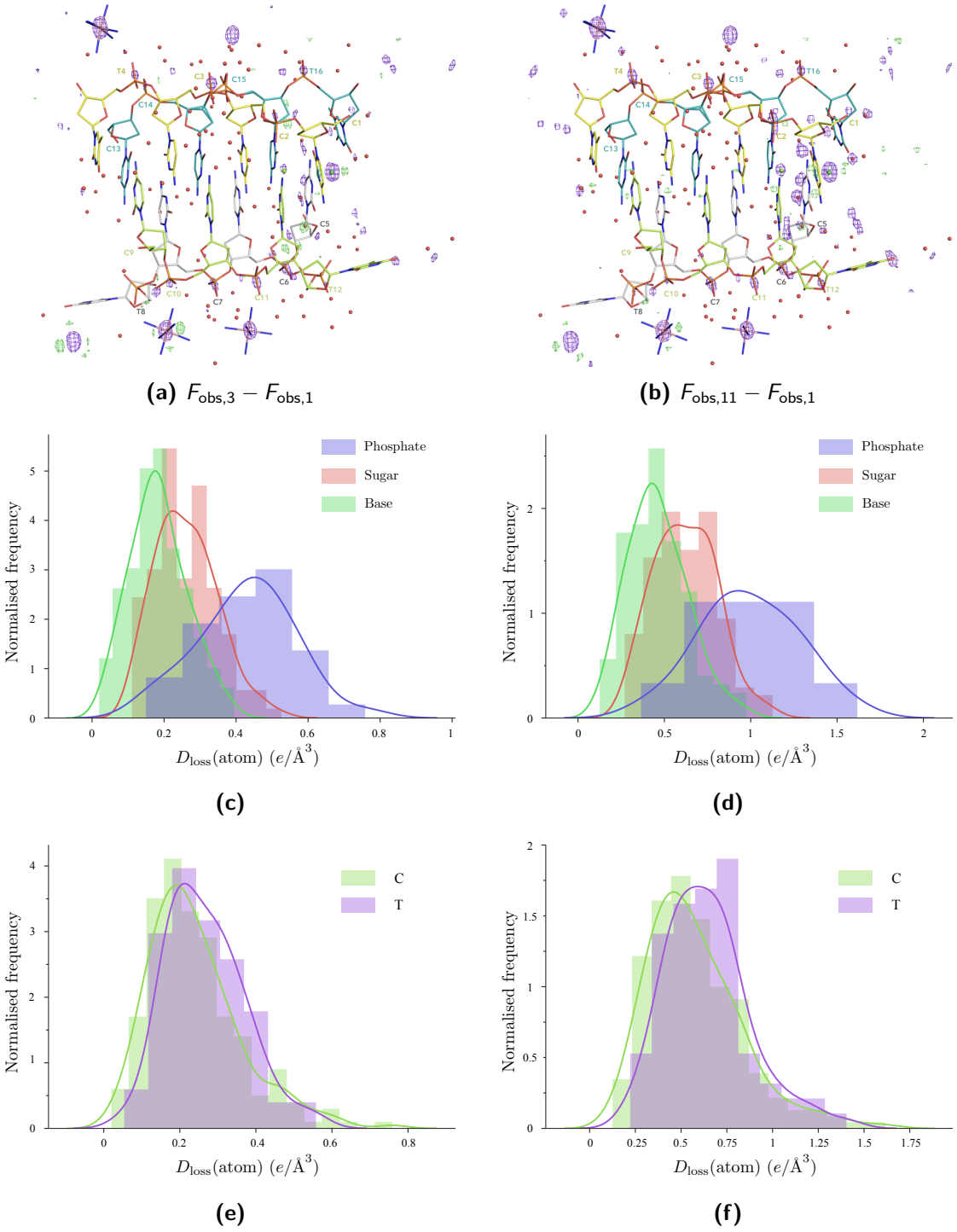
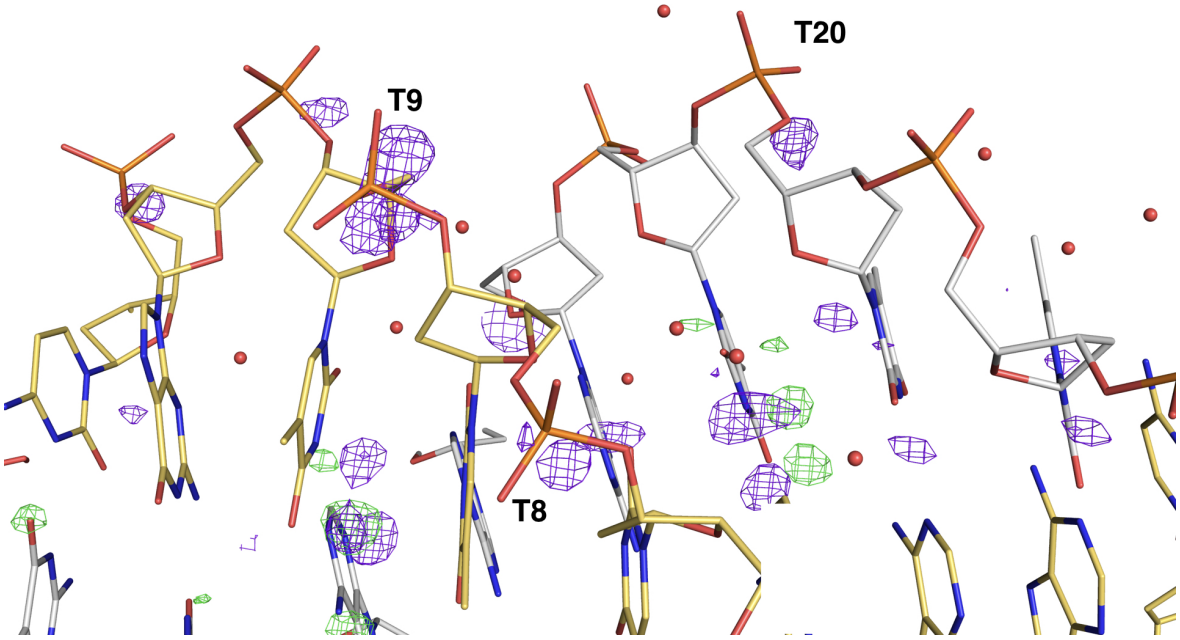
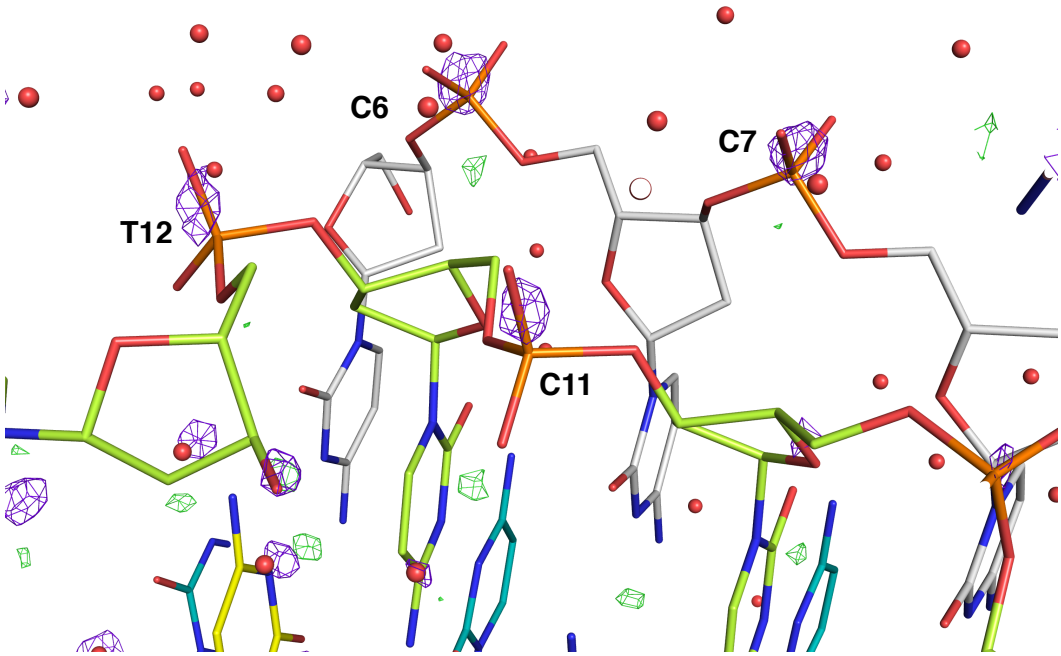


Figure 5.12: In (a) and (b), $F_{obs,n} - F_{obs,1}$ Fourier difference maps have been calculated for the CCCT tetrameric crystal, for datasets (a) $n = 5$ and (b) $n = 13$. All difference maps are contoured at $\pm 4\sigma$ in green/purple. In (c) and (d) the distribution of D_{loss}(atom) metric values over all refined DNA atoms have been grouped by DNA subunit type, with D_{loss} values derived from $F_{obs,n} - F_{obs,1}$ maps for (c) $n = 5$ and (d) $n = 13$. In (e) and (f), similarly, the distribution of D_{loss} values have been grouped by nucleobase type for $n = 5$ and $n = 13$. D_{loss} metric values have been calculated by the program *RIDL*, as presented previously in Section 2.5.



(a) CGCAAATTTGCG duplex



(b) CCCT tetramer

Figure 5.13: Examples of phosphate electron density loss in $F_{\text{obs},11} - F_{\text{obs}1}$ maps for the (a) CGCAAATTTGCG duplex and (b) CCCT tetramer crystals. For both crystals the difference map corresponding to the highest dose dataset ($n = 11$) have been calculated, and shown contoured at $\pm 4\sigma$ in green/purple. For visual clarity in both (a) and (b), carbon atoms situated on different strands are coloured differently.

5.3.2.2 The presence of Fourier ripples

As illustrated in Figure 5.10a-b, large $F_{obs,n} - F_{obs,1}$ map peaks were observable around the ordered Ba^{2+} and Ru atoms in the TCGGCGCCGA duplex asymmetric unit. The rings of positive and negative $F_{obs,n} - F_{obs,1}$ map density surrounding each metal site are a characteristic signature of Fourier ripples. In $2mF_{obs} - DF_{calc}$ maps, Fourier ripples can occur when unobserved structure factor amplitudes are not included in the Fourier summation used to calculate the electron density map (truncation errors). Indeed, in the current experiment the TCGGCGCCGA crystal diffracted beyond the recorded 1.4 Å limit enforced by the crystal-detector distance at beamline I02, DLS.

It can be speculated that for a metal-free protein/DNA crystal, the $F_{obs,n} - F_{obs,1}$ maps should be affected less by such truncation errors, since the omission of unobserved structure factor *differences* is equivalent to setting a Fourier summation coefficient to zero (i.e. no change in the intensity of a reflection), which is approximately true for a high resolution reflection. However, the above statement is only true if there is no damage to each Ru and Ba atom (either change in oxidation state or increased thermal motion), since as shown in Figure 5.14, these atoms have non-negligible atom scattering factors up to a high resolution of 1 Å. It thus appears highly likely that damage to the Ru and Ba metal atoms has introduced ripple effects, and consequently these have masked the original observation of damage. It is suggested that resolution extension (via the Free Lunch Algorithm (Usón *et al.*, 2007), for instance) could be a suitable strategy to mitigate such a Fourier truncation effect. However, this has not been performed in the current thesis.

Investigation of the nature of such $F_{obs,n} - F_{obs,1}$ map ripples is beyond the scope of the current work. However, it must be ensured that ripple effects do not dominate the rest of the $F_{obs,n} - F_{obs,1}$ map. In Figure 5.15 the results of two strategies that have been tested are shown. In Figure 5.15a, $F_{obs,n} - F_{obs,1}$ difference maps have been calculated with the resolution truncated further to 2 Å from 1.4 Å. By further truncating the $F_{obs,n} - F_{obs,1}$ map Fourier summation, the aim was to induce further Fourier ripples in the $F_{obs,n} - F_{obs,1}$ map and determine whether other sites of electron density loss persisted. Interestingly, Fourier ripples are diminished at 2 Å, particularly around the ordered Ba^{2+} sites. Electron density loss at the phosphate backbone and several ordered water molecules persisted. A similar effect was also observed when the $F_{obs,n} - F_{obs,1}$ map resolution was truncated to 2.5 Å and 3 Å.

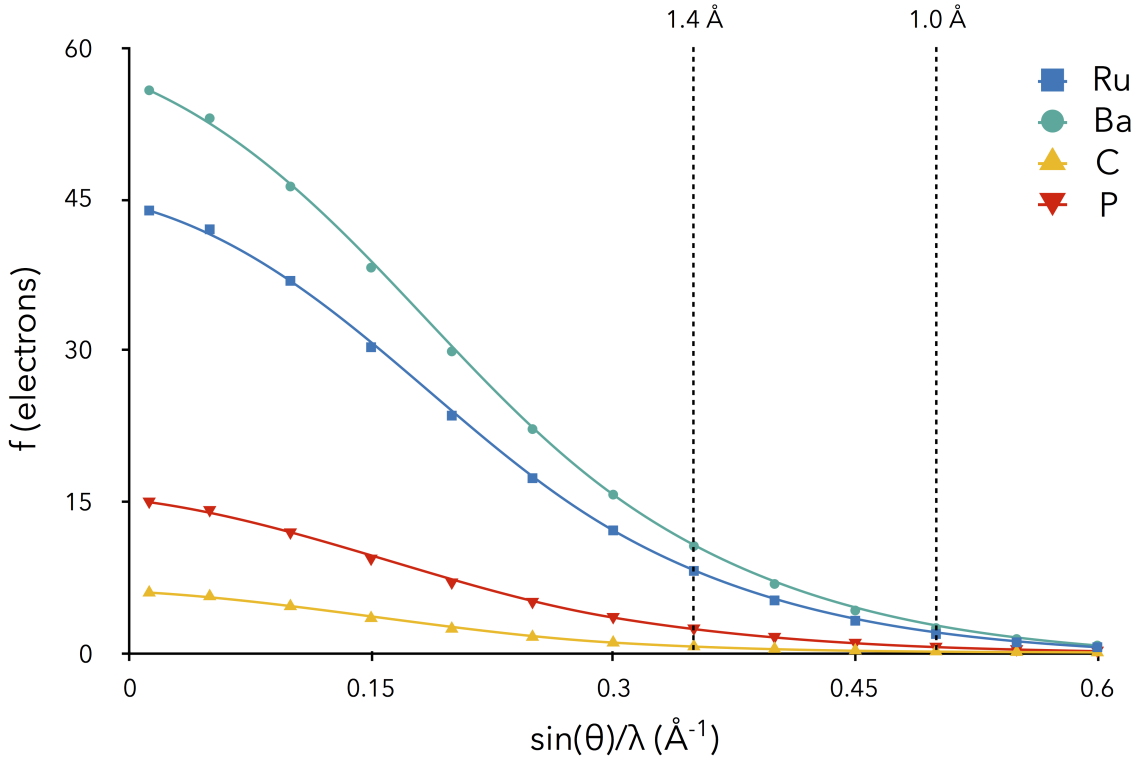


Figure 5.14: The atomic scattering factor f (in units of electrons) as a function of resolution (via Bragg's law) for Ru, Ba, P and C atoms, all of which are present in the TCGGCGCCGA crystal. Scattering factors have been calculated from Cromer-Mann coefficients using the online tool *http://www.ruppweb.org/new_comp/scattering_factors.htm*, assuming a fixed *B*-factor of 10 Å^2 and 100 % occupancy for each tested atom type.

In a second approach, a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map was instead calculated from the difference in diffraction data corresponding to the first and second full 360° rotation of the crystal. This is in contrast to all other $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps presented in this section for the TCGGCGCCGA crystal, for which each dataset corresponds to exactly 90° of crystal rotation. *AIMLESS* was used to scale and merge each 360° set of images separately in an identical procedure to that described in Section 5.2.3. This strategy was performed in an attempt to reduce the omission of Fourier summation terms by sampling a full 360° of reciprocal space for each dataset. In Figure 5.15b, Fourier ripples at the Ba^{2+} positions are still observed in this map. Sites of negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density at DNA phosphate subunits also persist.

Overall, these observations provide evidence that the effect of Fourier ripples at metal centres shown in Figure 5.10 is short ranged, however further work is required to prove this.

5.3.3 Relationship between phosphate disordering and DNA geometry

This section aims to distinguish whether a relationship exists between the magnitude of electron density disordered at each DNA phosphate group and local geometry along the DNA sequence. This is motivated by the example shown in Figure 5.16 of the TCGGCGCCGA sequence, for which D_{loss} metric values at each phosphate P atom varied across the length of the sequence.

The results presented in this section are restricted to the TCGGCGCCGA duplex DNA due to the higher resolution (1.4 Å) and superior refinement quality ($R_{\text{work}}/R_{\text{free}}$: 0.11/0.13) for that structure. However, similar conclusions were drawn for CGCAAATTTGCG duplex, which also exists as a B-form double helical assembly. The I-motif CCCT crystal structure will not be discussed, since the short 4 bp sequence length limited the validity of any conclusions.

The *3DNA* online server (Lu and Olson, 2003) was used to analyse the three-dimensional topology of each structure, starting from the refined `.pdb` coordinate models generated in Section 5.2.3 corresponding to each initial dataset. *3DNA* derives parameters to uniquely define the relative orientation and position of two bases forming a base pair, and the local stacking geometry of adjacent base pairs. In addition, the regularity of the helical structure is quantified by *3DNA* at each sequence position.

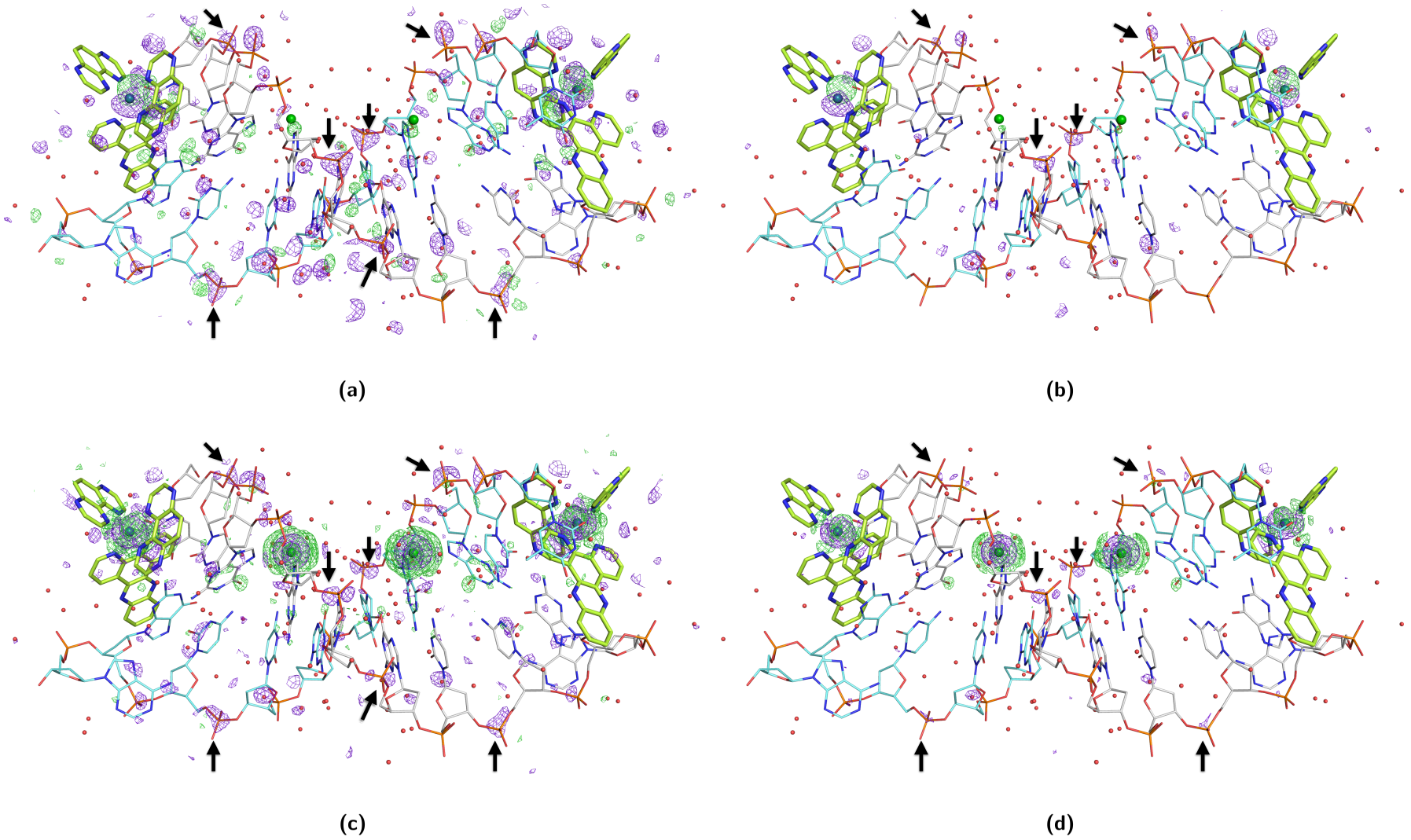


Figure 5.15: Analysis of Fourier ripples in $F_{obs,n} - F_{obs,1}$ maps for the TCGGCGCCGA duplex crystal. (a-b) The $F_{obs,5} - F_{obs,1}$ map has been calculated from diffraction data truncated to 2 Å. The resulting map has been contoured at (a) $\pm 3\sigma$ and (b) $\pm 4\sigma$ in green/purple. (c-d) A $F_{obs,n} - F_{obs,1}$ map has been calculated from the difference in diffraction data corresponding to the first and second full 360° rotation of the crystal (here to a maximum resolution of 1.4 Å). Each 360° of intensity data was merged separately using *AIMLESS* for the purpose of calculating this map. The resulting map has been contoured at (c) $\pm 3\sigma$ and (d) $\pm 4\sigma$ in green/purple. For each map, black arrows indicate phosphate backbone sites where significant electron density loss was visibly detected.

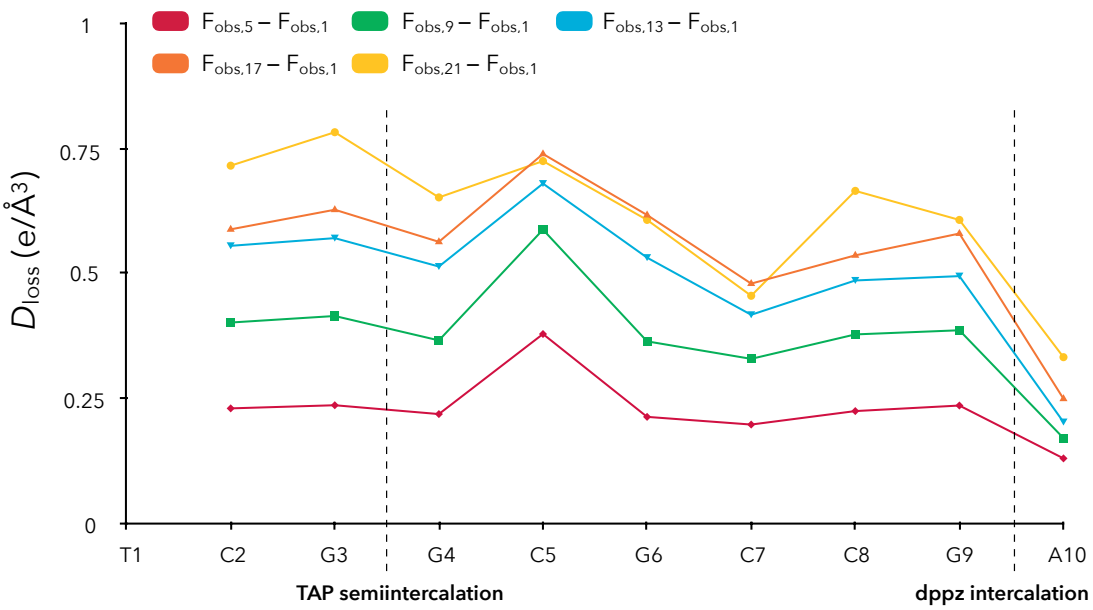
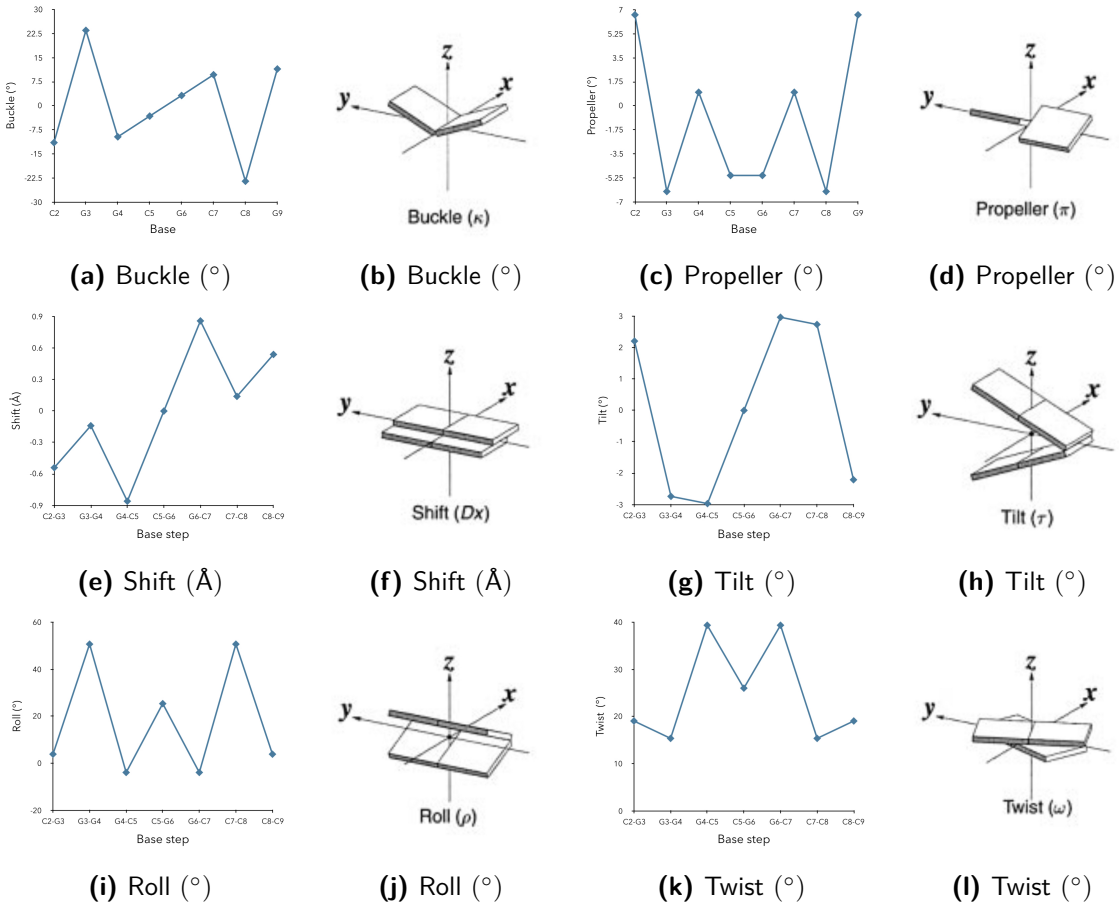


Figure 5.16: For TCGGCGCCGA crystal, the D_{loss}(atom) metric for phosphate backbone P atoms is shown as a function of sequence, with D_{loss} derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps for five higher dose datasets n = 5, 9, 13, 17 and 21. The sites of semiintercalation of the TAP ligand between bases G3 and G4, as well as dppz intercalation between bases G9 and A10 are marked. Due to the contorted geometry of the A10 nucleotide, dppz intercalation occurs between G9 and A10 of adjacent asymmetric units, not inside a single strand.

It is noted that the TCGGCGCCGA crystal consists of a single 10 base strand per asymmetric unit, as presented in the `.pdb` coordinate model; however, coordinates of the full duplex were required in *3DNA*. A custom *Python* script was written to generate the full biological assembly including the 10 base complementary strand. This was achieved using the 'REMARK 350' symmetry operations reported in the `.pdb` header information. The doubled stranded TCGGCGCCGA model was verified visually in *COOT* prior to submission to the *3DNA* server.

Figure 5.17 presents a subset of parameters that were identified to vary significantly along the sequence length for the TCGGCGCCGA structure. No clear correlation was determined between any individual geometry properties and the D_{loss} values shown in Figure 5.16, and there was insufficient evidence to conclude that phosphate disordering was dictated by local DNA geometry. A similar conclusion was also deduced for the CGCAAATTTGCG duplex, however the data have not been presented here.

The data in Figure 5.18 indicated the presence of a negative correlation between $F_{\text{obs,n}} - F_{\text{obs,1}}$ map disordering at each P position in TCGGCGCCGA and the isotropic *B*-factor attained in the first dataset coordinate model. This correlation became progressively higher with increasing dose.

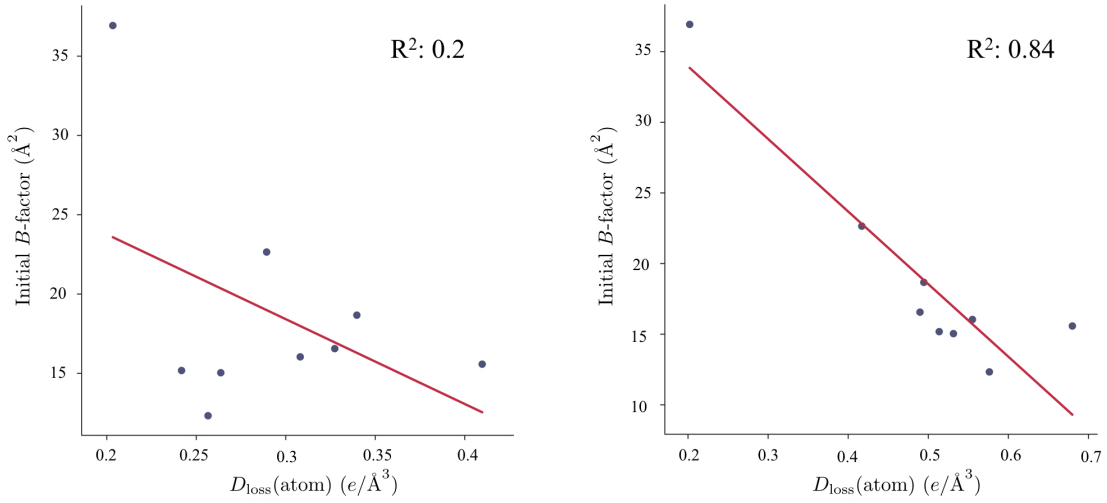


(a) Buckle (°) **(b)** Buckle (°) **(c)** Propeller (°) **(d)** Propeller (°)

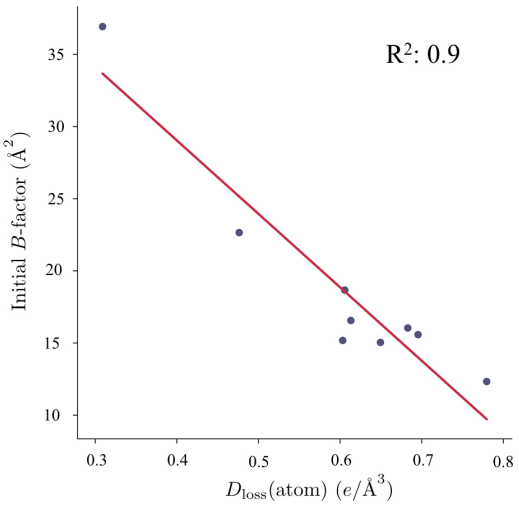
(e) Shift (Å) **(f)** Shift (Å) **(g)** Tilt (°) **(h)** Tilt (°)

(i) Roll (°) **(j)** Roll (°) **(k)** Twist (°) **(l)** Twist (°)

Figure 5.17: Parameters derived by *3DNA* defining the geometry and orientation between bases in each base pair, and between adjacent base pairs for the TCGGCGCCGA crystal. Only a subset of parameters have been shown for which values were observed to deviate significantly along the length of the TCGGCGCCGA sequence. As indicated by each x-axis, the *buckle* and *propeller* are defined per base pair, whereas all other parameters are defined between two adjacent base pairs. Schematics for each interaction parameter have been provided, taken directly from Lu and Olson (2003). In each schematic, each base is represented by a flat rectangular cuboid and the coordinate frame is such that the x vector points away from the minor groove edge of each base, and the y vector points towards the phosphate backbone.



(a) $F_{\text{obs,5}} - F_{\text{obs,1}}$ **(b)** $F_{\text{obs,13}} - F_{\text{obs,1}}$



(c) $F_{\text{obs,17}} - F_{\text{obs,1}}$

Figure 5.18: Correlation between D_{loss}(atom) metric and isotropic B-factor values attained in the first dataset coordinate model for P atoms in the TCGGCGCCGA duplex crystal. D_{loss} metric values are derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps corresponding to datasets (a) $n = 5$, (b) $n = 13$ and (c) $n = 17$. Linear R^2 values have been computed to assess the quality of the linear fit for each plot using the *SciPy* package in *Python*. Other intermediate datasets exhibited similar trends, such that, whereas no correlation was distinguishable initially, an increasing correlation became present with increasing dose.

This supports the hypothesis that the increasing observed phosphate disordering with dose is not sequence or position specific, but depends primarily on how well defined a phosphate is initially in the electron density map (i.e. has a relatively low *B*-factor). If a phosphate exhibits greater intrinsic disorder within the crystal lattice prior to irradiation, the extent of further disordering it can undergo during X-ray exposure is reduced. It is noted that this conclusion is restricted by the limited quantity of data shown in Figure 5.18, and further analysis would be required on larger nucleic acid sequences than in the current study in order to validate this claim.

5.4 Conclusion

The aim of the current chapter has been to investigate the radiation-susceptibility of DNA crystals in cryocooled MX experiments. This was in direct contrast to the results reported in Chapters 3 and 4, which exclusively considered oligonucleotides in pre-formed bacterial nucleoprotein complexes. In such complexes, nucleic acid was determined to be highly radiation insensitive relative to protein, with electron density loss coinciding with the phosphate backbone present only at high dose values (> 20 MGy). At such high doses, pronounced global damage effects (notably the loss in intensity of high resolution diffraction spots), were also prominent within the data.

A benefit of studying such complexes stems from the ability to compare damage to the nucleic acid directly to bound protein, for which radiation damage effects are far better characterised (Weik *et al.*, 2000; Fioravanti *et al.*, 2007; Sutton *et al.*, 2013). However, without an unbound nucleic acid control in these studies, it could not be distinguished whether nucleic acid was intrinsically robust to radiation damage effects, or whether each bound protein species contributed to this lack of susceptibility. The hypothesis was presented that protein-binding may physically stabilise the nucleic acid structure, or provide a sink for damaging radical species and electrons that are able to migrate in the crystal environment at 100 K. This chapter now aimed to determine whether DNA became susceptible to radiation damage effects in the absence of bound protein.

Three DNA crystal forms have been considered, of variable sequence length, metal content and conformational form (B-form and I-motif). Site-specific damage effects were investigated in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps, and the D_{loss}(atom) metric (Section 2.5) has been used as a tool to objectively quantify changes in these maps. Consistent with the results for the C.Esp1396I-DNA and TRAP-

RNA complexes presented in Chapters 3 and 4, the phosphate backbone atoms of each structure exhibited greater electron density loss relative to the basic and ribose DNA subunits. For a controlled experiment, $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps have only been investigated for higher dose datasets n corresponding to an identical region of crystal as the first dataset, to ensure that an identical crystal volume and region of reciprocal space were sampled between datasets. Unfortunately, this criterion limited the minimum dose difference between datasets that could be rigorously investigated. As such, the accumulated dose for the onset of such observable effects was calculated to be 29.5 MGy (relative to 4.29 MGy at the first dataset), 15.3 MGy (relative to 3.1 MGy) and 11.2 MGy (relative to 2.3 MGy) for the TCGGCGCCGA, CGCAAATTTGCG and CCCT crystal sequences. It is noted that due to the high metal content in the TCGGCGCCGA crystal (containing both Ru and Ba^{2+}), an unavoidably large absorbed dose rapidly accumulated in that crystal form. The reported dose values for phosphate damage onset are therefore of the same order as for complexes described in Chapters 3 and 4. No evidence has been provided that these are strict lower dose bounds for such phosphate damage effects, and further experiments on DNA crystals irradiated at lower flux values than the current experiment would be required to determine this.

An attempt to establish a correlation between the magnitude of electron density disordering at individual phosphate sites and local DNA geometry was unsuccessful, and no differential dependence on base type could be distinguished. For the TCGGCGCCGA and CGCAAATTTGCG duplex crystals, the initial intrinsic disorder of each phosphate in the crystal lattice appeared to play a dominant role in the dose-dependent electron density loss that was subsequently measured in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps, since a negative correlation did exist between isotropic B-factor of P atoms attained in the first dataset coordinate model and D_{loss} values attained in higher dose datasets. It thus appears that in principle, if all phosphates were initially equally ordered in the DNA backbone, radiation-induced disordering could occur at any position with equal probability. However, in practice, phosphates exhibit variable degrees of intrinsic disorder, and logically only those that are initially well ordered can then be disordered by irradiation.

The three DNA crystals were also found to be highly resistant to global damage effects, with the TCGGCGCCGA and CCCT crystal forms exhibiting greater insusceptibility than the CGCAAATTTGCG duplex crystal. Approximate DWD half-dose values, $D_{1/2}$, of 52 MGy and 40 MGy were determined for the TCGGCGCCGA and CGCAAATTTGCG crystals, however for the CCCT

crystal, the relative diffraction intensity decay, I_n/I_1 had not decayed to 50 % by the end of the experiment (42.3 MGy) and an accurate half-dose could therefore not be extracted. With the exception of the CGCAAATTTGCG crystal, these $D_{1/2}$ values do marginally exceed the 43 ± 3 MGy limit that was experimentally determined as an upper bound for protein crystals at 100 K (Owen *et al.*, 2006). Further investigation would be required in order to determine whether this trend persists for wider range of nucleic acid crystals, and to rigorously establish a $D_{1/2}$ dose limit for DNA at 100 K. Work in this area is currently being conducted (D. Barrett, unpublished), including the determination of the relative global damage susceptibilities of DNA crystals at 100 K and RT.

In proteins, the anecdotal notion that site-specific damage precedes global damage is in part due to the folded structure of proteins, which permits isolated radiation-induced bond cleavage events to occur at the interior of the protein, whilst still maintaining crucial crystal contacts and overall isomorphism between unit cells. However, for nucleic acid crystals this may not be true. Due to the high frequency of phosphate groups per asymmetric unit in DNA and their importance in maintaining the DNA assembly, the presence of backbone bond cleavage events reproducibly across unit cells would inevitably lead to fragmentation of the DNA double helix and rapid conformational heterogeneity throughout the crystal volume. In the current investigation, the diffracting power of the crystal (and thus crystal isomorphism) was preserved up to high dose values, implicitly suggesting that any site-specific backbone damage must also be negligible.

CHAPTER 6

Characterisation of site-specific radiation damage to Tyrosine

6.1 Introduction

During MX data collection from proteins at 100 K, site-specific damage is now widely believed to preferentially occur to particular protein moieties, and in a reproducible order between independent protein systems (Weik *et al.*, 2000; Fioravanti *et al.*, 2007; Sutton *et al.*, 2013). In particular, changes in the oxidation state of bound metal centres, and bond cleavage of disulphides, acidic residue carboxyl groups, and Met methylthiol side-groups have all been detected at doses of just several MGy at 100 K. Accurate knowledge of which sites are most prone to radiation damage effects (as well as the accumulated dose for the crystal) allows the experimenter to recognise where radiation damage has affected local protein conformation and interactions; if undetected, such damage sites can lead to poorly assigned geometry and incorrect biological conclusions to be drawn regarding the function of a protein. This is particularly relevant to ligand-binding and time-resolved studies, since active sites in general appear to be inherently radiation-sensitive (Dubnovitsky *et al.*, 2005), and are already frequently challenging to characterise due to ligands often being bound only at low occupancies. With knowledge of the preferential MX damage pathways, it is also sometimes possible to then employ appropriate small molecule radioprotectant compounds to mitigate damage at these sites, in order to preserve the integrity of the protein molecule, the crystal packing interactions between proteins in the crystal, and ultimately the life-dose of the irradiated crystal (De la Mora *et al.*, 2011; Allan *et al.*, 2013).

Tyrosine has also previously been thought to be radiation-sensitive at 100 K. During diffraction data collection from X-irradiated myrosinase crystals, Burmeister (2000) reported progressive accumulation of negative $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference density centred around the phenolic oxygen on an active site tyrosine residue, Tyr-330, coupled with observable positive difference density in the adjacent solvent region (Figure 6.1). In addition, a decaying −OH group occupancy with increasing dose was reported. Myrosinase functions to hydrolyse glucosinolates, predominantly in the Brassicaceae family, with Tyr-330 situated in a glucose binding site and exhibiting a hydrogen bond interaction to an acidic residue, Glu-409 in the absence of bound glucose (Burmeister *et al.*, 1997). Tyr-330 is predicted to reorientate to accommodate glucose insertion, with the Tyr hydroxyl group then positioned adjacent to the sugar ring oxygen (with no possibility to then form a hydrogen bond to Glu-409). In the Burmeister (2000) study, an ordered glycerol molecule

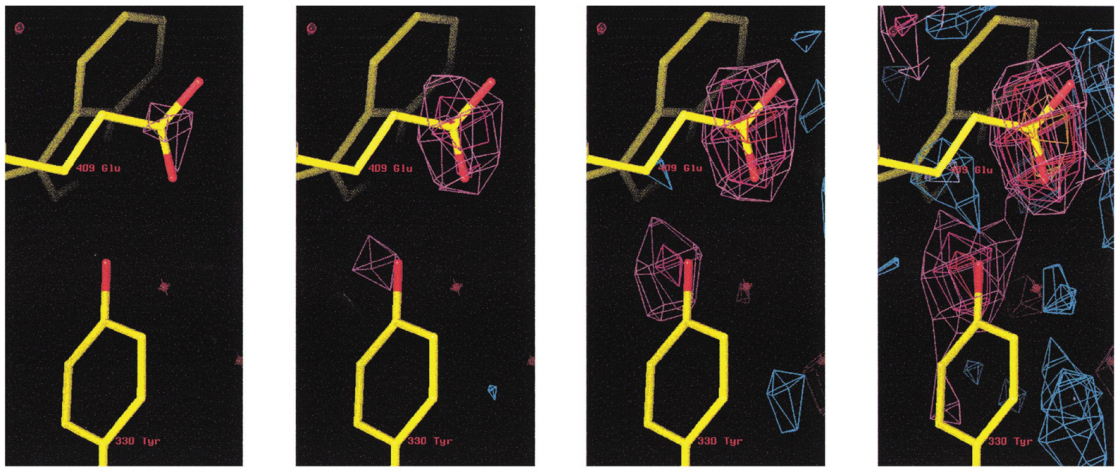


Figure 6.1: The original observations of electron density loss in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps (calculated at 2.4 Å localised to the −OH group of Tyr-330 in myrosinase, as reported in Burmeister (2000). Figure has been taken directly from Burmeister (2000). Progressive decarboxylation to Glu-409 was also demonstrated. The dose for each higher dose dataset n were reported in Burmeister (2000) to be (a) 9×10^{15} ph/mm^2, (b) 18×10^{15} ph/mm^2, (c) 27×10^{15} ph/mm^2 and (d) 54×10^{15} ph/mm^2. Note that these doses have been reported in units of flux density, and not in the more commonly used units in MX of MGy. For each dataset, $F_{\text{obs,n}} - F_{\text{obs,1}}$ electon density maps have been contoured at -0.25 (orange), -0.2 (red), -0.15 (pink), -0.1 (magenta), 0.1 (cyan), 0.15 (light blue), 0.2 (turquoise) and 0.25 e/Å^3 (blue).

instead occupies the glucose-binding site, and Tyr-330 initially exhibits the hydrogen bond to Glu-409. Glu-409 was observed to undergo radiation-induced decarboxylation, leading Burmeister to postulate the existence of a correlation between Tyr −OH damage and such hydrogen bonding interactions. The loss of the Tyr-330-stabilising hydrogen bond would permit increased lability of the Tyr-330 aromatic ring, and would manifest itself in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps mostly clearly at positions furthest from the Tyr C_β hinge-point (i.e. at the Tyr −OH group). Burmeister also suggested the possibility that an −OH cleavage event could have occurred at this position (see the legend of Figure 3c in Burmeister (2000)).

Overall, the suggestion of Tyr −OH group cleavage in myrosinase was surprising, since whereas tyrosine residues are well known to be redox active, cleavage of the phenolic C−O bond has not been reported in the proposed mechanisms of tyrosine radiation chemistry. Indeed the redox couple involving tyrosine and the tyrosyl radical plays a key role in high potential redox catalysis for a wide range of natural enzymes including, for example, photosystem II (Rappaport and Diner, 2008), ribonucleotide reductase (Stubbe *et al.*, 2003), cyctochrome c oxidase (Kaila *et al.*, 2010), galactose oxidase (Whittaker, 2003) and prostaglandin synthase (Tsai and Kulmacz, 2010), suggesting a robustness of the phenolic linkage. Moreover, density functional theory (DFT) cal-

culations have suggested that the C−O bond strength is high, at approximately 110 kcal/mol and that it is not significantly different in the oxidized species (I. Carmichael, private communications). In contrast, in a neutral disulphide, the typical S−S bond strength is much lower at 55 kcal/mol (Benson, 1978), and this strength is lowered further to only 23 kcal/mol, in the reduced anionic species.

An extensive literature on the aqueous radiation chemistry of tyrosine-containing solutions at room temperature also suggests the relative persistence of the tyrosyl radical. For example, from pulse radiolysis studies it is clear that attack by the strongly oxidizing hydroxyl radical, which is formed by the radiolysis of solvent water, leads predominantly to substitution on the aromatic ring. In the presence of oxygen, the eventual products are 2,4- and 3,4-dihydroxyphenylalanines (Chrysochoos, 1968; Solar *et al.*, 1984), with minor pathways also existing to produce ring-ring dimerisation and H-abstraction from the C_β carbon (Solar *et al.*, 1984). Other, less potent, one-electron oxidants also initially lead to the formation of the tyrosyl radical, opening pathways leading to a variety of Tyr modifications, all of which preserve the phenolic C−O linkage (Houée-Levin *et al.*, 2015). In the solid state, soft X-ray irradiation of the respective powders showed phenylalanine degrading more rapidly than tyrosine, with the latter exhibiting a prominent stability against radiation damage (Zubavichus *et al.*, 2004). Also, γ-radiolysis at low temperature cleaves the former to give a benzyl radical, but again only produces the tyrosyl radical in the latter (Fasanella and Gordy, 1969).

However, despite the ambiguity in the reason behind Burmeister's finding of density loss around the tyrosyl oxygen in myrosinase, and the weight of evidence from radiation chemistry studies for tyrosyl persistence, as detailed above, such Tyr −OH electron density loss has subsequently been interpreted as −OH dissociation for protein crystals at 100 K. The Burmeister (2000) publication has now been cited $>$ 370 times*, and Tyr −OH cleavage is regularly cited as one of the four dominant non-metalloprotein specific damage events present in MX experiments conducted at 100 K (Garman, 2010; Dubnovitsky *et al.*, 2005; Owen and Sherrell, 2016). Moreover, other than for myrosinase, to the author's knowledge there have been no other reports of clear Tyr −OH damage supported by electron density for protein crystal systems irradiated at 100 K. It is noted that studies (e.g. Leiros *et al.* (2001) on trypsin) have reported a reduction in occupancy (and

*As of 19th July 2017, as reported by the National Center for Biotechnology Information (NCBI)

increased atomic *B*-factor) for the O_{η} atom in Tyr residues following irradiation, but these have provided insufficient data to reproduce this observation.

The current chapter has three main aims. First, the originally deposited myrosinase diffraction data in conjunction with the Burmeister (2000) investigation are critically re-analysed to determine whether sites of Tyr −OH electron density loss are reproducible, with the use of the newly developed tool (*RIDL*, Section 2.5) that has been designed for objective quantification of per-atom dose-dependent density changes. Secondly, all other available MX radiation damage series that have previously been deposited in the PDB are retrieved and systematically screened for indications of Tyr −OH radiation damage with aid of the high-throughput analysis provided by *RIDL*, in order to better characterise the reproducibility of Tyr −OH electron density loss across independently irradiated protein crystals. Finally, a feasible mechanism is developed to explain any reproducible observations of Tyr −OH group electron density loss, that does not rely upon bond rupture of the robust C−O linkage.

6.2 Reanalysis of the myrosinase damage series (Burmeister, 2000)

This section aims to re-assess the quality of the originally deposited myrosinase damage series (Burmeister, 2000), in order to determine whether Burmeister's report of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map difference density isolated at the −OH group of Tyr residues is valid.

6.2.1 Acquisition of the MX damage series data

The coordinate (`.pdb` file) and structure factor information (in a `.mmcif` file format) were retrieved from the PDB (accession codes: *1dwa*, *1dwf*, *1dwg*, *1dwh*, *1dwi* and *1dwj*. The `.mmcif` format was converted to `.mtz` format with CCP4 program *CIF2MTZ*. Merged diffraction data had been deposited to a maximum resolution of 2.0 Å for all datasets. For the original publication, two coordinate models (*1dwi* and *1dwj*) had been deposited for the highest dose state, corresponding to occupancy and positional refinement, respectively, that had been performed by Burmeister (2000). Only the *1dwi* coordinate model was used for the current analysis.

To provide a direct comparison of the quality of the original refinement, coordinate and structure

factor files were additionally extracted from the online *PDB_REDO* server (Joosten *et al.*, 2014), both after initial file curation by *PDB_REDO* (ten cycles of rigid body refinement in *REFMAC* (Murshudov *et al.*, 2011)) and also after full model re-building in *PDB_REDO*.

6.2.2 Evaluation of myrosinase model building quality

For the first dataset in the myrosinase series (PDB accession code: *1dwa*), visual inspection of the $2mF_{obs} - DF_{calc}$ map retrieved from the Electron Density Server (EDS) revealed regions of poor fits of this model for ordered solvent and side-chain conformations to the electron density (overall map real-space correlation coefficient (RSCC): 67 %). Poor fits to $2mF_{obs} - DF_{calc}$ map density were visible for several Tyr aromatic groups (most notably Tyr-215 as shown in Figure 6.2a, for which the RSCC was calculated to be only 58.1 %).

In the Burmeister (2000) damage study, the initial low-dose coordinate model was generated by only rigid body and overall anisotropic *B*-factor refinement (but no positional refinement), starting from a 1.2 Å myrosinase model derived from a different crystal that had been used for a previous experiment (Burmeister *et al.*, 2000) that aimed to structurally characterise bound cofactors of myrosinase (PDB: *1e4m*). Indeed all-atom structural alignment of the *1dwa* and *1e4m* structures using the CCP4 program *GESAMT* resulted in a Q-score of 1.00 and low RMSD score of 0.007 Å. This sub-optimal refinement strategy may account for the poor density map-model RSCC calculated for the original data. Most strikingly, the R_{work}/R_{free} as provided by the `.pdb` header originally deposited in the PDB (0.165/0.179) did not agree with those recalculated by the wwPDB validation report associated with PDB deposition (0.436/0.447).

For the *1dwa* structure, it was determined that refinement in *PDB_REDO* significantly improved the overall quality of the built model. Initially *PDB_REDO* reported a poor agreement between R_{work}/R_{free} as provided by the `.pdb` header originally deposited in the PDB (0.165/0.179) and those recalculated directly from the deposited data (0.191/0.201) prior to any model re-building in *PDB_REDO*. To recalculate these values, the *PDB_REDO* server performed a 0-cycle round of rigid body refinement using *REFMAC*, followed by an additional 10 cycles of rigid body refinement, This is the standard procedure implemented in *PDB_REDO* when the R_{work}/R_{free} reported in the original `.pdb` header are not reproducible to within 5 % after the single 0-cycle *REFMAC* run.

Table 6.1: Distribution of real-space Z_{obs} (RSZO) scores calculated by *EDSTATS* (Tickle, 2012) for the *1dwa* structure retrieved directly from PDB and retrieved from *PDB_REDO* (Joosten *et al.*, 2014) after 10 cycles of rigid body refinement. RSZO is a metric for model precision and is calculated as RSZO $=$ mean$(\rho_{\text{obs}})/\sigma(\Delta\rho)$, in which, $\sigma(\Delta\rho)$ is the standard uncertainty of the $F_{\text{obs}} - F_{\text{calc}}$ map for *1dwa*. The percentage of residues in *1dwa* with lower RSZO scores than specific thresholds (between 0.5 and 3) are provided. The RSZO score is unbounded from above, with larger positive values indicating greater residue to density fit. For an individual residue, *EDSTATS* recommends a default < 1 rejection threshold. For the *1dwa* coordinate model retrieved directly from the PDB, a significantly larger number of residues attain lower (poorer) RSZO scores, compared to the *1dwa* data retrieved from *PDB_REDO* instead.

	Retrieved from the PDB			Retrieved from *PDB_REDO*		
RSZO score	Main (%)	Side (%)	Overall (%)	Main (%)	Side (%)	Overall (%)
0.5	0.2	3.0	2.8	0.0	0.0	0.0
1.0	3.0	18.6	16.6	0.2	1.8	1.8
1.5	22.2	42.3	42.3	0.8	6.4	6.0
2.0	75.8	80.2	90.4	1.4	13.8	12.2
2.5	99.2	96.1	99.8	3.4	19.1	17.0
3.0	100.0	97.5	100.0	6.2	24.8	22.4

The 10 cycles of rigid body refinement for *1dwa* resulted in a significant increase in overall map RSCC to 92.0 %, and improvement in the fit of individual Tyr side-chains to 2mF_{obs} − DF_{calc} map density (Tyr-215 RSCC at 90.8 %). Corresponding real-space Z_{obs} scores calculated by *EDSTATS* (Tickle, 2012) both pre- and post- further rigid body refinement are provided in Table 6.1 for assessment of the structure-wide improvement in model to map fit following re-refinement.

Further full model rebuilding in *PDB_REDO* again improved the refinement statistics ($R_{\text{work}}/R_{\text{free}}$: 0.13/0.16 with RSCC: 97 %). The improvement in electron density fit for the example case of Tyr-215 is shown in Figure 6.2.

6.2.3 Dose recalculation

Prior to reanalysis of site-specific damage in myrosinase, it was crucial to obtain an accurate estimate of the accumulated dose at each discrete dataset. In Burmeister (2000), 'dose' values had originally been quoted in units of photons/mm^2 (flux density), and consequently required conversion into the correct dose units of MGy. Diffraction-weighted doses, DWD, have been calculated in *RADDOSE-3D* for this damage series using experimental details as reported in Burmeister (2000).

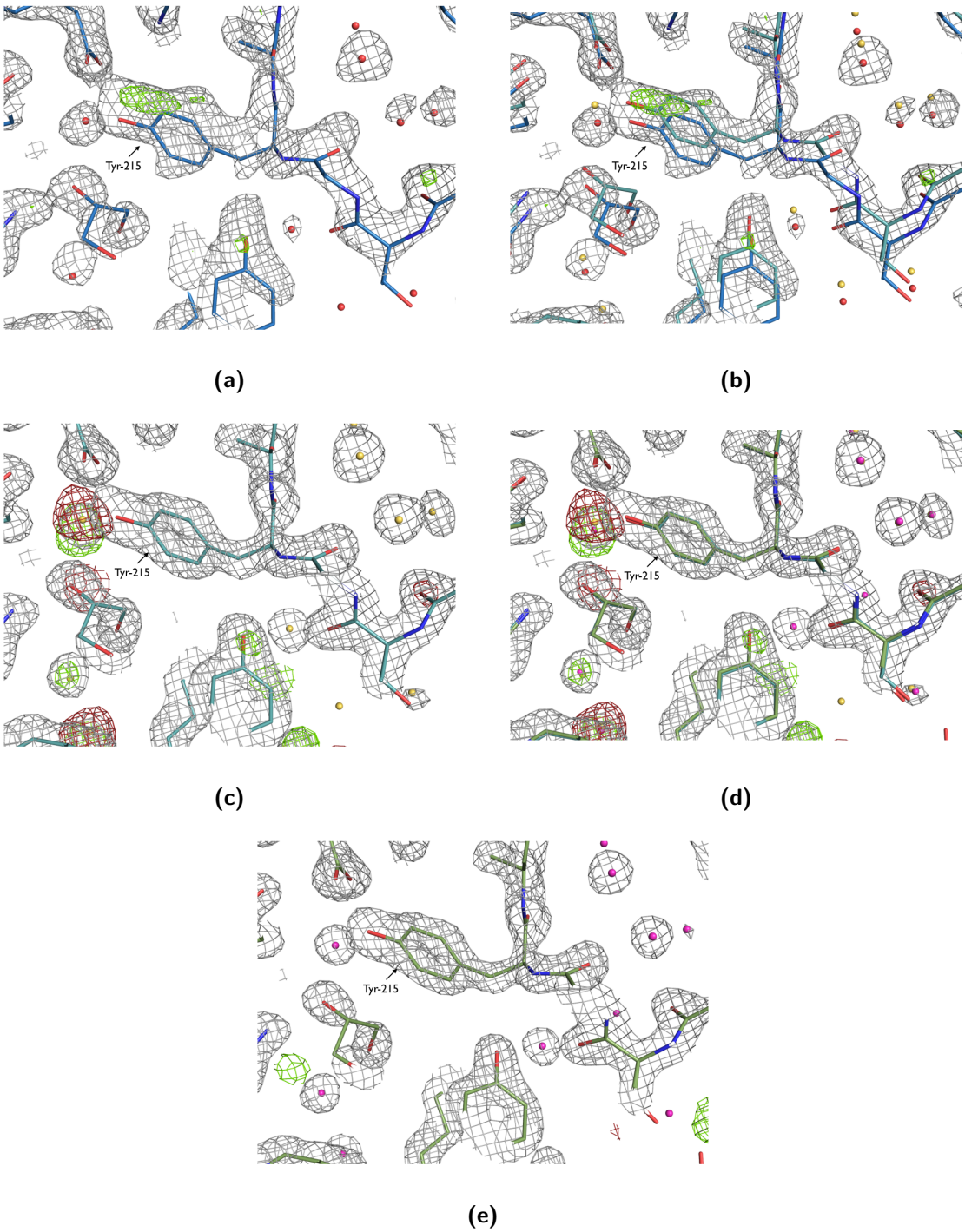


Figure 6.2: $2mF_{obs} - DF_{calc}$ electron density maps and $mF_{obs} - DF_{calc}$ maps for the *1dwa* myrosinase model at successive stages of the *PDB_REDO* re-refinement pipeline, centred at residue Tyr-215. All $2mF_{obs} - DF_{calc}$ and $mF_{obs} - DF_{calc}$ maps were generated over the asymmetric unit with *FFT*, and are contoured at 2σ (in grey) and $\pm 4\sigma$ (green/red) respectively. (a) The original *1dwa* model extracted directly from the PDB (with ordered water shown in red). In (b), the coordinate model after 10 cycles of rigid body refinement with *PDB_REDO* has been superimposed on the coordinate model and electron density maps in (a). In (c), the re-refined *1dwa* model (with ordered water shown in yellow) and electron density maps extracted from *PDB_REDO* following 10 cycles of rigid body refinement are shown. In (d), the coordinate model after full model re-building with *PDB_REDO* has been superimposed on the coordinate model and electron density maps in (c). In (e), the fully re-refined *1dwa* model (with ordered water shown in pink) and electron density maps extracted from *PDB_REDO* after full model re-building are shown. Full model re-building with *PDB_REDO* resulted in correct placement of both the Tyr-215 aromatic group and the ordered solvent molecules.

The following necessary parameters have been used to recalculate dose values. The crystal was quoted to have diameter of 0.1 mm and a length of 0.3 mm as aligned along the spindle axis. The 2.7 M ammonium sulphate component of the crystallisation buffer was included in the X-ray absorption calculation. Diffraction data were collected on beamline ID14-3 at the ESRF at 100 K, with a beam energy of 13.08 keV, and collimated beam size of 0.15 mm $\times$ 0.15 mm. The flux density was measured to be between $2.0 - 2.5 \times 10^{12}$ photons mm^{-2} s^{-1} using a windowless silicon PIN diode. Five datasets, each of 92 images, were collected over the same angular range, with an oscillation of 1° per image, and a per-image exposure time of $2.04 - 2.46$ s depending on the incident photon flux. Between each dataset, the crystal was additionally exposed to an X-ray burn of 8.9×10^{15} photons/mm^2 for datasets 1 to 3 and 26.7×10^{15} photons/mm^2 before the final dataset.

In *RADDOSE-3D*, the cumulative DWD at the end of data collection was $\sim$ 27.8 MGy. Burmeister (2000) had reported that the diffraction intensity (normalised to the first dataset) had decayed to below 50 % by dataset 5, and consequently, the recalculated DWD values presented here fall below the 43 MGy $D_{1/2}$ dose-limit determined by Owen *et al.* (2006).

6.2.4 Evaluation of recalculated $F_{obs,n} - F_{obs,1}$ maps

$F_{obs,n} - F_{obs,1}$ maps were generated using *FFT* (Ten Eyck, 1973) as part of the *RIDL* pipeline (as presented in Section 2.5). All $F_{obs,n} - F_{obs,1}$ maps were calculated using phases derived from the initial dataset (*1dwa*) coordinate model following full model re-building in *PDB_REDO*.

Consistent with the original findings of Burmeister (2000), interpretable negative difference density in $F_{obs,n} - F_{obs,1}$ maps located at the Tyr-330 –OH group was observed above a $\pm 5\sigma$ map threshold for the highest dose datasets ($n = 4$, 5) (Figure 6.3). An additional negative difference peak coincident with the Tyr-215 –OH group was also clear above a $\pm 5\sigma$ threshold for datasets $n = 4$, 5 (Figure 6.4a). This finding was not reported in the original study. It is suggested here that in that work the poor density fit for Tyr-215 (as discussed above) resulted in no difference peak being attributed to Tyr-215. Indeed, upon our re-inspection of the $F_{obs,5} - F_{obs,1}$ map from the original data extracted directly from the PDB, a nearby difference peak was revealed; however, it was not directly aligned with the Tyr-215 –OH group (Figure 6.4b). Tyr-215 and Tyr-330

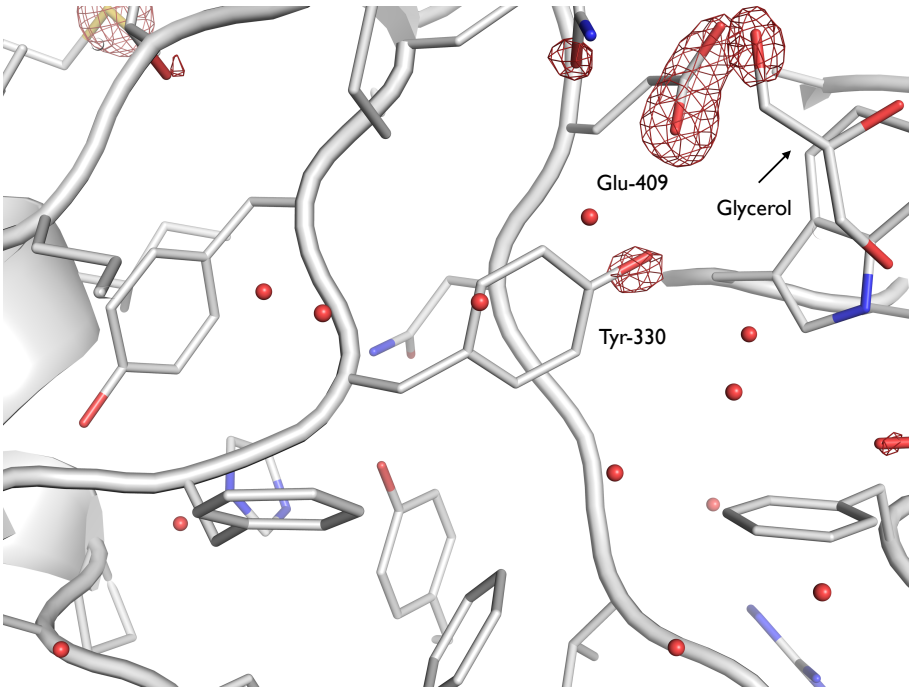


Figure 6.3: Reproduction of specific radiation damage to Tyr-330 and Glu-409 in the Burmeister (2000) myrosinase damage series. $F_{\text{obs},5} - F_{\text{obs},1}$ difference map, contoured at $\pm 5\sigma$ in green/red, overlaid on the initial coordinate model in white retrieved from *PDB_REDO* (PDB accession code: *1dwa*). Coordinate model and structure factor amplitudes were retrieved from *PDB_REDO*, and $F_{\text{obs},5} - F_{\text{obs},1}$ maps were generated by *FFT* as part of the *RIDL* pipeline. Electron density loss is observable around a glycerol oxygen, presumably also due to the dislocation of the glycerol upon decarboxylation of Glu-409.

both occupy similar hydrated pockets in myrosinase, with both being involved in hydrogen bond interactions (to Glu-153 and Glu-409, respectively), and both being within 5 Å from an ordered glycerol molecule. As such, the similar −OH electron density loss for Tyr-215 and Tyr-330 is perhaps not surprising.

Significant density loss peaks (apparent above a $\pm 4\sigma$ map threshold) were also identifiable in the highest dose $F_{\text{obs},5} - F_{\text{obs},1}$ map centred around several other Tyr −OH groups in myrosinase: Tyr-390 above $\pm 5\sigma$, and Tyr-400, Tyr-181 and Tyr-152 above $\pm 4.5\sigma$). Both the relative atomic *B*-factor increase and B_{Damage} atomic displacement metrics exhibited strong agreement with these observations, with Tyr-330, Tyr-215 and Tyr-390 flagged as most highly sensitive compared with other Tyr −OH groups (Figure 6.5a). For each atom, the B_{Damage} metric is derived from the isotropic *B*-factor, but deconvoluted from the atomic packing density (Gerstel *et al.*, 2015); it thus provides an indicator of increased side-chain mobility with dose. For the full set of myrosinase tyrosine O_η oxygen atoms, a strong positive correlation was also found between D_{loss} and change in B_{Damage} between the first and higher dose datasets (Figure 6.5b, linear $R^2 > 0.7$ for all higher

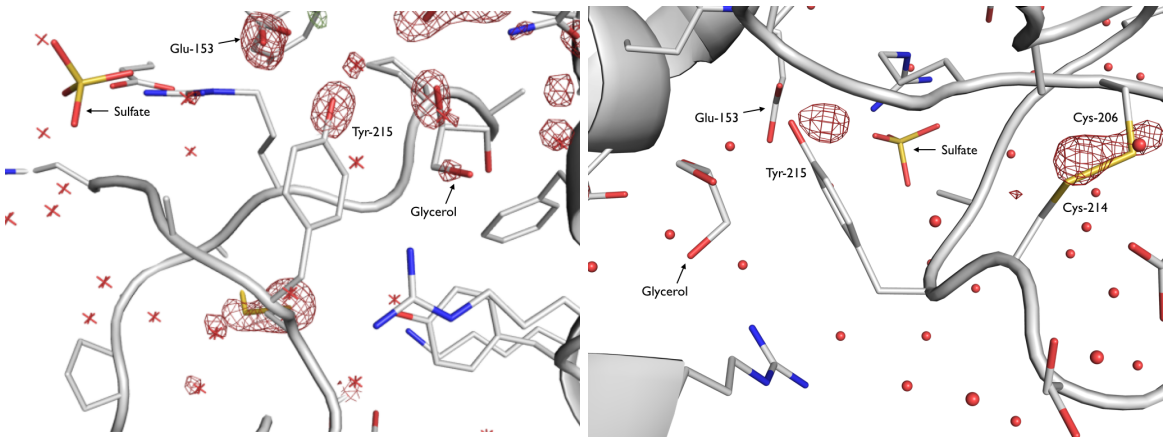
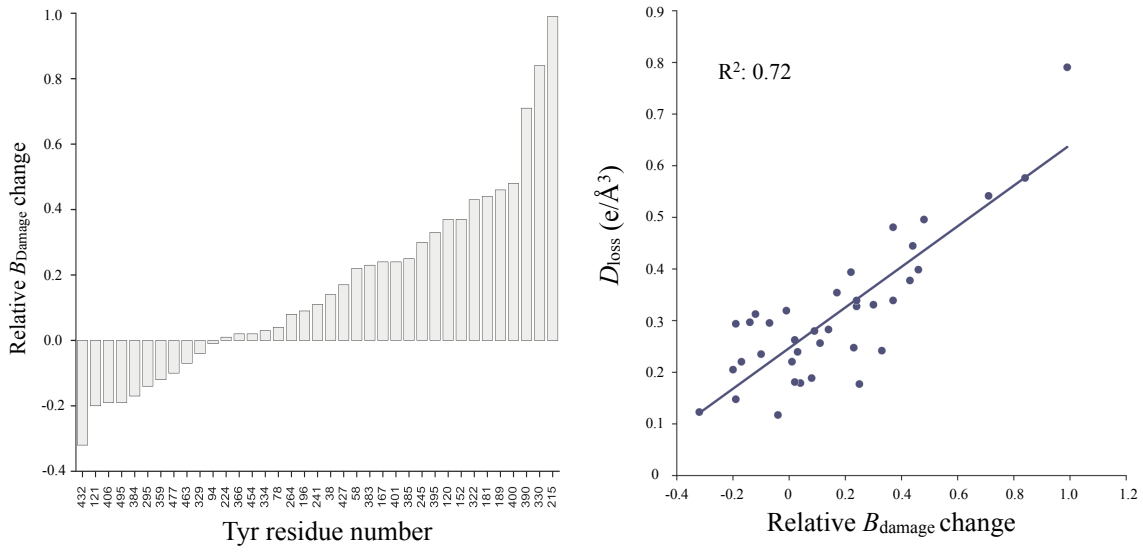


Figure 6.4: Specific radiation damage to Tyr-215 in the myrosinase damage series (Burmeister, 2000). In (a), the $F_{\text{obs},5} - F_{\text{obs},1}$ difference map, with structure factor amplitudes, $|\mathbf{F}_{\text{obs},1}|$, retrieved from *PDB_REDO*, is overlaid on the re-refined initial coordinate model in white. Phases were derived from the re-refined *1dwa* coordinate model. Clear negative difference density is now observable centred at the position of the Tyr-215 −OH group. In (b), $F_{\text{obs},5} - F_{\text{obs},1}$ difference map is overlaid on the original initial coordinate model *1dwa* retrieved from the PDB (shown in white). Structure factor amplitudes have been retrieved directly from the PDB for the difference map calculation, with phases derived from the original *1dwa* coordinate model. Negative difference density is observable adjacent to the Tyr-215, however does not align with the −OH group. In both cases, the $F_{\text{obs},5} - F_{\text{obs},1}$ maps have been generated by *FFT* within the *RIDL* pipeline, and are overlaid over the coordinate models at $\pm 5\sigma$ contouring levels in green/red.

dose datasets).

Figure 6.6 illustrates the behaviour of both the D_{loss} and density-weighted D^{-}_{mean} metrics calculated by the *RIDL* pipeline (Section 2.5) for the top three Tyr −OH damage sites as reported in the results shown in Figure 6.5. Unsurprisingly, metric values do not agree well when calculated from both data retrieved from *PDB_REDO* following model rebuilding, and also for data retrieved directly from the PDB. However Tyr-215 is consistently flagged as the most radiation-susceptible Tyr group in myrosinase. For data extracted from the PDB, the poor placement of Tyr-390 appears to have led to density loss that is indistinguishable from average metric values attained over all refined atoms in myrosinase.

Overall, it can thus be concluded that although *PDB_REDO* provided a significant improvement in $R_{\text{work}}/R_{\text{free}}$ statistics for the original lowest-dose myrosinase structure (*1dwa*), the Tyr −OH damage signatures reported by Burmeister (2000) could still be replicated, and furthermore, additional sites of damage could be detected (e.g. Tyr-390).



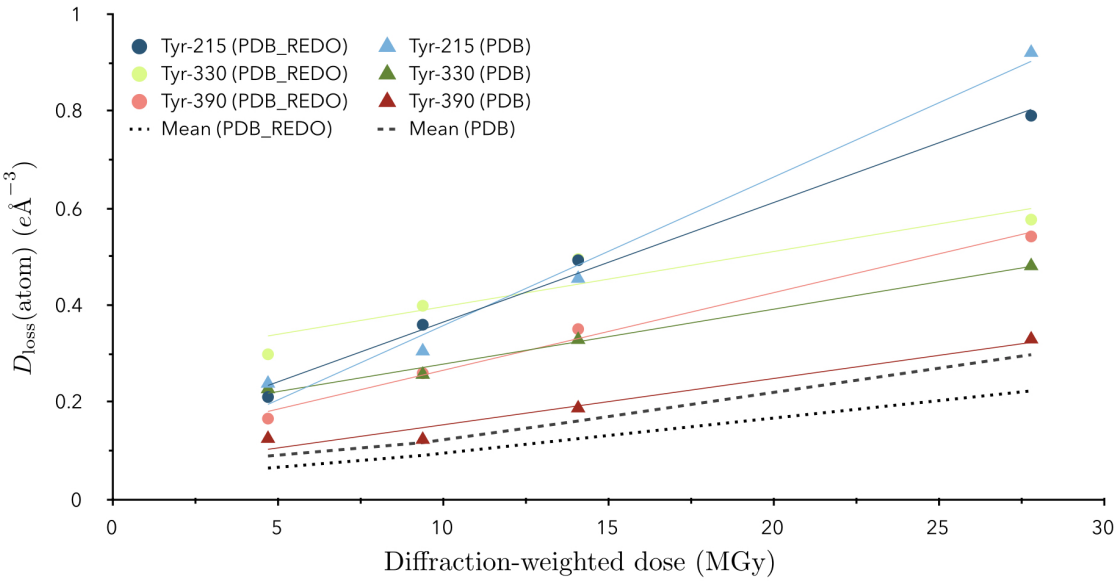
(a) **(b)**

Figure 6.5: (a) B_{Damage} change for each Tyr O_η atom in the highest dose myrosinase structure (PDB accession code: *1dwi*) relative to the lowest-dose *1dwa* model, with all coordinate models retrieved from *PDB_REDO*. (b) Relationship between D_{loss}(atom) and change in B_{Damage} relative to the *1dwa* structure, for each Tyr −OH atom at the highest dose $n = 5$ myrosinase dataset. For each atom a, the relative change is calculated as $[B_{\text{Damage}}(a^{1dwi}) - B_{\text{Damage}}(a^{1dwa})]/B_{\text{Damage}}(a^{1dwa})$. To account for the dependence of B_{Damage} on non-unity atomic occupancies in the originally deposited data, all occupancies in the *1dwa* and *1dwi* coordinate models were set to 1 and a further round of isotropic *B*-factor refinement was performed in *phenix.refine* prior to calculating B_{Damage}. Linear regression has been performed using the *SciPy* package in *Python*.

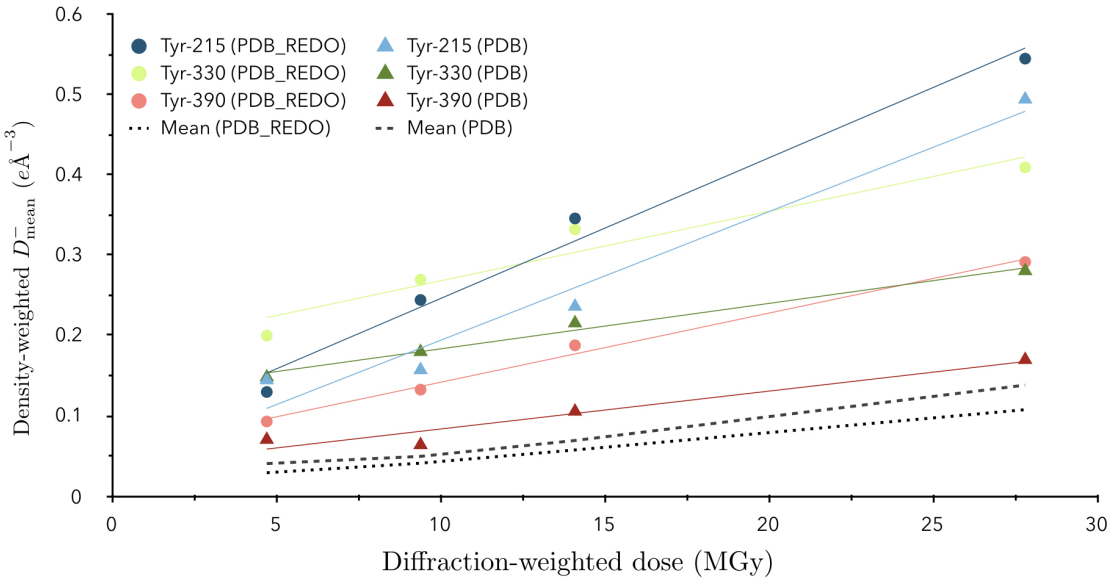
6.3 Investigation of radiation-induced trends between proteins

To better characterise the reproducibility of Tyr −OH group electron density loss in irradiated protein crystals, this section will present a systematic investigation on a wide range of MX structures at 100 K. A subset of previously conducted MX radiation damage investigations have included deposition of readily accessible coordinate and structure factor information in the PDB, providing an ideal database of protein damage series. Each damage series has been refined independently by different research groups, mitigating the risk of systematic and subjective errors in the refinements between different structures. Moreover, the deposited damage series have all *a priori* satisfied the PDB validation criteria, and can thus be deemed of suitable quality for the current study.

Located proteins were also of varying size (5.7−204 kDa), each with unique nuances of buffer composition (potential free radical scavengers) and macromolecular composition (monomeric versus multimeric protein, presence of nucleic acid, glycans and metal cofactors). These factors allowed a range of experimental conditions to be sampled.



(a)



(b)

Figure 6.6: The D_{loss} and density-weighted D^{-}_{mean} metrics for the top three Tyr −OH damage sites in myrosinase, as indicated by B_{Damage} analysis (Figure 6.5). Metrics have been calculated using the *RIDL* pipeline (introduced in Section 2.5), and have been calculated for data retrieved from *PDB_REDO* following full model rebuilding and re-refinement (denoted '*PDB_REDO*' in each legend), and also retrieved directly from the PDB (denoted 'PDB' in each legend above). The mean metric values attained across all refined atoms are also shown at each dataset.

6.3.1 Acquisition of MX damage series data from *PDB_REDO*

To the best of the author's knowledge, a complete list of MX damage series with both coordinate models and structure factor files deposited in the PDB was collated. Of these, a selection was made using criteria designed to ensure a suitable sample set for the current analysis:

- Only MX experiments definitely conducted at 100 K were included
- Studies with reported accumulated doses < 1 MGy at highest dose dataset were omitted
- Intermediate datasets with reported doses < 1 MGy in the remaining series were excluded

A 1 MGy lower dose limit was imposed in the current analysis to isolate datasets for which Tyr damage events may be expected. Whereas disulphide radicalisation and cleavage has been observed in non-metalloproteins below 1 MGy, Glu/Asp decarboxylation has not (Sutton *et al.*, 2013). To justify the chosen 1 MGy lower bound, the complete Sutton *et al.* (2013) lysozyme damage series (reported dose range $0.07 - 1.05$ MGy) was additionally included as a control for comparison. In Section 6.3.2.2, Figure 6.8m confirms that no significant damage to Glu, Asp and Tyr was detected for lysozyme below 1 MGy.

Table 6.2 provides a full list of damage series located in the PDB that were used in the current study.

For each MX damage dataset series, PDB-format coordinate model and MTZ-format structure factor files were retrieved from the *PDB_REDO* online server, and files were retrieved after initial *PDB_REDO* file curation, but prior to any further model rebuilding. At this stage, coordinate models had been subjected to either 0 or 10 cycles of rigid body refinement with *REFMAC*, dependent on whether *PDB_REDO* flagged possible R_{free} bias in the PDB-deposited data. This procedure was chosen to extract data as close to that used by the authors of the associated radiation damage studies as possible, whilst ensuring that all data were in a consistent format. Accordingly, for the initial structure in each damage series, with the exception of myrosinase, the recalculated R_{work} statistics were all reported by *PDB_REDO* to be within $2\,\%$ of those reported in the originally deposited coordinate file header (and for structures where there was no *PDB_REDO* flagged potential R_{free} bias, R_{free} was also within $1\,\%$ of the original reported values);

Table 6.2: Selection of MX damage series (all conducted at 100 K) previously deposited in the PDB, complemented with one unpublished insulin damage series supplied by Dr. Jonathan Brooks-Bartlett. For series in which intermediate datasets fall below the 1 MGy selection criterion, PDB accession codes are only provided for datasets included in the current analysis. (Table continues on next page)

Protein	Host species	PDB accession codes	Resolution of first dataset (Å)	Quoted total molecular weight (kDa)	Unique Tyr residues in asymmetric unit	Associated publication	Dose calculation program
Acetylcholinesterase	*Torpedo californica*	*1qid*, *1qie*, *1qif*, *1qig*, *1qih*, *1qii*, *1qij*, *1qik*, *1qim*	2.05	60.7	17	Weik *et al.* (2000)	manual
Thermolysin	*Bacillus thermoproteolyticus*	*3p7p*, *3p7q*, *3p7r*, *3p7s*, *3p7t*, *3p7u*, *3p7v*, *3p7w*	2.2	34.6	28	Juers and Weik (2011)	*RADDOSE*
Myrosinase	*Sinapis alba*	*1dwa*, *1dwf*, *1dwg*, *1dwh*, *1dwi*	2.0	62.4	37	Burmeister (2000)	manual
Malate dehydrogenase	*Haloarcula marismortui*	*2j5k*, *2j5q*, *2j5r*	2.0	131.6	32	Fioravanti *et al.* (2007)	*RADDOSE*
Chicken egg white lysozyme	*Gallus gallus*	*2ybh*, *2ybi*, *2ybj*, *2ybl*, *2ybm*, *2ybn*	2.0	14.6	3	De la Mora *et al.* (2011)	*RADDOSE v2*
Chicken egg white lysozyme	*Gallus gallus*	*4h8x*, *4h8y*, *4h8z*, *4h90*, *4h91*, *4h92*, *4h93*, *4h94*, *4h9a*, *4h9b*, *4h9c*, *4h9e*, *4h9f*, *4h9h*, *4h9i*	1.2	14.5	3	Sutton *et al.* (2013)	*RADDOSE v2*
Pancreatic elastase	*Sus scrofa*	*3mnb*, *3mnc*, *3mns*, *3mnx*, *3mo3*, *3mo6*, *3mo9*, *3moc*	1.2	26.1	11	Petrova *et al.* (2010)	*RADDOSE*
Pancreatic elastase	*Sus scrofa*	*2blo*, *2blq*	1.33	26.2	11	Nanao *et al.* (2005)	*RADDOSE*

Table 6.2: (cont.) Selection of MX damage series (all conducted at 100 K) previously deposited in the PDB, complemented with one unpublished insulin damage series supplied by Dr. Jonathan Brooks-Bartlett. For series in which intermediate datasets fall below the 1 MGy selection criterion, PDB accession codes are only provided for datasets included in the current analysis.

Protein	Host species	PDB accession codes	Resolution of first dataset (Å)	Quoted total molecular weight (kDa)	Unique Tyr residues in asymmetric unit	Associated publication	Dose calculation program
Insulin	*Bos taurus*	*2bn3*, *2bn1*	1.4	5.8	4	Nanao *et al.* (2005)	*RADDOSE*
Lysozyme	*Gallus gallus*	*2blx*, *2bly*	1.4	14.7	3	Nanao *et al.* (2005)	*RADDOSE*
Ribonuclease A	*Bos taurus*	*2blp*, *2blz*	1.4	13.7	6	Nanao *et al.* (2005)	*RADDOSE*
Thaumatin	*Thaumatococcus daniellii*	*2blr*, *2blu*	1.4	22.3	8	Nanao *et al.* (2005)	*RADDOSE*
Trypsin	*Bos taurus*	*2blv*, *2blw*	1.2	24.0	10	Nanao *et al.* (2005)	*RADDOSE*
C.Esp1396I-DNA complex	*Escherichia coli*	*4x4b*, *4x4c*, *4x4d*, *4x4e*, *4x4f*, *4x4g*, *4x4h*, *4x4i*	2.8	59.7	8	Bury *et al.* (2015)	*RADDOSE-3D*
Trp RNA-binding attenuation protein (TRAP) -RNA complex	*Escherichia coli*	*5eeu*, *5eev*, *5eew*, *5eex*, *5eey*, *5eez*, *5ef0*, *5ef1*, *5ef2*, *5ef3*	1.98	204.1	22	Bury *et al.* (2016b)	*RADDOSE-3D*
Phosphoserine aminotransferase	*Bacillus alcalophilus*	*2bhx*, *2bi5*, *2bi9*, *2bia*	1.68	81.8	28	Dubnovitsky *et al.* (2005)	*RADDOSE*
GH7 family cellobiohydrolase	*Daphnia pulex*	*5mcc*, *5mcd*, *5mce*, *5mcf*, *5mch*, *5mci*, *5mcj*, *5mck*, *5mcl*, *5mcm*, *5mcn*	2.0	96.6	34	Bury *et al.* (2017)	*RADDOSE-3D*
Insulin	*Bos taurus*	N/A	1.38	5.8	4	Unpublished	*RADDOSE-3D*

see Table 6.3 for details. For the myrosinase damage series, the coordinate model following full model rebuilding with *PDB_REDO* was used in the current analysis, since as discussed in Section 6.2, the 2mF_{obs} − DF_{calc} map-model fit was poor for the original PDB-deposited *1dwa* model.

For each coordinate model, further rigid body refinement was performed in *REFMAC* for each higher dose dataset, now using the coordinate model from the initial dataset coupled with the structure factor amplitudes, $|\mathbf{F}_{obs,n}|$, corresponding to the higher dose dataset. This ensured that an identical set of atoms was modelled for each dataset in a damage series, whilst accounting for whole molecule translation and rotation within the unit cell, as a result of any changes in unit-cell dimensions with increasing dose. The resulting coordinate models were used throughout the analysis presented here.

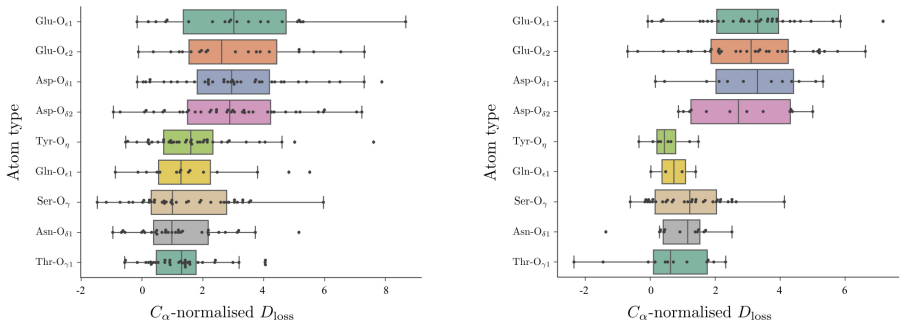
6.3.2 Damage comparison strategies between deposited structures

$F_{obs,n} - F_{obs,1}$ maps were generated using *FFT* (Ten Eyck, 1973) as part of the *RIDL* pipeline (as presented in Section 2.5). Here, $F_{obs,n} - F_{obs,1}$ maps were calculated from the differences of reflection amplitudes between the first and each successive dataset and with a fixed phase set, $\phi_{calc,1}$, derived from model refinement against the first dataset. Manual inspection of maps was first performed in *COOT* to ensure that key specific damage manifestations (e.g. clear negative difference density surrounding disulphide bonds) were consistent with reports in the original publications pertaining to the particular damage series.

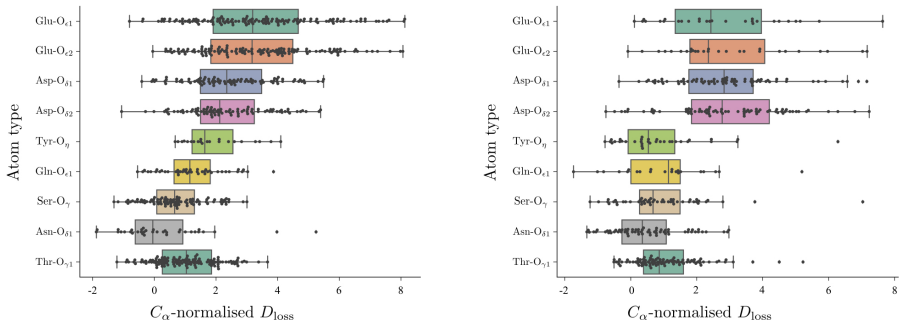
For the analysis of multiple MX damage series across distinct protein structures, a systematic visual interpretation of $F_{obs,n} - F_{obs,1}$ maps would be impractical, with a high risk of subjective bias by the investigators. The D_{loss}(atom) metric provided an objective dose-dependent atom-by-atom quantification of electron density loss. D_{loss} is derived from $F_{obs,n} - F_{obs,1}$ difference maps at each subsequent accumulated dose (n = 2, 3,...) as the maximum density loss value in the local vicinity of each refined atom. D_{loss} thus provides a suitable indicator to screen for electron density loss proximal to the Tyr −OH group across deposited structures.

Table 6.3: R_{work} and R_{free} statistics corresponding to the lowest dose structures for each MX damage series located in the PDB and *PDB_REDO*. Values are as reported in original coordinate file header retrieved from the PDB, and from *PDB_REDO* following *REFMAC* rigid body refinement cycles in *PDB_REDO* (either 0 or 10 rounds depending on whether potential R_{free} bias was flagged by *PDB_REDO*). The percentage difference in R_{work} and R_{free} between the two sources is also provided, as are any warning flags raised by *PDB_REDO* during data handling.

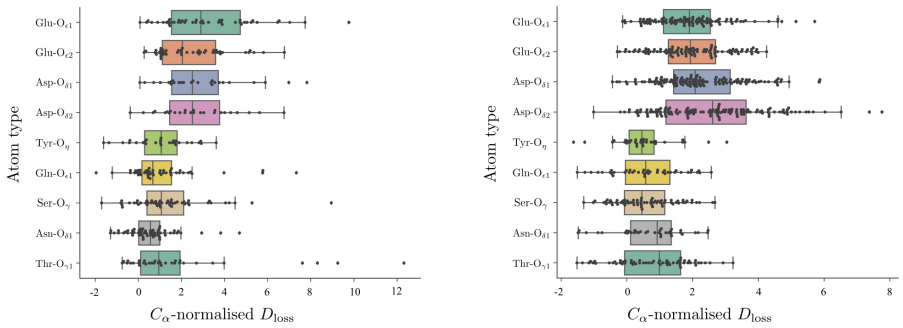
Publication	Protein	PDB code	Original PDB R_{work}/R_{free}	*PDB_REDO* (rigid body refinement): R_{work}/R_{free}	% difference in R_{work}/R_{free}	*PDB_REDO* notes
Weik *et al.* (2000)	Acetylcholinesterase	*1qid*	0.22/0.23	0.21/0.21	-0.49/-1.84	Possible R_{free} bias flagged
Juers and Weik (2011)	Thermolysin	*3p7p*	0.15/0.21	0.15/0.20	0.15/-0.76	-
Burmeister (2000)	Myrosinase	*1dwa*	0.17/0.18	0.19/0.20	2.6/2.23	Possible R_{free} bias flagged
Fioravanti *et al.* (2007)	Malate dehydrogenase	*2j5k*	0.23/0.27	0.22/0.22	-0.09/-4.86	New R_{free} set (5 %)
De la Mora *et al.* (2011)	Lysozyme	*2ybh*	0.20/0.24	0.20/0.21	-0.32/-3.07	New R_{free} set (10 %)
Sutton *et al.* (2013)	Lysozyme	*4h8x*	0.19/0.20	0.18/0.19	-0.76/-0.82	-
Petrova *et al.* (2010)	Elastase	*3mnb*	0.11/0.14	0.11/0.14	0.07/-0.09	-
Nanao *et al.* (2005)	Elastase	*2blo*	0.11/0.14	0.11/0.14	0.06/-0.04	-
Nanao *et al.* (2005)	Insulin	*2bn3*	0.13/0.16	0.13/0.16	-0.05/-0.09	-
Nanao *et al.* (2005)	Lysozyme	*2blx*	0.13/0.17	0.13/0.17	0.09/-0.11	-
Nanao *et al.* (2005)	Ribonuclease A	*2blp*	0.15/0.18	0.15/0.17	0.09/-0.24	-
Nanao *et al.* (2005)	Thaumatin	*2blr*	0.13/0.16	0.12/0.15	-0.39/-0.22	-
Nanao *et al.* (2005)	Trypsin	*2blv*	0.11/0.13	0.11/0.13	0.09/0.23	-
Bury *et al.* (2015)	C.Esp1396I-DNA	*4x4b*	0.21/0.26	0.22/0.27	1.85/0.89	-
Bury *et al.* (2016b)	*Trp* RNA-binding Attenuation Protein	*5eeu*	0.19/0.23	0.19/0.23	-0.01/-0.28	-
Dubnovitsky *et al.* (2005)	Phosphoserine aminotransferase	*2bhx*	0.18/0.23	0.17/0.18	-0.97/-5.07	Possible R_{free} bias flagged
Bury *et al.* (2017)	Glycoside hydrolase family 7 protein	*5mcc*	0.15/0.17	0.16/0.18	0.04/0.06	-



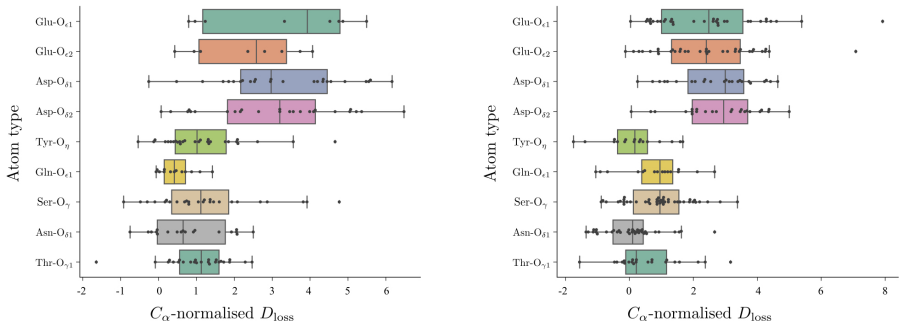
(a) Myrosinase **(b)** C.Esp1396I-DNA



(c) TRAP-RNA **(d)** GH7 family protein



(e) Phosphoserine aminotransferase **(f)** Malate dehydrogenase



(g) Thermolysin **(h)** Acetylcholinesterase

Figure 6.7: The distribution of C_α-normalised D_{loss}(atom) values for side-chain oxygen atoms in MX damage series retrieved from *PDB_REDO* prior to model rebuilding. C_α-normalisation was performed as discussed in Section 2.5.8, such that in each plot $D_{\text{loss}}^{C_\alpha} = 0$ corresponds to the average metric value attained by the protein backbone. Metric values have been provided for only the highest dose $F_{\text{obs,n}} - F_{\text{obs,1}}$ map that could be calculated for each independent damage series. Only a subset of protein damage series have been shown at a lower bound mass cut off of 30 kDa to ensure the presence of enough data to plot suitable box plots. Tukey boxplots have been generated using the *Seaborn* package in *Python* such that the whiskers represent 1.5× the interquartile range. In addition, raw data are overlaid over each boxplot as points.

6.3.2.1 Challenges of comparison between independent damage series

Due to variability in experimental procedures and refinement strategies amongst damage series, rigorous comparison of damage between proteins was non-trivial. As an example, due to the maximum attained resolution differing between each protein system (from 1.2 Å to 2.8 Å), proteins had undergone non-uniform refinement protocols prior to retrieval from *PDB_REDO* depending on the optimum model refinement strategy (e.g. NCS if appropriate, *B*-factor restraints) at a particular resolution. Consequently, a *B*-factor-derived metric, such as the relative atomic *B*-factor increase (Weik *et al.*, 2000) or B_{Damage} (Gerstel *et al.*, 2015) was deemed to be sub-optimal for comparing damage series. Instead, a metric derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps (e.g. D_{loss}) instead depends more directly on experimentally-determined diffraction intensities, thus reducing dependence on refined models[†], and was deemed more appropriate.

Figure 6.7 illustrates the distribution of D_{loss} metric values for side-chain oxygen atoms in independent protein damage series. Only oxygen atoms in the structures have been compared with Tyr O_η oxygen, to maintain a constant total number of electrons per compared atom. In a protein at a specific dose, neglecting variations in intrinsic disorder between different side-chains, two oxygen atoms O_A and O_B should exhibit identical D_{loss}(atom) values if the yields of O_A and O_B damage events[‡] throughout the protein crystal are the same. To aid comparison between proteins, for each protein, D_{loss} values have also been normalised to the average of D_{loss} values attained by the set of protein C_α backbone atoms, as discussed in Section 2.5.8. Thus, for each protein, $D_{\text{loss}}^{C_\alpha} = 0$ corresponds to electron density loss comparable to backbone of that protein, and positive values indicate radiation-sensitivity. For each protein, only D_{loss} metric values corresponding to the highest dose $F_{\text{obs,n}} - F_{\text{obs,1}}$ map have been plotted, however since the accumulated dose reported for each protein differs, the numerical $D_{\text{loss}}^{C_\alpha}$ metric values should not strictly be compared between proteins.

The challenge of rigorous comparison between proteins is further compounded by uncertainties in the validity of quoted dose values. In the publications relating to each protein radiation damage

[†]As has been remarked previously in this thesis, of course the phases used to calculate each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map depend directly on the refined coordinate model and thus the chosen refinement strategy. However, as has been demonstrated previously in Section 2.5.9.3 (see Figure 2.33), the results from *RIDL* appear to be relatively insensitive to differing refinement strategies employed on later datasets.

[‡]Here a damage event is considered to be either radiation-induced disordering or cleavage.

series, dose values have been independently calculated using a variety of tools [e.g. *RADDOSE v1* (Murray *et al.*, 2004), *v2* (Paithankar *et al.*, 2009) and *v3* (Paithankar and Garman, 2010), and *RADDOSE-3D* (Zeldin *et al.*, 2013b), or unspecified]. Moreover, differing dose metrics (e.g. diffraction-weighted dose, maximum dose, average dose) have been quoted, depending on the particular program used. As such, direct comparison of damage events between protein crystals *reported* to be at identical doses was deemed to be highly sensitive to miscalibration between dose scales.

An attempt to recalculate all dose values, using *RADDOSE-3D* and the DWD metric, was unsuccessful due to the inadequate data collection strategy information reported in numerous original publications. For an example where most experimental conditions were documented, the Fioravanti *et al.* (2007) study did not provide explicit details of the beam FWHM at ESRF beamline ID23-EH1. In the current investigation, *RADDOSE-3D* was run iteratively with FWHM varied between 30 µm × 100 µm (using other parameters specified in that publication), resulting in a high variation in the computed DWD values between 1 MGy and 7.5 MGy for a 180° wedge of data corresponding to a single dataset. In the original publication, only a 'maximum dose' of 1.2 MGy had been reported for one dataset, as calculated by *RADDOSE*.

Due to these difficulties, attempts to recalculate doses for each damage series are not presented here, and all doses are as reported in the original publications. To account for the non-uniformity in dose metrics used, the following damage analysis strictly refrains from direct comparison between proteins at any quoted dose value.

6.3.2.2 A discrete damage ranking scheme for proteins

A discrete ranking scheme has been implemented to mitigate uncertainties when comparing density map-derived metric values between independent protein damage series, as follows:

- For a damage series consisting of a single low- and high- dose dataset, D_{loss} values correspond to the single $F_{\text{obs},2} - F_{\text{obs},1}$ map, and each atom type A (e.g. the set of Tyr O_η oxygen atoms) is ranked with respect to the mean D_{loss} value attained by that atom type relative to other atom types in that protein (rank: $R_A \geq 0$). The atom type with the

highest overall mean D_{loss} metric attains the rank $R_A = 0$.

- For a damage series containing multiple higher dose datasets, atom type ranks were first determined per-dataset ($R_{A,i}$ for $i = 2, ..., n$) and the products of these per-dataset ranks ($R_{A,2} \times R_{A,3} \times ... \times R_{A,n}$) were then ranked to provide an across-dataset rank for each atom type in that protein.

Figure 6.8 illustrates the resulting ranks of radiation-induced electron density loss for different side-chain oxygen moieties in the damage series investigated. In the case of Glu and Asp, a single damage ranking value has been computed as a average across both oxygens in each side-group, since in general the labelling of the $O_{\epsilon 1}$ and $O_{\epsilon 2}$ of oxygens in Glu (and similarly for Asp) is arbitrary and has no physical meaning. In close accord with the previous radiation damage literature, clear Glu and Asp decarboxylation was indicated by uniformly high ranks for Glu and Asp oxygens. As indicated in Figure 6.8d for phosphoserine aminotransferase, Glu and Asp are ranked as less susceptible than typically observed, with the highest rankings instead attained by atoms comprising the lysine-pyridoxal-50-phosphate Schiff base present in the protein. The Schiff base active site was reported in the original publication (Dubnovitsky *et al.*, 2005) to undergo severe radiation-induced relaxation of chemical strain, with pronounced reorientation of the Schiff base cofactor. In Figure 6.8m, the poor rankings (> 73) for Glu and Asp in lysozyme corresponding to the Sutton *et al.* (2013) study are strongly consistent with the low maximum dose (1.05 MGy) reported for that damage series. Whereas disulphide bond breakage was reported in that study, the dose is below the onset of detectable Glu/Asp decarboxylation.

In contrast, the Tyr O_η atom was consistently ranked lower, and of a similar order as seen in the control group of Gln, Asn, Ser and Thr O atoms for each protein structure, which are not suspected to be cleaved. Indeed in 14/18 damage series, at least one out of the control group has a rank higher than that of Tyr O_η. In particular, in Figure 6.8a for the myrosinase series (Burmeister, 2000), Gln-$O_{\epsilon 1}$ receives a higher ranking than Tyr O_η. Interestingly, the most radiation-susceptible Gln in myrosinase (Gln-146) is also situated in the active site occupied by Tyr-215, with the Gln $N_{\epsilon 2}$ atom at a distance of 6.2 Å from the Tyr O_η atom and only 2.9 Å from the central oxygen of the ordered glycerol occupying the binding site. No covalent bond cleavage is anticipated for Gln, and it is anticipated that the decarboxylation of Glu-153 and Asp-150 located in this active site (as well as reorientation of the glycerol) account for the disordering of

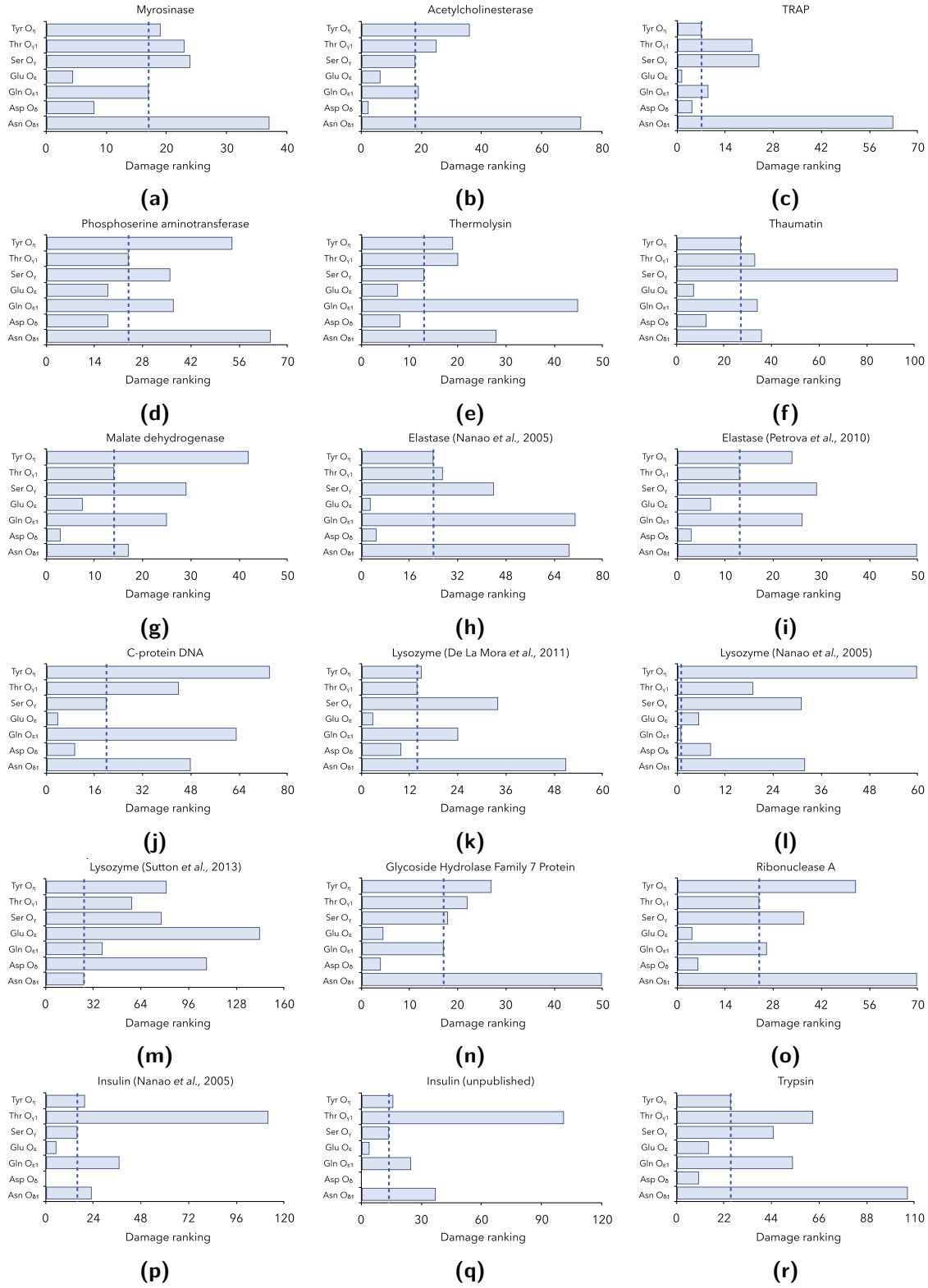


Figure 6.8: Electron density loss rankings for distinct side-chain oxygen atom types occurring in each protein structure as presented in Table 6.2. For each protein, the ranking scheme starts at 0 for the most damaged atom species, thus a large ranking number indicates a low D_{loss}(atom) value. For Glu and Asp, a single damage ranking value has been reported as the average over both oxygens in each side-group (denoted as Glu-O_ϵ and Asp-O_δ, respectively). For protein types corresponding to multiple publications, the publication names have been provided. In each plot, the dashed red line indicates the lowest rank assigned to a non Glu or Asp atom for that protein: only for TRAP (Bury *et al.*, 2017), thaumatin (Nanao *et al.*, 2005), elastase (Nanao *et al.*, 2005) and trypsin (Nanao *et al.*, 2005) is Tyr −OH ranked highest relative to the Gln, Asn, Thr and Ser atoms included as controls.

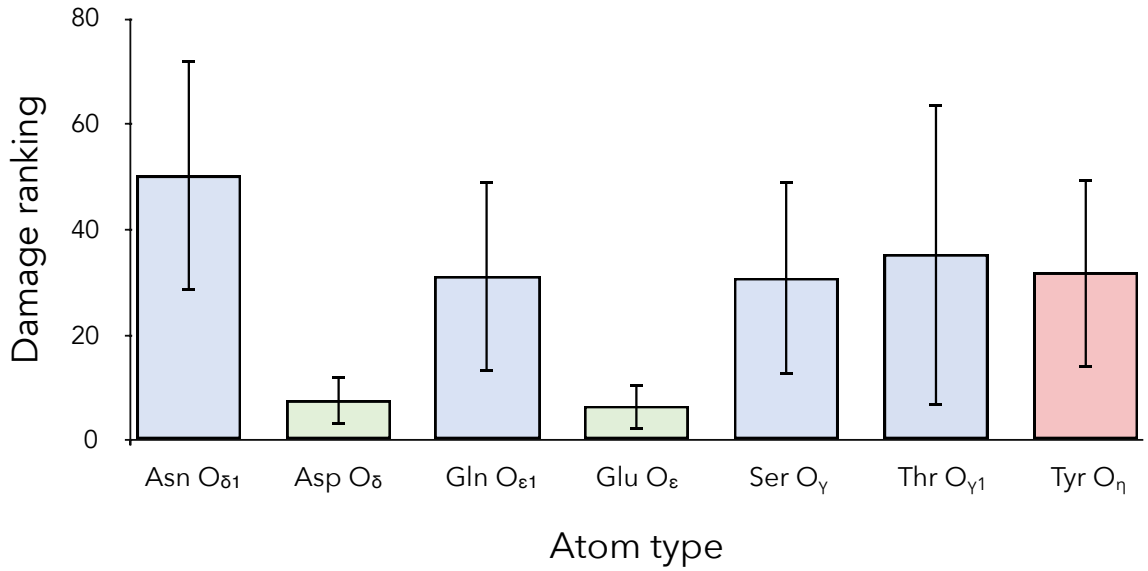


Figure 6.9: Combined electron density loss rankings for distinct side-chain oxygen atom types occurring for protein structures as presented in Table 6.2. For each atom type, each bar represents the average ranking achieved for proteins shown in Figure 6.8. Again for Glu and Asp, a single damage ranking value has been reported as the average over both oxygens in each side-group (denoted as Glu-O_ϵ and Asp-O_δ, respectively). Error bars indicate the standard deviation for each atom type. The lysozyme damage series corresponding to the Sutton *et al.* (2013) study has been omitted from the calculation, since the maximum dose reported for that study (1.05 MGy) was deemed to be too low for any detectable Glu/Asp decarboxylation events. Similar to Figure 6.2, the ranking scheme starts at 0 for the most damaged atom species, thus a large ranking number indicates a low D_{loss}(atom) value.

the Gln side-group. In Section 6.4, it will argued that this justification can also be used to explain electron density loss to the Tyr −OH group.

6.3.2.3 Overall ranking trends across proteins

The data in Figure 6.9 indicate a trend in damage rank per atom type across protein structures. Overall, despite unavoidable variability between the included protein systems (e.g. differing maximum resolution, crystallisation buffers etc.), the implemented D_{loss} ranking scheme was capable of detecting specific damage trends throughout the investigated proteins. Irrespective of the feasibility of a radiation-induced Tyr −OH damage event for crystalline proteins at 100 K, Tyr −OH electron density loss was in general ranked significantly lower than for Glu and Asp carboxylates for doses > 1 MGy, at which decarboxylation events were generally detected.

However, individual sites of pronounced Tyr −OH electron density loss have been detected as

shown in Figure 6.7, such as Tyr-215 and Tyr-330 in myrosinase, symmetry-related Tyr-62 residues in TRAP, Tyr-153 in the glycoside hydrolase family 7 (GH7) protein investigation described in Section 2.3, and Tyr-211 in thermolysin (Juers and Weik, 2011). The next section will investigate the mechanism behind these observations.

6.4 A hypothesis for the observed Tyr electron density loss

The aim of this section will be demonstrate that published reports of negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference density at a small subset of Tyr residues are consistent with a model of radiation-induced reorientation of the full Tyr side-group, without the necessity of invoking bond cleavage.

6.4.1 Comparison of damage signatures between residues

Evidence will first be presented to suggest that the Tyr and Phe aromatic rings undergo comparable positional displacement with increasing dose.

The distribution of D_{loss} values (so-called damage signatures) for topologically similar residue types have been directly compared. These are shown in Figure 6.10 for the highest dose datasets for myrosinase, *Trp* RNA-binding attenuation protein (Section 4.2), malate dehydrogenase (Fioravanti *et al.*, 2007) and the glycoside hydrolase family 7 protein (Section 2.3), all of which contain a statistically sufficient sample size of atoms for histogram analysis (each > 62 kDa). For each residue, the damage signature plot contains D_{loss} information for all non-hydrogen atoms of that residue at a specific dose. In Figure 6.10, at each dose, D_{loss} values have been normalised to the average D_{loss} value attained at that particular dose by the set of all C_α backbone atoms in the structure. As a result, signatures were observed to be typically unimodal distributions, with undamaged atoms attaining near zero values and right-tails indicating the presence of radiation-sensitive atom types in a residue.

For each protein, Asp and Asn exhibited qualitatively distinct behaviour, with the extended right-hand tail of the Asp distribution indicative of radiation-induced carboxylate loss, in contrast to the radiation-insensitive amine functional group of Asn (Figure 6.10 a, e, i and m). The same trend is true for Glu and Gln (Figure 6.10 c, g, k and o).

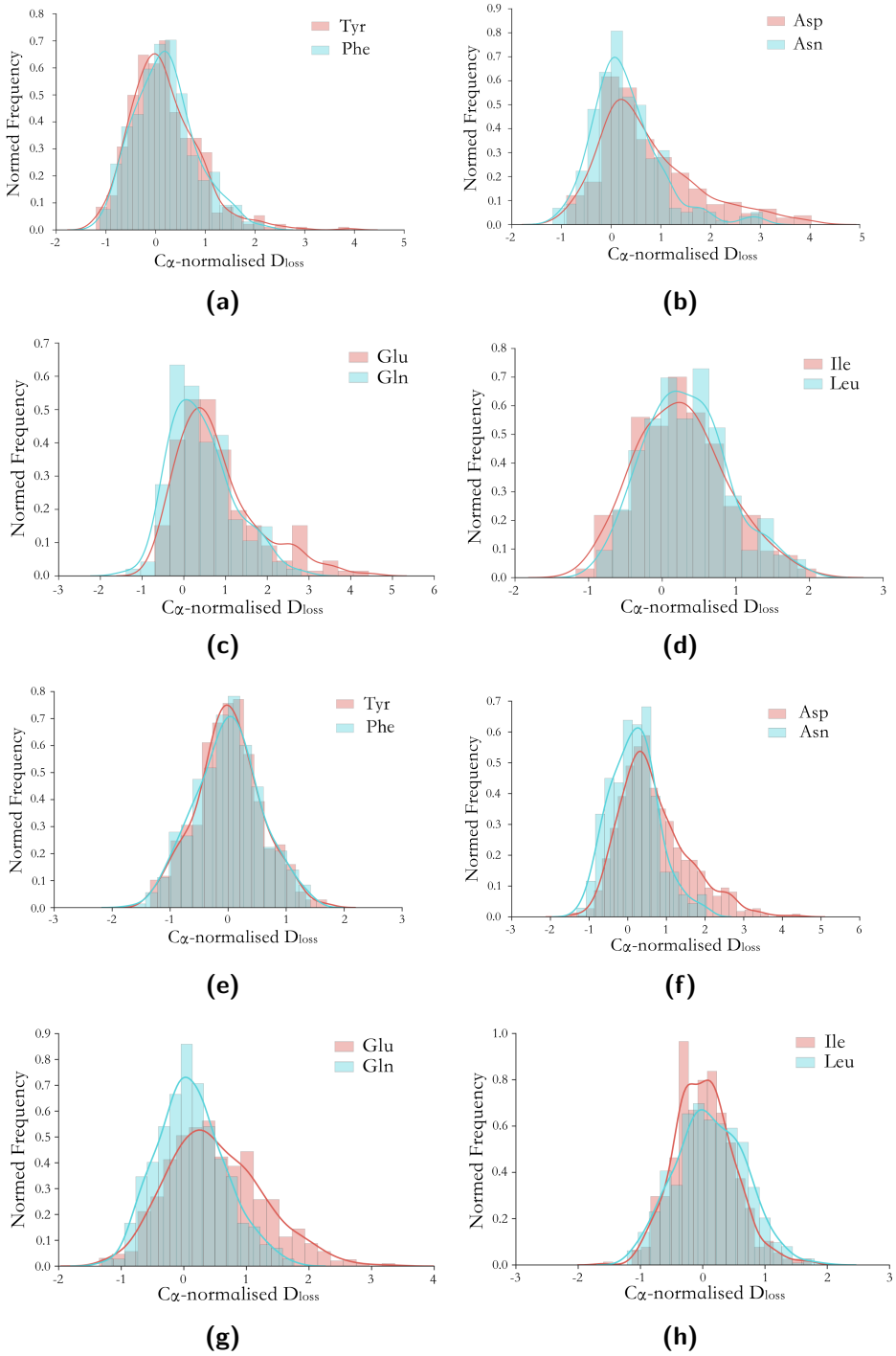


Figure 6.10: C_α-normalised D_{loss} damage signature histogram plots over all atoms of specific residue types for (a-d) myrosinase and (e-h) malate dehydrogenase. In each plot, two residue types are directly compared. Gaussian kernel density estimates (KDE), which are non-parametric approximations of the probability density function underlying each D_{loss} distribution, are overlaid on the histogram plots. For each protein, the D_{loss} values derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ map corresponding to the highest dose dataset are displayed; other doses in each series exhibit qualitatively similar behaviour (data not shown).

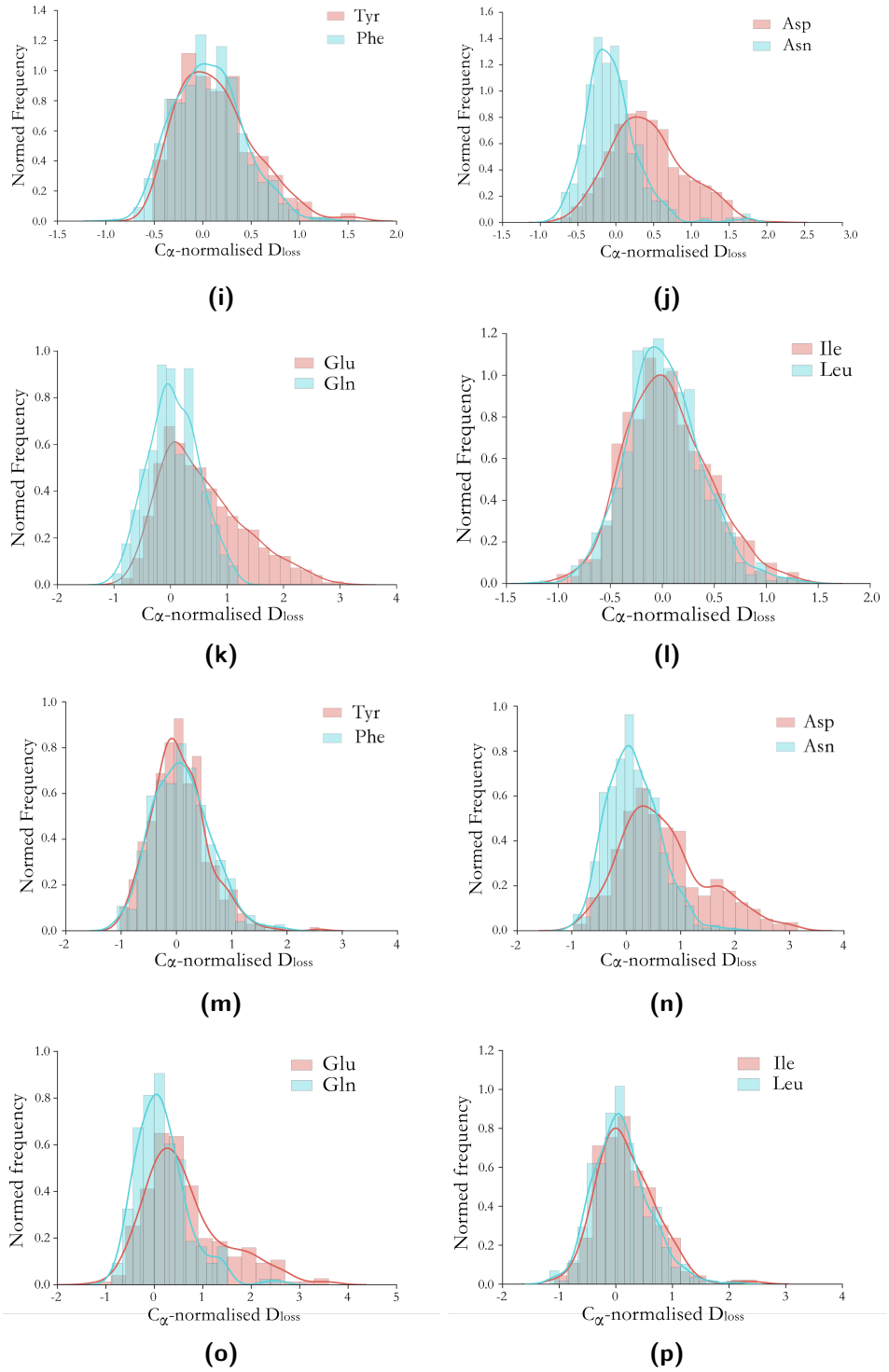


Figure 6.10: (cont.) C_α-normalised D_{loss} damage signature histogram plots over all atoms of specific residue types for (i-l) *Trp* RNA-binding attenuation protein and (m-p) glycoside hydrolase family 7 protein. In each plot, two residue types are directly compared. Gaussian kernel density estimates (KDE), which are non-parametric approximations of the probability density function underlying each D_{loss} distribution, are overlaid on the histogram plots. For each protein, the D_{loss} values derived from $F_{\text{obs,n}} - F_{\text{obs,1}}$ map corresponding to the highest dose dataset are displayed; other doses in each series exhibit qualitatively similar behaviour (data not shown).

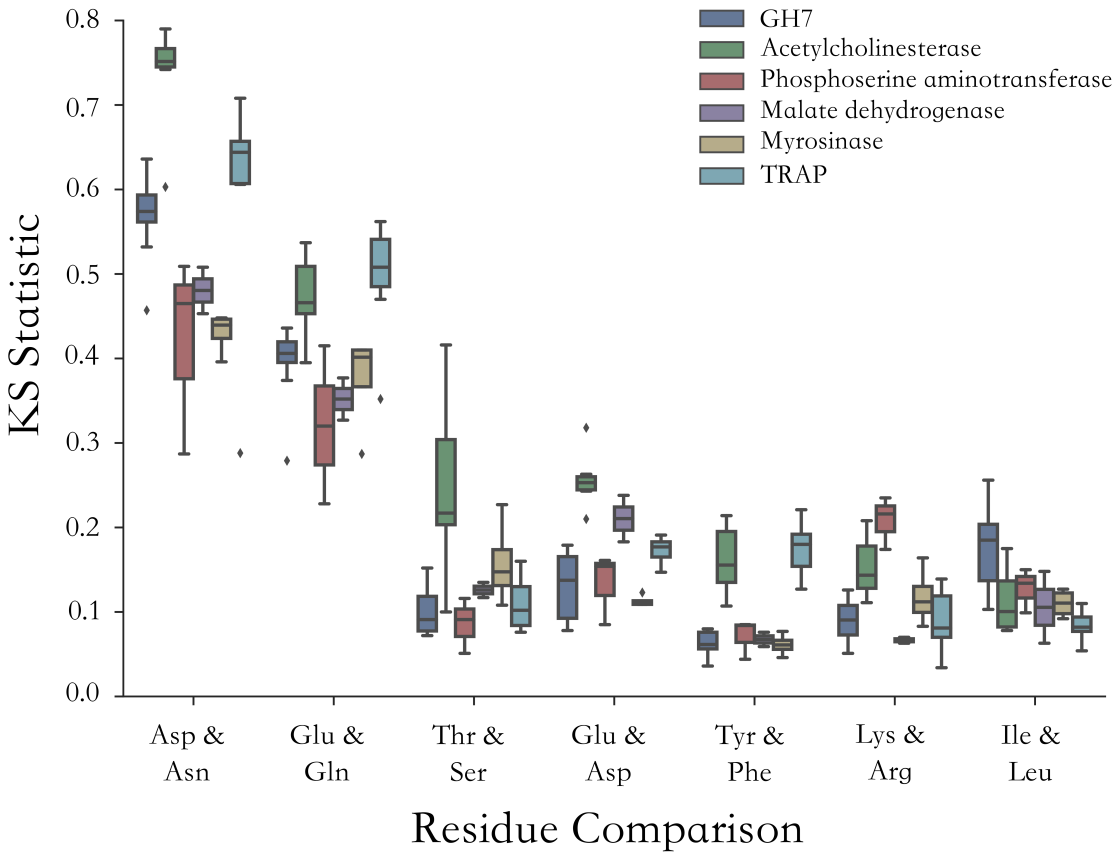


Figure 6.11: Kolmogorov-Smirnov (KS) test statistics comparing similarity of D_{loss} damage signatures (Figure 6.10) for different residue types in large protein structures ($>$ 60 kDa). For each damage series, box plots illustrate the distribution in KS statistic values across all the datasets in that series. A higher KS statistic indicates less similarity between the distribution of D_{loss} values for a specified residue pairing.

Conversely, Phe and Tyr exhibited qualitatively indistinguishable D_{loss} signatures (Figure 6.10b, f, j, n). Ile and Leu also showed qualitatively indistinguishable D_{loss} behaviour (Figure 6.10d, h, l, p), as expected, since both have similar aliphatic side-chain topologies that are believed to be relatively radiation-insensitive in protein crystals at 100 K, even at doses of several MGy.

For each residue pairing, the two-sample Kolmogorov-Smirnov (KS) test has been employed as a non-parametric statistic to test the null hypothesis that the D_{loss} values of both residue types have been drawn from the same distribution. A non-parametric test was deemed suitable, since it was unclear what underlying probability distribution a set of D_{loss} values should follow. The Kolmogorov-Smirnov test statistic between two residue types r_1 and r_2 is defined as:

$$D_{m,n} = \sup_x |F_{1,m}(x) - F_{2,n}(x)| \tag{6.4.1}$$

where, m and n are the number of atoms in a structure of residue type r_1 and r_2 respectively, and $F_{1,m}(x)$ and $F_{2,n}(x)$ are the empirical distribution functions (an estimation of the cumulative distribution function) for residue types r_1 and r_2. The KS-test statistic, $D_{m,n}$, thus provides a quantifier for the similarity between the D_{loss} signatures for two residue types, with a higher $D_{m,n}$ value indicating greater dissimilarity.

As can be seen in Figure 6.11, the Tyr-Phe KS-test statistic was consistently low and of the same order as the Leu-Ile control across the range of investigated proteins. The radiation effects to Phe and Tyr were thus deemed to be highly similar in protein crystals at 100 K in the MGy dose regime, suggesting that comparable spatial displacement of the aromatic rings is present for both residue types. It must be noted that this alone does not rule out cleavage of the C−O bond in Tyr.

6.4.2 Tyr O_η versus C_ζ correlation analysis

To attempt to discount cleavage of the C−O bond in Tyr, D_{loss} values for Tyr O_η have been compared directly with those for covalently bound Tyr-C_ζ atoms. Here, linear regression fitting has been restricted to structures deemed to contain statistically valid sample sizes of atoms (> 60 kDa).

For myrosinase, at the highest dose analysed (Figure 6.12a) a high positive correlation exists between Tyr-O_η and C_ζ D_{loss}, (linear R^2 : $0.72 - 0.82$ for dose range), which is of the order of both the Asp C_γ & $O_{\delta 1}$ (Figure 6.12c) and Glu C_δ & $O_{\epsilon 1}$ (Figure 6.12e) (with linear R^2 : $0.81 - 0.91$ and $0.83 - 0.91$ respectively for dose range). The observed high correlations for Asp and Glu is strongly consistent with full oxidative cleavage of the carboxylate. The comparably high correlation between D_{loss} for Tyr-C_ζ and Tyr-O_η supports a hypothesis of radiation-induced disordering of the overall tyrosyl aromatic ring, fixed as a covalently bound unit. For explicit cleavage of the phenolic C−O bond, a lower correlation would be anticipated, similarly to those reported between the Asp C_γ & C_β (Figure 6.12b) and Glu C_δ & C_γ (Figure 6.12d) atoms (linear R^2 : $0.11 - 0.23$ and $0.42 - 0.59$ respectively across the dose range). These reduced correlations are expected since the Asp-C_β and Glu-C_γ atoms are not predicted to be cleaved during side-chain decarboxylation. Since the Glu-C_γ atom may change orientation following decarboxylation, it is

unsurprising that the correlation between Glu C_δ & C_γ is higher than between Asp C_γ & C_β.

Similarly, for the TRAP damage series, the Asp C_γ & $O_{\delta 1}$ and Glu C_δ & $O_{\epsilon 1}$ D_{loss} behaviour was more correlated than that of Asp C_γ & C_β and Glu C_δ & C_γ (Figure 6.13b-d). For TRAP, the Tyr O_η and C_ζ behaviour was only weakly correlated across the large dose range studied ($1.3 - 25.0$ MGy) (Figure 6.13a). It is suggested here that the low correlation observed for Tyr could be a consequence of the 11-fold symmetry around each TRAP ring, as illustrated previously in Figure 4.17 of Chapter 4. Each of the two TRAP rings within the asymmetric unit contains eleven symmetry-related copies of a single Tyr residue (Tyr-62), with all eleven Tyr-62 copies around the non-RNA bound TRAP ring being involved in hydrogen bond interactions with a nearby glutamate (Glu-50). Consequently, the Tyr scatter plot may not contain an adequate sampling of Tyr residues that is suitable for linear regression analysis.

In contrast, as indicated in Figure 6.14, other proteins (malate dehydrogenase, acetylcholinesterase and phosphoserine aminotransferase) exhibited negligible D_{loss} correlation between Tyr O_η and C_ζ atoms. However, Tyr −OH groups were not flagged as highly radiation-sensitive in these structures (Figure 6.8, average Tyr −OH damage ranks: 42, 36 and 54 for these proteins respectively, with the highest individual Tyr −OH positions being above 290, 138 and 235 at all tested doses). Consequently, a correlation would not be expected between O_η and C_ζ, with noise dominating at low D_{loss} values. In summary, whereas the correlation between Tyr O_η and C_ζ atom electron density loss is inconsistent between the investigated proteins, in a subset of cases for which negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map peaks have been detected near Tyr −OH groups (myrosinase and TRAP), a positive correlation is present.

6.4.3 Occupancy refinement of Tyr groups in myrosinase

To confirm whether the original observations of Tyr −OH density loss in myrosinase were consistent with full aromatic ring displacement, secondary aromatic ring conformations (initially with 50 % occupancy) were built into the *1dwa* coordinate model retrieved from *PDB_REDO* following model rebuilding (Figure 6.15). This was performed for a subset of Tyr residues, including radiation-labile residues (Tyr-330, Tyr-215 and Tyr-390) and control groups where no $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density was detected near the Tyr O_η atom (Tyr-121, Tyr-495, Tyr-366). In each case, the

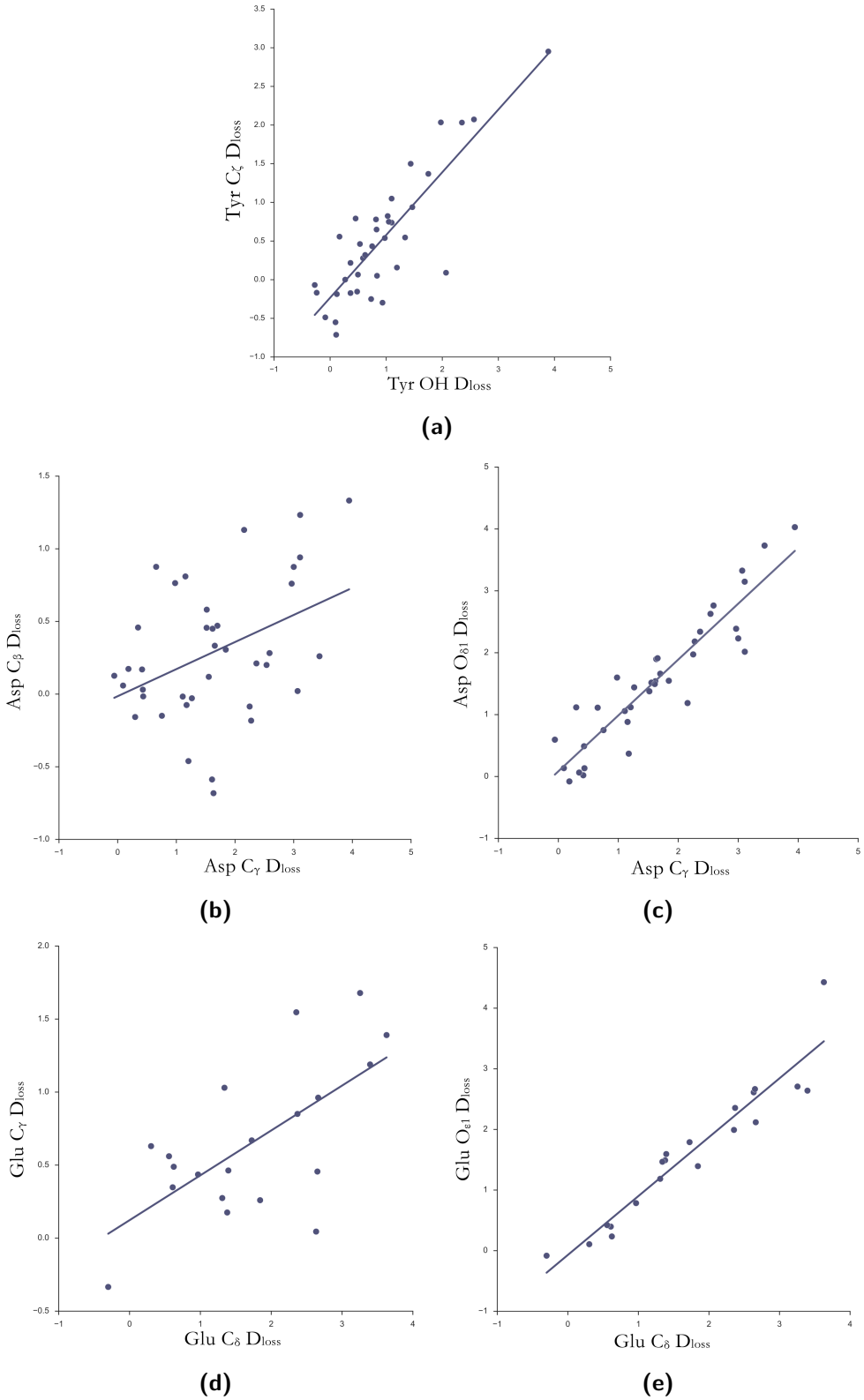


Figure 6.12: Scatter plots to compare C_α-normalised D_{loss}(atom) behaviour for covalently-bound atoms in the side-chains in (a) Tyr, (b,c) Asp and (d,e) Glu residues, for the highest dose myrosinase dataset ($n = 5$) (Burmeister, 2000). The linear coefficient of determination, R^2 has been computed for each scatter plot using the *Python* package *SciPy*.

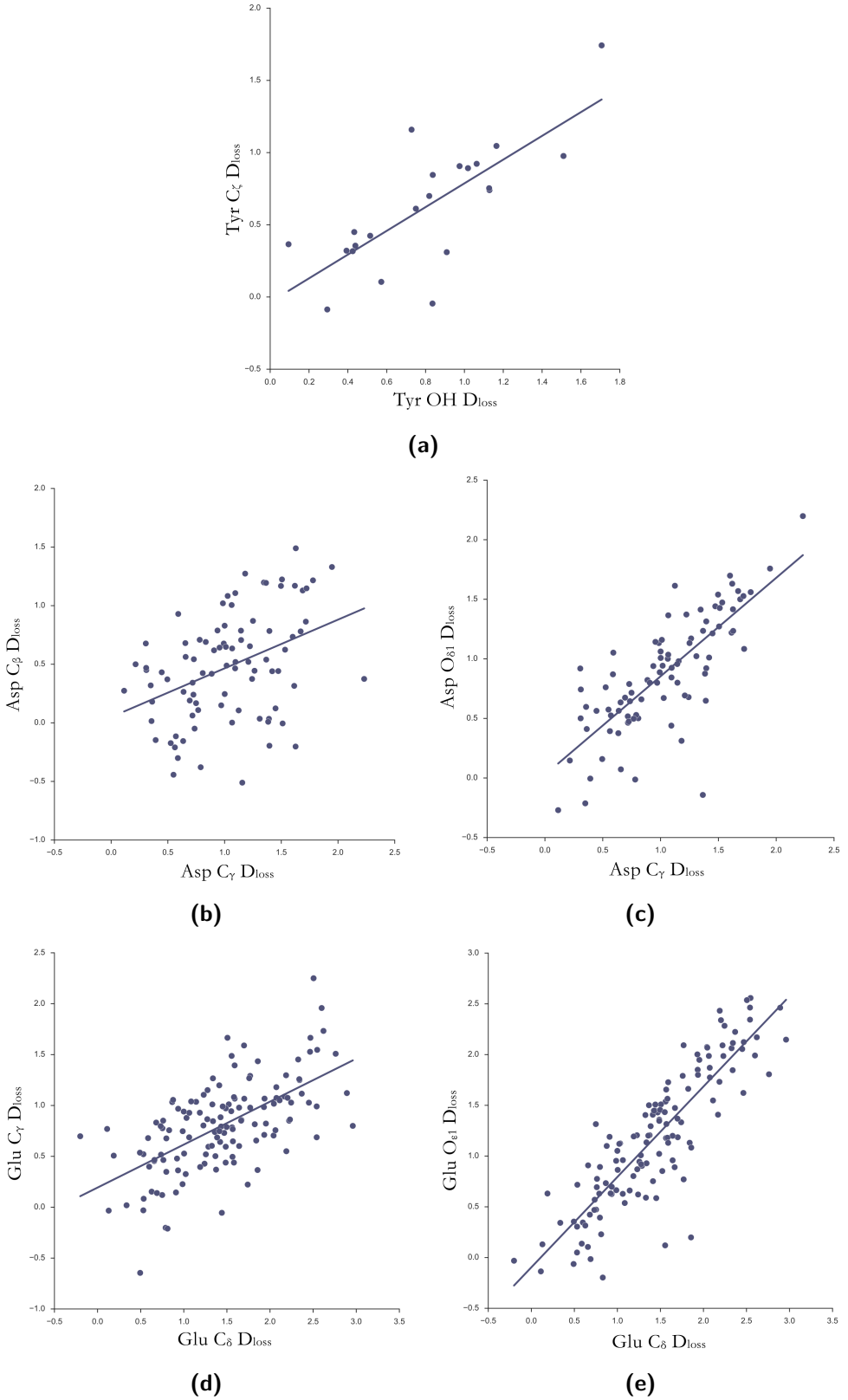


Figure 6.13: Scatter plots to compare C_α-normalised D_{loss}(atom) behaviour for covalently-bound atoms in the side-chains in (a) Tyr, (b,c) Asp and (d,e) Glu residues, for the highest dose *Trp* RNA-binding attenuation protein (TRAP) dataset ($n = 10$), as reported in Section 4.2. The linear coefficient of determination, R^2 has been computed for each scatter plot using the *Python* package *SciPy*.

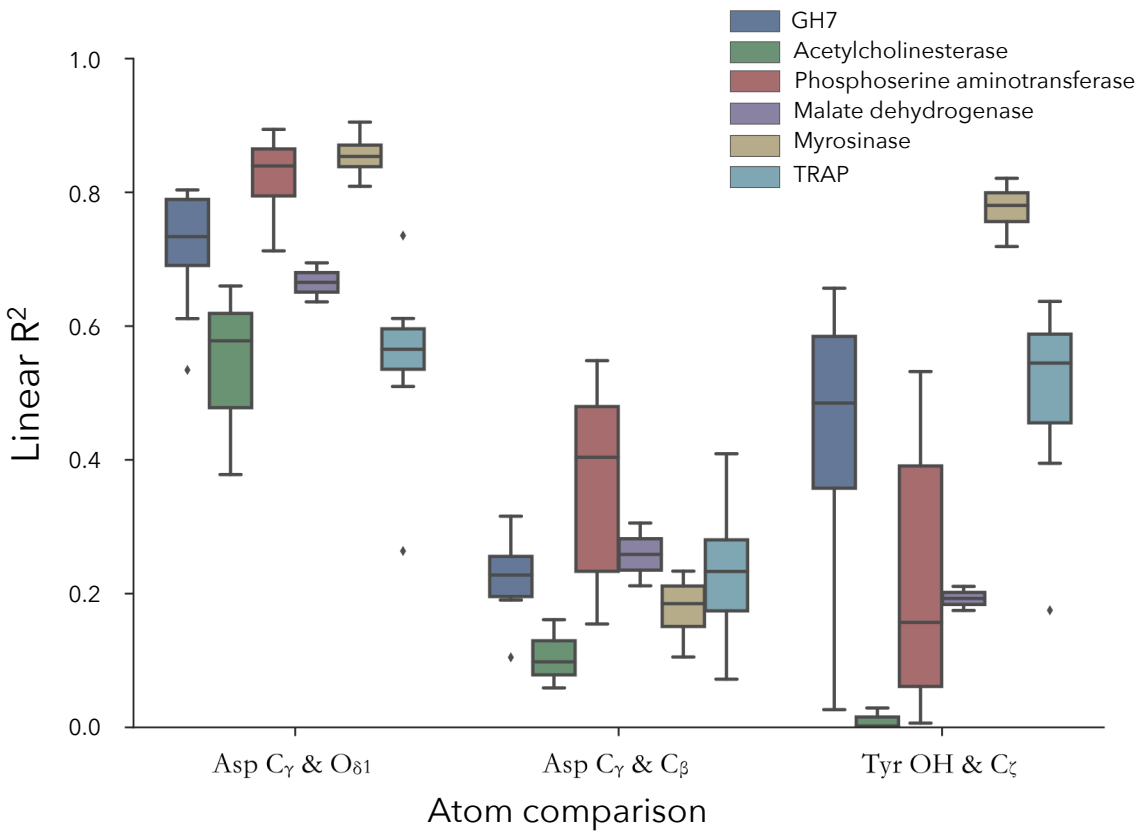


Figure 6.14: The linear coefficient of determination, R^2, quantifying the linear correlation between C_α-normalised D_{loss} metric values for selected covalently bound atom pairs in large protein structures ($>$ 60 kDa). Tukey Box plots illustrate the variation in R^2 value between each dataset for each separate damage series. Whiskers represent 1.5$\times$ the interquartile range for each protein. R^2 values have been calculated using the *Python* package *SciPy*.

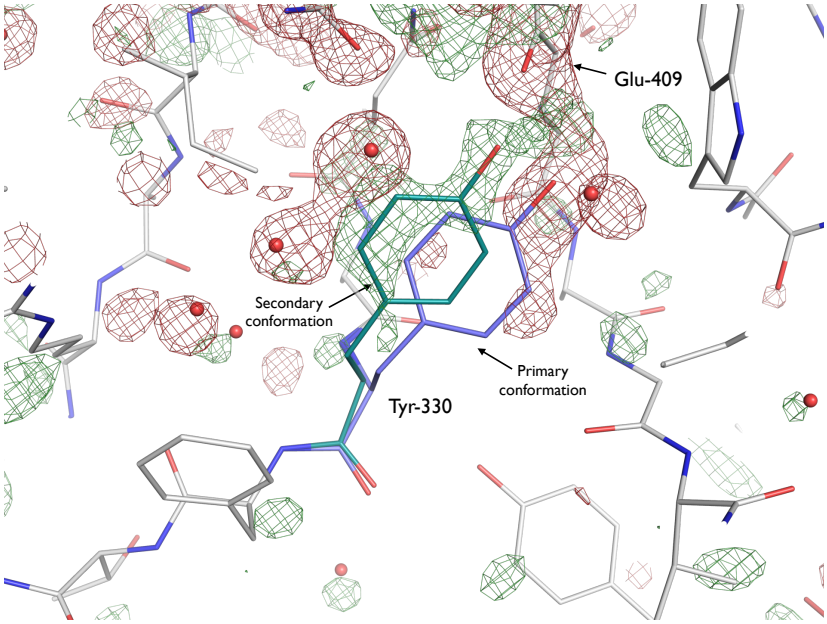


Figure 6.15: Example case where a secondary Tyr-330 side-chain conformation (occupancy: 50 %) was built into positive $F_{obs,4} - F_{obs,1}$ map difference density manually using *COOT*. $F_{obs,4} - F_{obs,1}$ map contoured at $\pm 3\sigma$ in green/red is overlaid on the initial myrosinase coordinate model (PDB accession code: *1dwa*).

$F_{obs,4} - F_{obs,1}$ map was used to guide placement of the secondary conformation, using positive difference map density proximal to the aromatic ring at low σ-levels. In practice, the '*real-space refine*' tool was used in *COOT* to manually reorientate each Tyr side-group, whilst still maintaining good side-chain geometry. This tool also worked for the control Tyr groups, where no significant $F_{obs,n} - F_{obs,1}$ density (either positive or negative) was present.

For the model from each dataset n, occupancy refinement was performed with *phenix.refine*, using the modified *1dwa* coordinate model and structure factor amplitudes, $|\mathbf{F}_{obs,n}|$. Atomic *B*-factors were held constant throughout occupancy refinement. With increasing dose, there was an observed shift in the relative occupancies of the two Tyr-330 conformations (Figure 6.16), with a similar observation being made for Tyr-215, Tyr-390 and Tyr-432 (secondary conformations not shown). The growth of each secondary conformation occupancy is consistent with partial reorientation of each Tyr ring with dose. For the control group (Tyr-121, Tyr-495 and Tyr-366), no significant change in occupancy of the primary conformation was observed with dose. Tyr-366 occupies both conformations with non-zero occupancy in the first dataset, however their proportions do not fluctuate with dose.

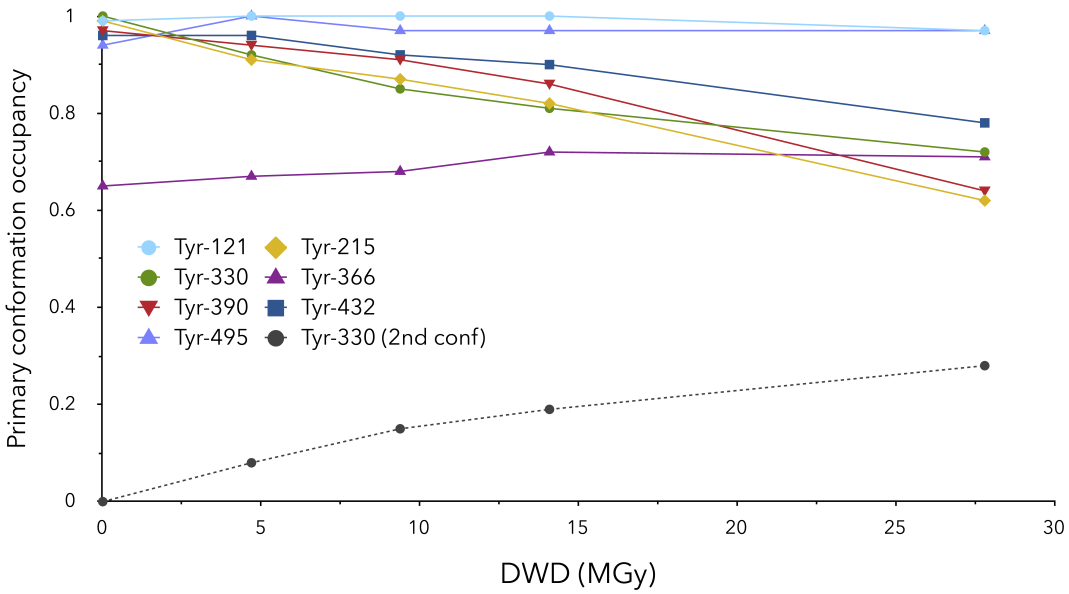


Figure 6.16: For a selected subset of Tyr residues in myrosinase, secondary conformations have been built manually using *COOT* (see Figure 6.15) and the occupancy of each primary side-chain conformation has been plotted as a function of dose, following occupancy refinement (in *phenix.refine*, with atomic *B*-factors held constant) of the model pertaining to each successive dataset in the damage series. Both groups suspected to be radiation-labile (Tyr-330, Tyr-215 and Tyr-390) and control groups where no $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map density was detected near the Tyr O_{η} atom (Tyr-121, Tyr-495, Tyr-366) have been included. For clarity, the secondary conformation for Tyr-330 has also included, to demonstrate that the occupancies for the primary and secondary conformation sum to 1 at each dose. The secondary conformations for other Tyr groups have not been shown. Interestingly, Tyr-366 occupies both conformations with non-zero occupancy in the first dataset, however these values do not fluctuate with dose.

6.4.4 Simulation of Tyr damage with $F_{calc} - F_{calc}$ maps

It is necessary to confirm that a model of displacement of the full Tyr aromatic ring is consistent with negative $F_{obs,n} - F_{obs,1}$ map density being most prominent at the Tyr O_η atom. This section now aims to simulate artificial Tyr ring motion in a protein, to determine what visual signature is left in the resulting $F_{obs,n} - F_{obs,1}$ map.

As in Section 6.4.3, a secondary Tyr-330 conformation was built into the *1dwa* myrosinase coordinate model retrieved from *PDB_REDO* using *COOT*. The original Tyr-330 conformation was deleted from the model, to create a modified *1dwa* model, post X-ray irradiation. A custom script was written in *Python* to calculate a $F_{calc,modified} - F_{calc,original}$ map between the calculated structure factor amplitudes corresponding to the original *1dwa* model and newly altered model. This has been designed to artificially simulate movement of the Tyr-330 side-group, as would be observed in a real $F_{obs,n} - F_{obs,1}$ map. To calculate the $F_{calc,modified} - F_{calc,original}$ map, the script first generates calculated structure factor amplitudes, $|\mathbf{F}_{calc}|$, for both the original and modified *1dwa* coordinate models using the CCP4 program *SFALL* and then calls *RIDL* to generate a $F_{calc,modified} - F_{calc,original}$ map using *FFT*, with the observed structure factor amplitude information replaced by the calculated equivalents, $|\mathbf{F}_{calc,modified}|$ and $|\mathbf{F}_{calc,original}|$. A further improvement to the current approach (not presented here) would be to include a noise term to each $|\mathbf{F}_{calc}|$ amplitude prior to calculation of the $F_{calc,modified} - F_{calc,original}$ map, in order to simulate random noise present in a true $F_{obs,n} - F_{obs,1}$ map; this was deemed unnecessary for the purpose of the current work.

The results of this simulation of damage are shown in Figure 6.17. Since this is merely a simulation of damage to Tyr-330, and the rest of the structure was unaltered, the map σ contouring units are meaningless. However it is clear that the most significant electron density loss site coincides with the Tyr O_η atom in Figure 6.17d. It should be noted that for simplicity the secondary conformation (post-irradiation) is modelled at 100 % occupancy, and consequently a positive difference peak is also clearly present at the new position of the Tyr O_η atom in Figure 6.17. In a more realistic scenario, in the presence of additional Tyr disorder, such a pronounced positive peak would not be anticipated.

Overall, these results are in strong agreement with the hypothesis proposed originally by Burmeis-

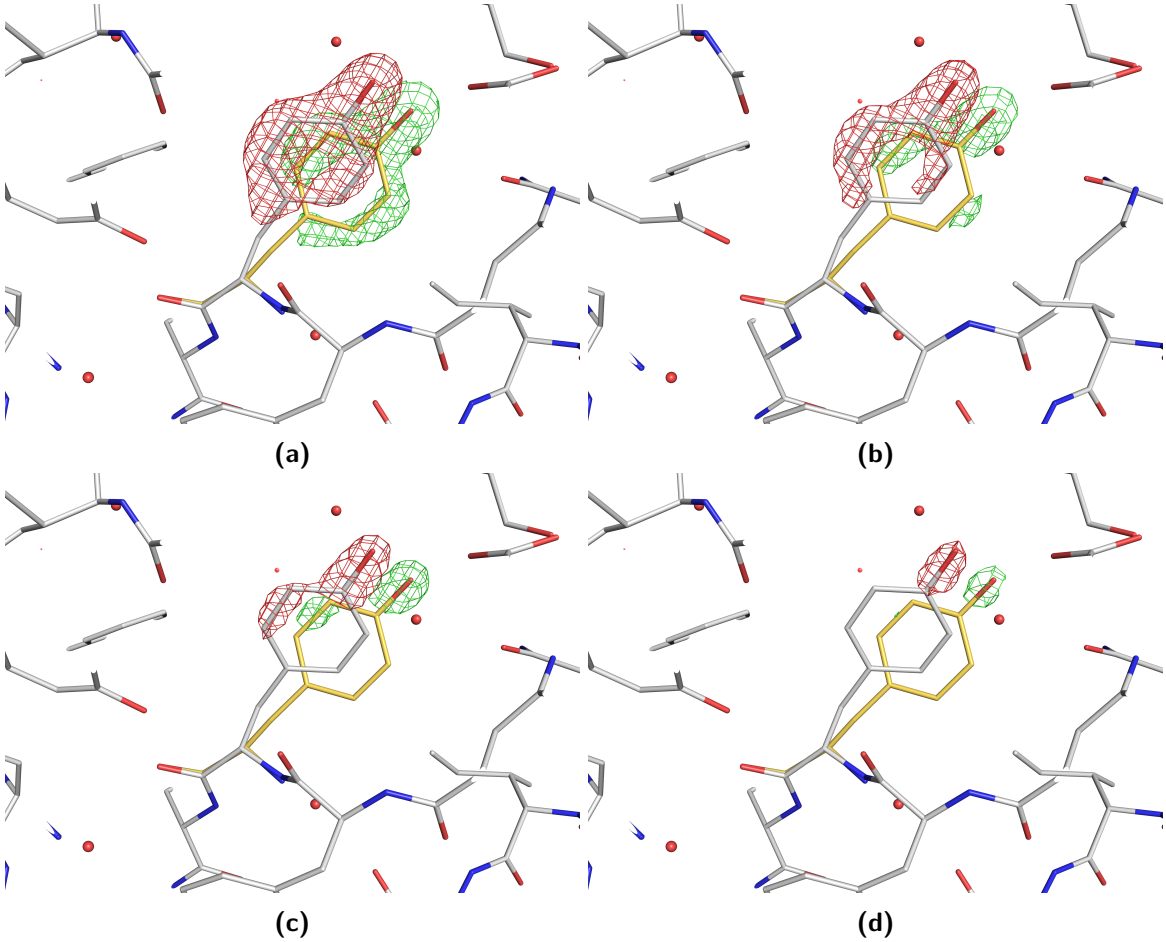


Figure 6.17: A simulation of radiation-induced conformational change of Tyr-330 in myrosinase. The original *1dwa* myrosinase coordinate model retrieved from *PDB_REDO* is shown in white, with a new artificially determined conformation for Tyr-330 shown in yellow. The $F_{\text{calc,modified}} - F_{\text{calc,original}}$, calculated as the difference in the calculated structure factor amplitudes, $|\mathbf{F}_{\text{calc}}|$, between the original (white) and modified (yellow) model, is overlaid in green/red at (a) $\pm 40\sigma$, (b) $\pm 50\sigma$, (c) $\pm 60\sigma$ and (d) $\pm 70\sigma$.

ter (2000), in which "hinge-like motion of the tyrosine ring" around the C_β atom would be "most prominent for the hydroxyl group".

6.5 Automated approaches to detection of radiation-induced motion

Whereas the above section indicates that radiation-induced motion of Tyr groups can be distinguished by manual building of secondary conformations coupled with occupancy refinement, this process is laborious and requires informed insight into where a possible Tyr conformation can be built without clashing with neighbouring atoms. This section aims to investigate the possibility

of automated approaches to detect such radiation-induced conformational changes, to enable a high-throughput screening across proteins.

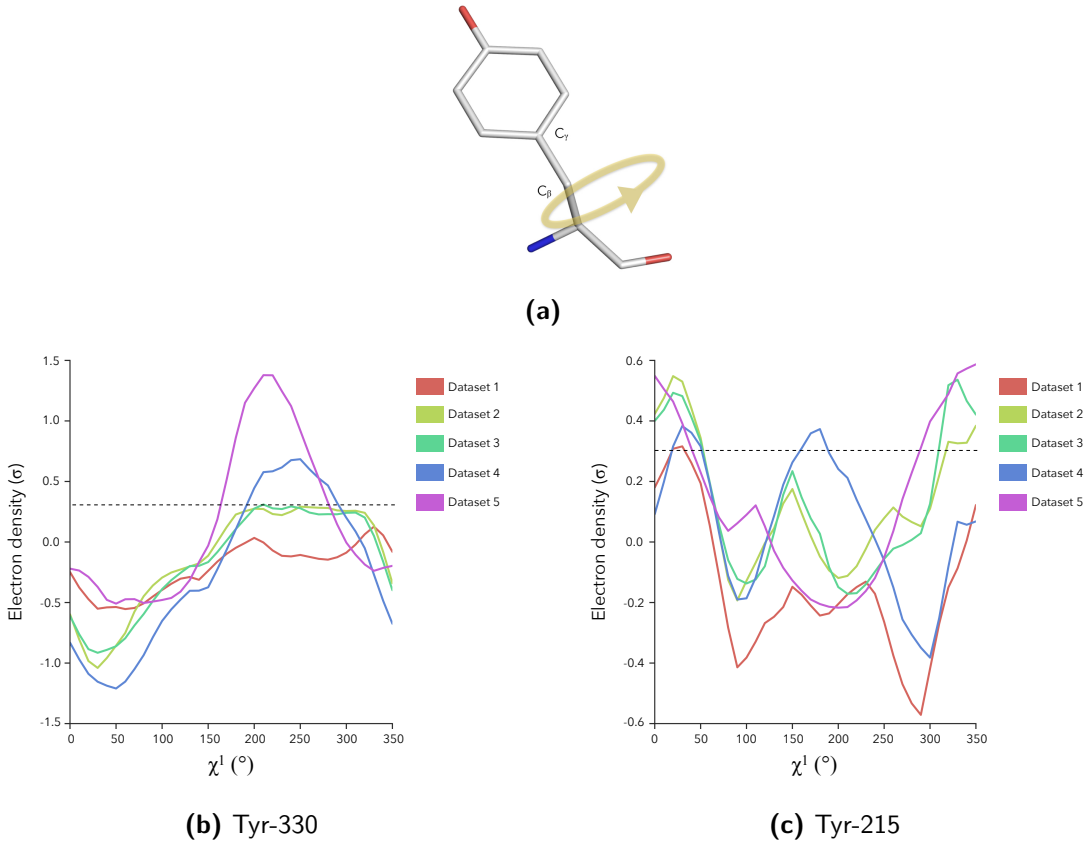
6.5.1 Automatic detection with *RINGER*

The program *RINGER* (Lang *et al.*, 2010) was designed to identify lowly populated alternate conformations, by systematically sampling electron density around the dihedral angles of protein side chains. In a study on 485 representative proteins extracted from the PDB at variable temperatures, *RINGER* found evidence for non-modelled secondary conformations above the map noise level for $\sim 45\,\%$ of side-chains (Lang *et al.*, 2014), and thus appeared highly suited in the current investigation as a tool for high throughput screening of radiation-induced changes in protein MX damage series.

RINGER was initially trialled for the re-refined myrosinase damage series discussed above, using the $2\mathrm{m}F_{\mathrm{obs}} - \mathrm{D}F_{\mathrm{calc}}$ map at each dataset. The results of *RINGER*, restricted to Tyr residues, are presented in Table 6.4. Interestingly, in contrast to the results detailed in Section 6.4.3 above, the formation of unmodelled alternative conformations for the top two Tyr damage sites in myrosinase (Tyr-215 and Tyr-330) was not detected using *RINGER* at low $2\mathrm{m}F_{\mathrm{obs}} - \mathrm{D}F_{\mathrm{calc}}$ map σ-thresholds (above a default $+0.3\sigma$ cutoff as specified in the Lang *et al.* (2010) publication).

Figure 6.18 shows the electron density (σ) as a function of χ^1 angle. In the original Lang *et al.* (2010) publication presenting *RINGER*, such plots are used to identify unmodelled conformations lying above a $+0.3\sigma$ level. Here the electron density is sampled in a ring around an axis perpendicular to the covalent bond between the C_β and the C_γ atoms of each Tyr group (Figure 6.18a). In the case of Tyr-330, only for datasets 4 and 5 are *single* peaks above $+0.3\sigma$ located, corresponding to the primary conformation, with not even a single conformation detected for the other datasets. Unexpectedly for dataset 1, the electron density height does not vary significantly with χ^1 angle, and a fixed angle for the primary conformation could not be distinguished using *RINGER*.

Furthermore, there appeared to be only poor agreement between datasets in the myrosinase damage series, with many Tyr residues unexpectedly flagged in only a subset of non-adjacent datasets (Table 6.4). Indeed in Figure 6.18b, the variation of electron density with χ^1 angle



(a)

(b) Tyr-330

(c) Tyr-215

Figure 6.18: Variation in electron density (σ) sampled around the χ^1 angle (illustrated in (a) by the yellow ring) of (b) Tyr-330 and (c) Tyr-215 for each dataset in the myrosinase damage series (Burmeister, 2000). Data have been generated using *RINGER* and plotted using the *Seaborn* package in *Python*. The dashed horizontal line indicates the default $+0.3\sigma$ map threshold for detection of conformations, as specified in Lang *et al.* (2010).

Table 6.4: Number of conformations detected by *RINGER* per Tyr side-group for each dataset in the myrosinase damage series (Burmeister, 2000). An entry of 1 indicates that the primary conformation of the residue was detected above the default $+0.3\sigma$ map threshold for detection of conformations, as specified in Lang *et al.* (2010). A dash indicates that no conformations satisfied this condition. *RINGER* samples electron density around the χ^1 angle, in a perpendicular ring between the C_β and the C_γ atoms of each Tyr group.

	Dataset						Dataset				
Residue	1	2	3	4	5	Residue	1	2	3	4	5
Tyr-78	-	-	-	-	2	Tyr-359	2	2	-	-	-
Tyr-120	2	-	-	-	2	Tyr-384	-	2	2	-	-
Tyr-121	2	-	-	2	-	Tyr-385	2	3	4	3	4
Tyr-152	-	-	2	-	-	Tyr-390	2	2	2	-	-
Tyr-167	-	-	2	-	2	Tyr-395	-	2	2	2	-
Tyr-181	2	-	-	-	-	Tyr-400	-	-	-	2	2
Tyr-215	-	2	2	2	1	Tyr-401	-	-	-	-	2
Tyr-241	2	2	2	2	3	Tyr-406	2	3	3	4	-
Tyr-260	-	-	-	-	2	Tyr-427	-	2	2	-	-
Tyr-264	-	2	2	2	2	Tyr-432	-	-	-	-	3
Tyr-322	2	2	-	2	3	Tyr-454	-	2	3	-	-
Tyr-330	-	-	-	1	1	Tyr-477	-	-	-	-	2
Tyr-334	-	-	-	2	2	Tyr-495	2	2	2	3	-

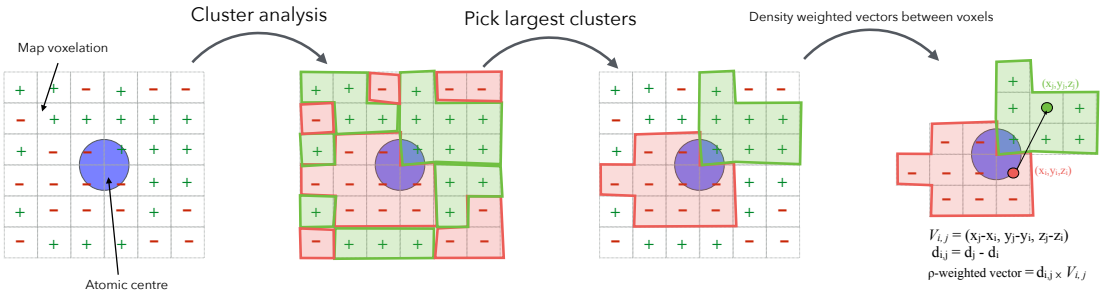


Figure 6.19: Schematic to demonstrate the partitioning of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxels assigned to an atom by cluster analysis. The schematic has been presented in only two dimensions here for simplicity. A k-means algorithm is used to locate clusters of negative and positive map values using the *scikit-learn* package in *Python* (Pedregosa *et al.*, 2011). The two clusters corresponding to the most positive and negative map values, V^{+}_{atom} and V^{-}_{atom}, are then isolated, and density-weighted vectors between individual voxels in V^{+}_{atom} and V^{-}_{atom}, $d_{i,j} \times V_{i,j}$, are calculated.

for Tyr-215 exhibited minimal trend between datasets. It thus appeared that *RINGER* was an unsuitable high-throughput tool for screening for radiation-induced lability. This is a reasonable limitation of the tool, since *RINGER* was not specifically designed to handle diminishing quality $2mF_{\text{obs}} - DF_{\text{calc}}$ maps at high dose values.

6.5.2 Automatic detection of electron density shifts with *RIDL*

This section aims to demonstrate that the *RIDL* pipeline developed in Section 2.5 can be modified for the purpose of detecting radiation-induced density shifts for atoms.

In order to calculate damage metrics per atom, *RIDL* extracts a set of $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxels, V_{atom}, assigned to each atom (Figure 6.19a). Since a conformation change, for instance to Tyr-330 in myrosinase, is characterised by the presence of a clustering of negative and positive $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density in the vicinity of the atom (Figure 6.15), if the Cartesian x, y, z coordinates are known for each voxel, such clusters of negative and positive density can be automatically located using *RIDL*. In fact, voxels in a $F_{\text{obs,n}} - F_{\text{obs,1}}$ `.map` file are formatted as a one-dimensional list. However, *RIDL* can determine the x, y, z coordinates for each voxel assigned to an atom, with knowledge of both the unit cell dimensions and map grid sampling coarseness as reported in the `.map` file header information. This is now performed by a custom script in *RIDL*.

For a set of voxels, V_{atom}, assigned to an atom, k-means cluster analysis is then performed using

the *scikit-learn* package in *Python* (Pedregosa *et al.*, 2011) to isolate clusters of negative and positive $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxel values (Figure 6.19b). Clusters are determined using both the Cartesian x, y, z coordinates of each voxel and the corresponding density value, d (input in the form of a four-dimensional vector $[x, y, z, d]$ for each voxel). Formally, given a set of n voxels, V_{atom}, each with four-dimensional vector $\vec{x} = [x, y, z, d]$, k-means clustering aims to partition these n voxels into $k \leq n$ sets, $S = \{S_1, S_2, ..., S_k\}$ such that the *within-cluster sum of squares* expression:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in S_i} \|\vec{x} - \vec{\mu_i}\|^2 \tag{6.5.1}$$

is minimised, where $\vec{\mu_i}$ is the mean of points in S_i and $\| \cdot \|$ is the L^2-norm (Euclidean norm). Using a k-means clustering algorithm therefore requires that the number of clusters is known. By default, $n = 2$ negative clusters are specified in order to isolate a single spatial cluster of high negative values (and discard all other voxels), and a higher number of $n = 5$ positive clusters are specified to locate multiple regions of positive $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density.

Currently, only a single negative and positive cluster are taken into account (Figure 6.19c), corresponding to the most negative and positive regions of voxels around an atom, V^{-}_{atom} and V^{+}_{atom}, respectively. The aim is to quantify the magnitude of electron density shift between these two clusters. As indicated in Figure 6.19d, sets of inter-cluster vectors, $\vec{r_{i,j}}$ are then defined between each voxel v_i^- in V^{-}_{atom} and each voxel v_j^+ in V^{+}_{atom}. The length of each vector is then weighted by the difference in density values, $d_j - d_i$ between voxels v_j^+ and v_i^-. An overall vector, $\vec{r}$, between cluster V^{-}_{atom} and V^{+}_{atom} can then be defined as:

$$\vec{r} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} (d_j - d_i) \cdot \vec{r_{i,j}} \tag{6.5.2}$$

Calculating the L^2-norm of this vector provides the magnitude, denoted D_{shift}. A high D_{shift} value indicates the presence of significant positive and negative $F_{\text{obs,n}} - F_{\text{obs,1}}$ map density at an atom, and can thus be used to screen for sites of radiation-induced density shifts.

In Figure 6.20, for the case of O_η atoms of Tyr residues in myrosinase, the D_{shift} has been calculated by *RIDL* for the $F_{\text{obs,2}} - F_{\text{obs,1}}$ and $F_{\text{obs,5}} - F_{\text{obs,1}}$ map[§]. Here, the D_{shift} metric

[§]In practice, in *RIDL* the per-atom search radius (defined by the parameter α in the program *SFALL*) was increased to the maximum possible value $\alpha = 3$ in order to ensure a large enough voxel-sampling radius around each atom to also include positive $F_{\text{obs,n}} - F_{\text{obs,1}}$ map values in the solvent regions around each atom. The default

was able to distinguish conformational changes in the highest dose $F_{obs,5} - F_{obs,1}$ map most significantly at Tyr-215, Tyr-330 and Tyr-400, as anticipated. Notably, these Tyr groups are also flagged as radiation-labile even in the $F_{obs,2} - F_{obs,1}$ map.

It has thus been demonstrated that an automated approach may be developed to screen for radiation-induced lability in proteins. It is however noted that there are limitations in the design of the D_{shift} metric presented above. Specifically, no account has been made for other atoms in the Tyr side-group, which also must simultaneously move with the O_η atom, nor the feasible range of motion of each side-chain. In general, the D_{shift} metric thus presents a simple indicator of *possible* conformational shifts, but is predicted to be sensitive to false positive events. Improvements to the D_{shift} metric have been left as future work.

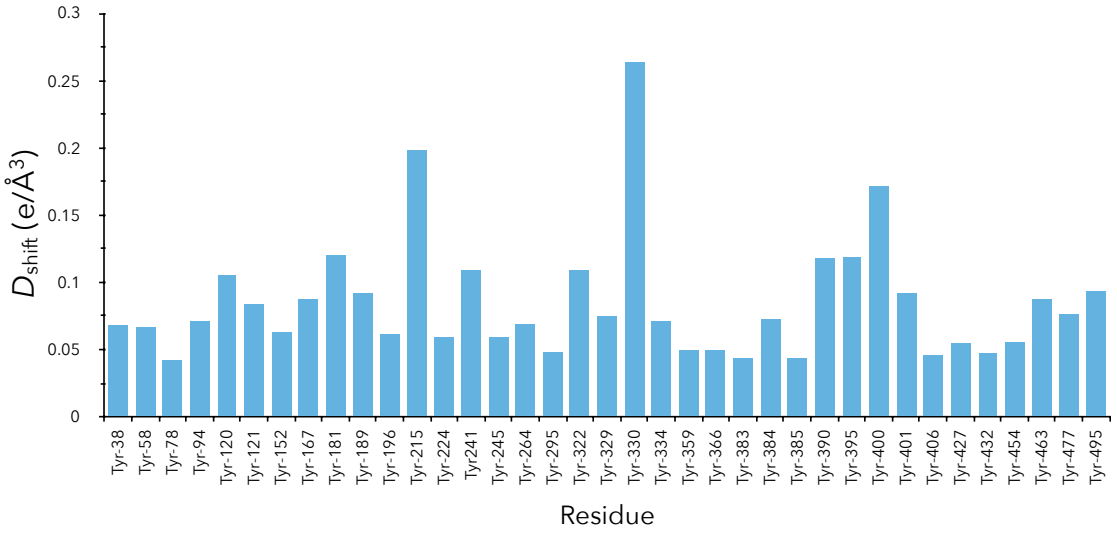
6.6 Factors affecting the Tyr −OH group

6.6.1 Availability of hydrogen bonds

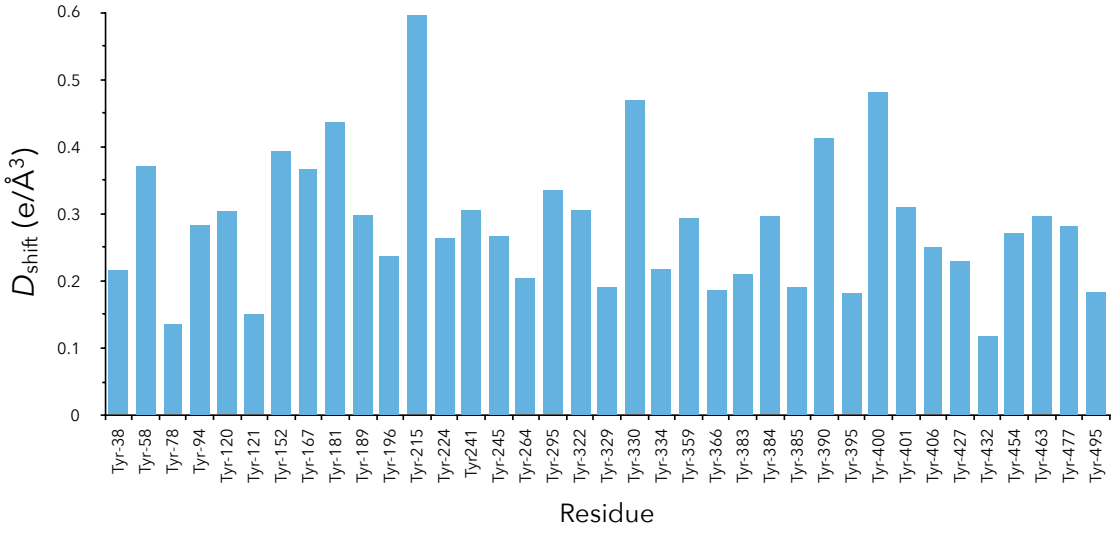
There has been anecdotal evidence to suggest a relationship between radiation-induced Tyr −OH electron density disordering and the presence of hydrogen bond interactions to Glu/Asp carboxyl side-groups (Burmeister, 2000). Indeed, as described in Section 4.3.5 for TRAP (a macromolecule with 11-fold rotational self-symmetry), eleven symmetrically related copies of Tyr-62 around the RNA non-bound protein all hydrogen bond to nearby Glu-50 residues, and all exhibit significantly higher D_{loss} metric values relative to the non-hydrogen bonded equivalent Tyr-62 residues on the RNA-bound ring (Figure 4.19). Furthermore, the most radiation-labile Tyr groups in myrosinase (Tyr-330 and Tyr-215) both exhibit such hydrogen bonding interactions (Figure 6.4 and 6.3).

This section aims to establish whether a relationship between radiation-induced Tyr −OH electron density disordering and the availability of proximal hydrogen bond interactions to acidic residues exists across independent protein structures. For each structure reported in Table 6.2 as retrieved from *PDB_REDO*, Tyr residues were grouped into two subsets: those predicted to interact with Glu/Asp carboxylates at distances under 4 Å (between the Tyr O_η oxygen and the carboxylate

value of $\alpha = 1$ resulted in no positive $F_{obs,n} - F_{obs,1}$ density being assigned to a subset of Tyr O_η atoms, and failure to calculate the D_{shift} metric. See Section 2.5.5 of Chapter 2 for a detailed discussion of the α parameter.

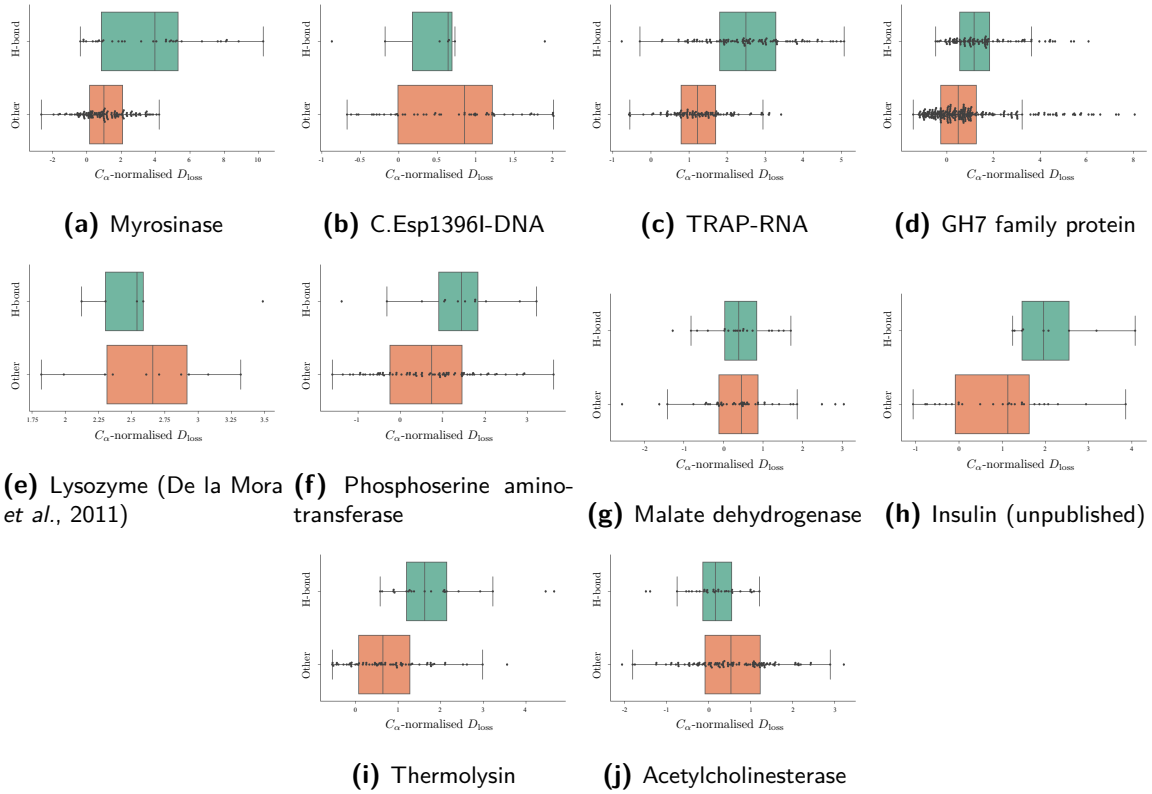


(a)



(b)

Figure 6.20: The D_{shift} metric for O_η atoms of Tyr residues in the myrosinase damage series (Burmeister, 2000). Metric values have been calculated for the (a) $F_{\text{obs},2} - F_{\text{obs},1}$ and (b) $F_{\text{obs},5} - F_{\text{obs},1}$ map by the *RIDL* pipeline, in order to detect sites of significant radiation-induced conformational shift.



(a) Myrosinase **(b)** C.Esp1396I-DNA **(c)** TRAP-RNA **(d)** GH7 family protein

(e) Lysozyme (De la Mora *et al.*, 2011) **(f)** Phosphoserine aminotransferase **(g)** Malate dehydrogenase **(h)** Insulin (unpublished)

(i) Thermolysin **(j)** Acetylcholinesterase

Figure 6.21: The effect of Glu/Asp carboxylate hydrogen bonding on the distribution of C$_\alpha$-normalised D_{loss}(atom) values for Tyr –OH groups in MX damage series retrieved from *PDB_REDO* prior to model rebuilding. C$_\alpha$-normalisation was performed as discussed in Section 2.5.8, such that in each plot $D_{\text{loss}}^{\text{C}_\alpha} = 0$ corresponds to the average metric value attained by the protein backbone. Each plot combines metric values derived from all higher dose $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps that could be calculated from each independent damage series. Only a subset of protein damage series have been shown which were deemed to contain enough data points for box plot analysis. Tukey boxplots have been generated using the *Seaborn* package in *Python* such that the whiskers represent 1.5× the interquartile range. In addition, raw data are overlaid over each boxplot as points.

oxygens), and all other Tyr residues. This calculation was performed using the CCP4 program *NCONT*, whilst accounting for crystal contacts between adjacent asymmetric units (Winn *et al.*, 2011). Tyr –OH density loss has been quantified using the D_{loss} metric as computed by *RIDL*.

For each protein, Figure 6.21 indicates the dependence of hydrogen bond interactions on C$_\alpha$-normalised D_{loss}. For clarity, box plots have only been generated for MX damage series for which a sufficient sample set of Tyr O$_\eta$ atoms were present. Consequently the relative small proteins analysed for the Nanao *et al.* (2005) RIP study have been omitted. In one protein, elastase (Petrova *et al.*, 2010), no Tyr residues were predicted to form such hydrogen bonds, and thus it has also been excluded.

Overall, a significant trend was not observed across proteins. Whereas Figure 6.21a & c for myrosinase and TRAP are consistent with the hypothesis that hydrogen-bonded Tyr groups are

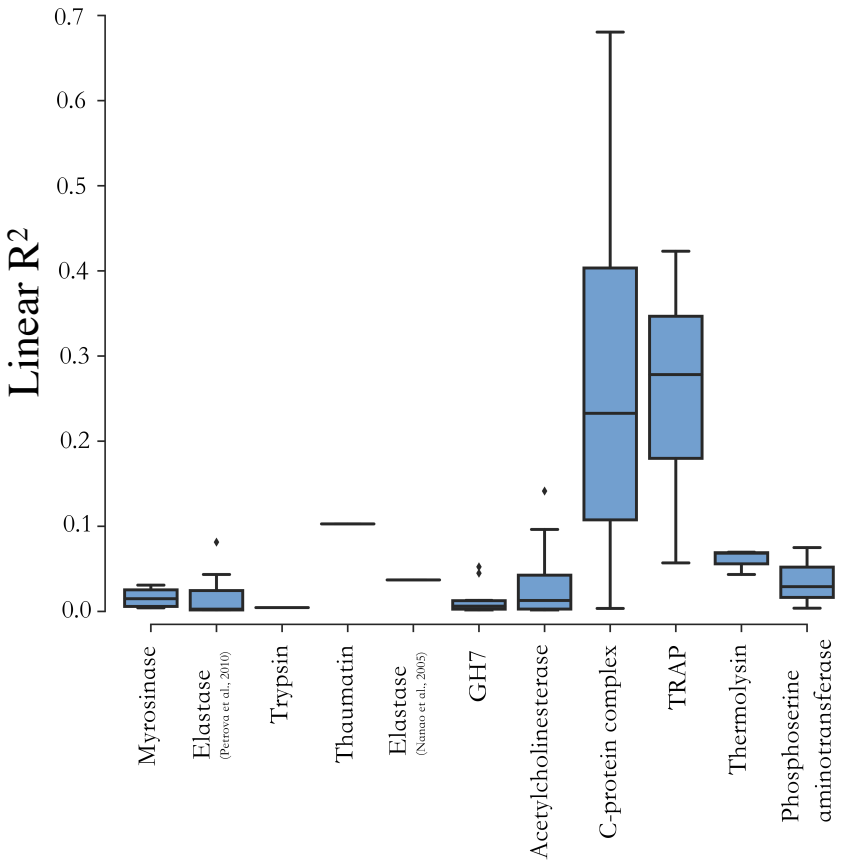


Figure 6.22: Linear R^2 coefficients calculated for each MX damage series for correlation between Tyr −OH D_{loss} metric values and solvent accessibility of each Tyr aromatic ring, following removal of all Glu/Asp carboxyl atoms from each coordinate model. R^2 values have been calculated for each dose in a damage series, and Tukey box plots represent distribution in R^2 across these multiple datasets. Whiskers represent 1.5× the interquartile range for each box plot. Only structures containing eight or more Tyr residues per asymmetric unit have been included.

more radiation-labile, there are numerous cases (e.g. malate dehydrogenase in Figure 6.21g) for which this is not true, and even proteins for which the contrary holds (e.g. acetylcholinesterase in Figure 6.21j). It is argued that the loss of a direct Tyr-Glu/Asp hydrogen bond may not in itself be sufficient for increased Tyr ring flexibility, and undoubtedly other factors will influence the Tyr local environment: the identity of the remaining radicalised Glu/Asp species post decarboxylation, the proximity of additional hydrogen bond acceptors, and aromatic ring stacking interactions. However, these individual factors are challenging to deconvolute and quantify.

In an attempt to parametrise the reduction in steric hindrance of the Tyr aromatic ring following decarboxylation of a nearby hydrogen bonded Glu/Asp, the solvent accessible area for each Tyr side-group was calculated, after first stripping all Glu and Asp carboxylate atoms from each protein coordinate model. This was performed using a custom script in *Python*, with per-atom solvent

accessibilities calculated using the CCP4 program *AREAIMOL*, whilst accounting for any crystal contacts between adjacent asymmetric units. The average solvent accessible area across all atoms in the Tyr side-group, as opposed to just the O_η atom, to provide a better indicator of steric hindrance for the full aromatic ring. Across proteins only a poor correlation was observed between Tyr $-$OH D_{loss} values and the resulting solvent accessibility (Figure 6.22).

Overall, insufficient evidence has been presented to conclude that a Glu/Asp hydrogen bond to the Tyr $-$OH group significantly increases the rate of disordering of the Tyr side-chain. It appears that this analysis is limited by the suitability of aromatic ring solvent accessible area as an implicit measure for steric hindrance. For example, considering the two most susceptible Tyr $-$OH groups of myrosinase, despite Section 6.4.3 providing evidence that these Tyr side-groups can reorientate, following loss of interacting Glu $-CO_2$ moieties, a negligible change in solvent accessible area was detected for Tyr-330, and only a small increase of 1.36 Å^2 was observed for Tyr-215.

6.6.2 Dependence on disulphide bond damage

It could also be suggested that a Tyr residue in close proximity to a disulphide bond may undergo enhanced radiation-induced disordering as a by-product of disulphide bond cleavage and resulting local changes to protein backbone conformation. A custom script was written in *Python* to calculate the closest distance between each Tyr O_η atom and the Cys S_γ atom in each disulphide bond. However, for the subset of disulphide containing proteins (Figure 6.23), no significant correlation was detected between the Tyr $-$OH D_{loss} metric and disulphide bond proximity (Å). There is thus insufficient evidence to suggest that this is a contributing factor.

6.7 Conclusion

This chapter has aimed to investigate the phenomenon of radiation-induced electron density loss to the $-$OH group of tyrosine residues in MX experiments at 100 K. Whereas previous reports (e.g. Burmeister (2000) on myrosinase) have subsequently been construed by the subsequent radiation damage literature as a clear indication of cleavage of the phenolic C$-$O linkage, this is highly unsupported by a large body of radiation chemistry studies on both aqueous and solid

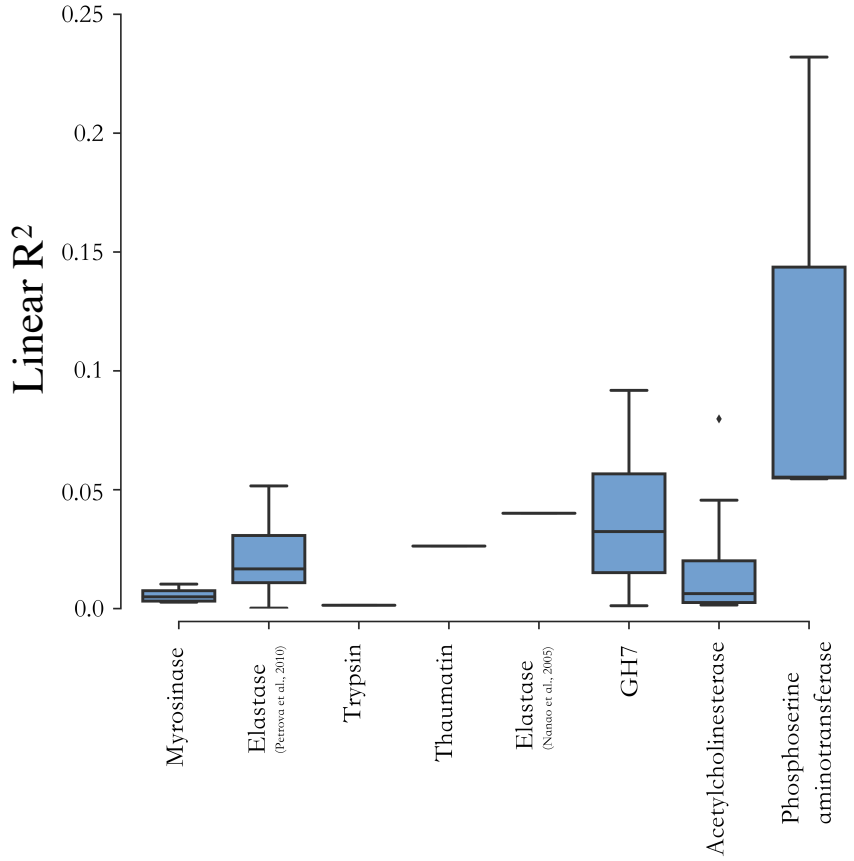


Figure 6.23: Linear R^2 coefficients calculated for each MX damage series for correlation between Tyr −OH D_{loss} metric values and the distance to the nearest disulphide bond. R^2 values have been calculated for each dose in a damage series, and Tukey box plots represent distribution in R^2 across these multiple datasets. Whiskers represent 1.5× the interquartile range for each box plot. Only structures containing eight or more Tyr residues per asymmetric unit have been included. The y-axis only extends to $R^2 = 0.25$, and thus all reported R^2 values indicate poor correlation.

state protein samples, in which the tyrosyl radical produced following deprotonation of the oxidised residue is deemed to form a highly stable redox couple with tyrosine.

Re-investigation of the original myrosinase MX damage series data independently reproduced the observations of Tyr-330 –OH group electron density loss in $F_{\text{obs,n}} - F_{\text{obs,1}}$ difference maps. It was determined that in the Burmeister (2000) damage study, rigid body and overall anisotropic *B*-factor refinement (but no positional refinement) was performed starting from a 1.2 Å myrosinase model instead derived from a different crystal used for another experiment (Burmeister *et al.*, 2000). Whereas it could be suggested that such Tyr damage observations were an artefact of the aforementioned sub-optimal refinement strategy, even with full re-refinement and model re-building with *PDB_REDO* (resulting in a significant increase in map quality), negative difference peaks surrounding the Tyr-330 –OH group were still detectable above noise in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps. Furthermore, since *PDB_REDO* re-refinement also provided an improved fit to $2mF_{\text{obs}} - DF_{\text{calc}}$ map density for several Tyr aromatic rings (e.g. Tyr-215), pronounced negative difference density (comparable with that at Tyr-330) also became interpretable around additional Tyr groups; a finding not originally reported by Burmeister (2000).

The next aim of the chapter was to determine the prevalence of such observations of Tyr –OH electron density loss across general protein structures. The available MX damage series deposited in the PDB gave a wide range of independently refined coordinate models corresponding to proteins of varying size ($5.7 - 204$ kDa) that had all been collected at 100 K, each with unique nuances of buffer composition (potential free radical scavengers) and macromolecular composition (monomeric or multimeric protein, presence of nucleic acid, glycans and metal cofactors). The D_{loss} per-atom quantification of radiation-induced electron density loss, as developed previously in Section 2.5, permitted objective and high-throughput investigation of Tyr –OH density loss events across these proteins, without the necessity to manually inspect $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps.

No restriction on resolution or crystallisation cocktail composition was enforced in the current sampling, and, accordingly, no attempt has been made in the current chapter to quantitatively compare raw D_{loss} values between proteins. However, as a testament to the strength of the D_{loss} metric, clear trends have emerged nonetheless, with radiation-induced Glu and Asp decarboxylation highlighted to be prevalent across the range of protein systems in our study at doses > 1 MGy. Prominent Tyr –OH density loss did not appear to be a widespread feature of general

protein structures, a finding that is consistent with the high energetic barrier for Tyr phenolic C−O bond scission. It is thus strongly suggested that tyrosine should not be generally quoted with Cys, Glu, Met and Asp residues as a radiation-sensitive locus in proteins at 100 K.

However, observations of pronounced Tyr −OH susceptibility have been observed (e.g. Tyr-215 in myrosinase and Tyr-62 in *Trp* RNA-binding attenuation protein, as reported in Section 4.3.5). The next aim of this chapter was to provide an energetically feasible model for such reports. To this end, Section 6.4 has proposed a simple model in which Tyr −OH density loss is a result of radiation-induced conformational change of the Tyr group, without the necessity of bond cleavage. Indeed, evidence has been presented to suggest that Tyr and Phe aromatic rings undergo comparable positional displacement with increasing dose, and Tyr O_η electron density loss is correlated with density loss for covalently bound C_ζ (for the structures myrosinase and TRAP for which high Tyr −OH D_{loss}(atom) values have been reported here). This model has further been reinforced through occupancy refinement of multiple Tyr side-chain conformations for selected residues in myrosinase; for radiation-sensitive Tyr groups, as the occupancy of the original conformation decreased with dose, the population of alternative positions was determined to increase.

There has been anecdotal evidence to suggest a relationship between radiation-induced Tyr −OH electron density disordering and the presence of hydrogen bond interactions to Glu/Asp carboxyl side-groups (Burmeister, 2000). Indeed, this was observed for hydrogen bonded Tyr residues in the large TRAP complex that was previously solved in Chapter 4, and has also been reported in this chapter for Tyr-215, Tyr-330 and Tyr-390 in myrosinase. This hypothesis is highly consistent with the model of radiation-induced Tyr reorientation, since upon X-ray irradiation, decarboxylation of the nearby Glu/Asp would disrupt the hydrogen bond that stabilises the Tyr group. As shown in Section 6.6, whereas select Tyr residues forming hydrogen bond interactions to proximal Glu/Asp carboxylates exhibited comparably high D_{loss} metric values (relative to non hydrogen bond-forming Tyr −OH groups), this behaviour was restricted to a subset of proteins (myrosinase, TRAP, thermolysin, insulin), and was not conserved amongst all such Tyr residues for these proteins. It is suggested that a multitude of compounding factors associated with a hydrogen bonding interaction must play a role in the density loss rate for each individual Tyr −OH group (e.g. readiness of Glu/Asp decarboxylation, lability of the Tyr aromatic group, steric hindrance both prior and post

decarboxylation). No quantitative correlation could be established for these individual factors, although it is suggested that this could be due to limitations in the parametrisation of such factors.

CHAPTER 7

Further methods for site-specific damage detection

7.1 Introduction

Chapter 2 of this thesis has presented an objective tool, *RIDL*, designed to quantify radiation-induced electron density loss of individual atoms in MX structures. The tool derives metric values from $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps calculated over the crystal asymmetric unit, and therefore requires the collection of multiple disjoint sets of diffraction images (1,...,n) from a single crystal, each of high completeness. The work in the current chapter will present two applications of the *RIDL* program.

7.1.1 Specific damage half-doses

For MX experiments on protein crystals at 100 K, an experimentally-determined dose of 30 MGy has been suggested as an appropriate upper limit, D_{ln2}, for experiments (Owen *et al.*, 2006). At this dose, it is proposed that for any protein crystal, the mean diffraction intensity, I_n, will have decayed to at least 70 % of the initially recorded value, I_1, and the biological information in the crystal likely to be compromised. However, unfortunately the D_{ln2} limit for cryocooled experiments has often been misinterpreted as a *safe dose* limit below which site-specific damage will be avoided (Nogly *et al.*, 2015).

X-ray induced changes to redox centres have been monitored by complementary techniques (e.g. UV-Vis absorption spectroscopy) and have been detectable in X-irradiated protein crystals at doses on the order of just tens of kGy at 100 K (Owen *et al.*, 2011) and 65 K (Meharenna *et al.*, 2010). For photosensitive proteins, spectroscopy can be used in conjunction with X-ray irradiation of a sacrificial crystal to determine a safe dose for X-ray crystallographic data collection, below which the yield of structural changes is insufficient to contribute to the final electron density map (Oghbaey *et al.*, 2016; Owen *et al.*, 2017). These calculations assume that a structural change of less than 20 % cannot significantly affect the electron density (Pearson *et al.*, 2007). In good accordance with this, the lowest dose reported for a radiation-induced structural change detected in an electron density map is higher than for UV-Vis spectral changes, at 0.06 MGy for covalently bound retinal in bacteriorhodopsin at 100 K (Borshchevskiy *et al.*, 2014).

The half-life dose values for the yield of particular oxidised species to decay to 50 % have been

determined for the cases* of copper nitrate reductase (1.5 MGy using UV-Vis and XAS (Hough *et al.*, 2008)), putidaredoxin (0.3 MGy using XAS (Corbett *et al.*, 2007)) and the haem group in met-myoglobin (0.21 MGy using UV-Vis (Beitlich *et al.*, 2007)). However, for standard MX data collection on general protein crystals at 100 K, advisory dose-limits to evade significant site-specific events to Cys, Glu and Asp residues are still lacking. Consequently, an aim of the work reported in this chapter has been to develop a protocol to model the decay in electron density as a function of dose for individual atoms, and thus calculate half-dose values, $D_{1/2}^{\text{specific}}$ for individual radiation susceptible species.

7.1.2 Specific damage detection in single MX datasets

The vast majority of MX data collections intent on structural solution do not typically collect a sufficient quantity of diffraction data to form multiple disjoint datasets of high completeness. Consequently, it is unclear how to readily generate $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps to screen for radiation-sensitive hotspots in proteins. The work described in this chapter will also demonstrate how a limited quantity of diffraction images can still be utilised to generate suitable Fourier difference maps in which damage signals can be detectable. As such, the yield of site-specific damage events accumulating over the course of the collection of a single set of diffraction images can be detected in electron density using *RIDL*. The aim of this work will be to provide a proof of principle, and it is hoped that such screening for site-specific damage becomes an eventual requirement associated with coordinate deposition in the PDB.

7.2 Specific damage half dose model

This section will introduce a protocol to determine half doses for electron density decay of radiation susceptible sites. In order to sufficiently characterise density decay with dose, multiple disjoint datasets of diffraction images with known dose values must be available, in addition to a refined coordinate model. These criteria significantly limited the suitability of most MX damage series deposited in the PDB for this study.

*all conducted at 100 K

The protocol described below has been applied to several test protein candidates, each of varying size and data resolution:

- **The GH7 family protein**. This MX damage series was reported previously in Section 2.3, where it was used to demonstrate the behaviour of the *PEAKS* and *RIDL* programs developed in Chapter 2. Due to its large size (96.8 kDa per asymmetric unit) and prevalence of radiation susceptible residues (Cys: 36, Glu: 28, Asp: 74, Met: 22), GH7 was deemed a suitable test system for the current investigation. Eleven disjoint diffraction datasets had been collected to an initial high resolution of 2 Å over a large dose range (DWD: $1.1 - 22.7$ MGy).
- ***Trp* RNA-binding attenuation protein (TRAP)**. This MX damage series was reported in Chapter 4 for the purpose of characterising site-specific damage in a model protein-RNA complex. This protein contains 220 Glu and Asp residues per asymmetric unit (however no Cys or ordered Met residues), and was therefore deemed a suitable test system to characterise the electron density decay of acidic residues. Ten disjoint diffraction datasets had been collected to an initial high resolution of 2 Å over a large dose range (DWD: $1.3 - 25.0$ MGy).

The primary aim has been to develop a suitable protocol for half-dose calculation. In order to rigorously establish strict lower bound half-dose values, a larger sample of distinct protein damage series would be required. Due to time restrictions, this has been left for future work.

7.2.1 The protocol

7.2.1.1 Per-atom density quantification using *RIDL*

The program *RIDL* provides a method to directly sample electron density localised around each refined atom in an input coordinate model. In the work described in previous chapters, *RIDL* has exclusively used $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ Fourier difference maps to detect regions of electron density change with respect to dose; however due to the modular architecture of the *RIDL* program, it was straightforward to instead generate $2\mathrm{m}F_{\mathrm{obs}} - \mathrm{D}F_{\mathrm{calc}}$ density maps using *RIDL*. The local

magnitude of $2mF_{obs} - DF_{calc}$ density map at each atom could then be monitored as a function of increasing dose, using an identical procedure as for $F_{obs,n} - F_{obs,1}$ maps described in Section 2.5.5, with the use of the CCP4 program *SFALL* to assign a localised search radius around each atom. For a detailed discussion on this per-atom density sampling, see that previous section.

Whereas in theory it would be possible to instead extract half-dose values based from $F_{obs,n} - F_{obs,1}$ maps, it is simpler here to consider the decay in the actual electron density (as provided by a $2mF_{obs} - DF_{calc}$ map) as opposed to the difference density at each dose. In particular, as is presented later in Section 7.2.1.5, the derivation of a physically-reasonable *model* for electron density decay with dose (and resulting formulation of the half-dose) is more easily perceived in terms of overall electron density for each atom at each dose.

In this section, unless otherwise specified, $2mF_{obs} - DF_{calc}$ map coefficients have been calculated for each dataset n in an MX damage series using the CCP4 program *SIGMAA*, using the observed structure factor amplitudes $|\mathbf{F}_{obs,n}|$ and calculated phases $\phi_{calc,n}$ and amplitudes $|\mathbf{F}_{calc,n}|$ derived from the refined coordinate model for dataset n (using the program *SFALL*). The refinement strategies used to generate the coordinate model for each dataset have been described previously in Section 2.3.3.1 and 4.2.3 (for the GH7 protein and TRAP, respectively) and has not been repeated here.

The standard per-atom metrics derived by *RIDL* (e.g. D_{loss}) were designed for the purpose of suitably quantifying electron density change around an atom site. For a $2mF_{obs} - DF_{calc}$ map, at each dose, the magnitude of electron density at each atomic site can similarly be quantified. In this report, the maximum and median density values in the vicinity of each atom, D_{max} and D_{median}, have been defined. Note that the definition of D_{max} is identical to that of D_{gain} for $F_{obs,n} - F_{obs,1}$ difference maps, as in Section 2.5.6:

$$D_{max}(\text{a,dose}) = \max_{v \in V_a}[\rho(v)] \tag{7.2.1}$$

$$D_{median}(\text{a,dose}) = \text{median}_{v \in V_a}[\rho(v)] \tag{7.2.2}$$

where V_a is the set of voxels assigned to an atom a, and $\rho(v)$ is the $2mF_{obs} - DF_{calc}$ electron density value calculated at voxel position v. D_{max} and D_{median} are simple indicators that were deemed suitable for the current demonstration of the protocol. Future work would aim to es-

tablish improved metrics to quantify local $2mF_{obs} - DF_{calc}$ behaviour. As discussed previously in Section 2.5.6, it is important to note that any integration of density values around an atom (either to provide the sum of density or mean value) is highly dependent on the assigned search radius and the number of voxels allocated to a particular atom. Consequently, the current analysis has refrained from the use of such a metric.

It is noted that the D_{max} and D_{median} metrics exhibit qualitatively very similar behaviour with dose throughput this section, however, both metrics have been presented to demonstrate that the eventual half-dose values computed using these two metrics are consistent (see Figure 7.8).

For each atom, the resulting metric values can be plotted as a function of dose. As anticipated, for radiation sensitive groups (e.g. Cys-S_γ, Glu-C_δ), metric values decay with increasing dose (Figure 7.1). For TRAP, metric values for non-damaged atoms (Glu-C_α and Asp-C_α) were also observed to decrease in with increasing dose (Figure 7.1 a and b). This is suggested to be a manifestation of global radiation damage, due to the deteriorating diffraction data quality with increasing dose. Interestingly, as shown in Figure 7.1c-d, metric values exhibited non-monotonic behaviour, with values observed to eventually increase at high dose values (> 15 MGy). This behaviour is not believed to be a true feature of the physical crystal, and will be examined further in the next section.

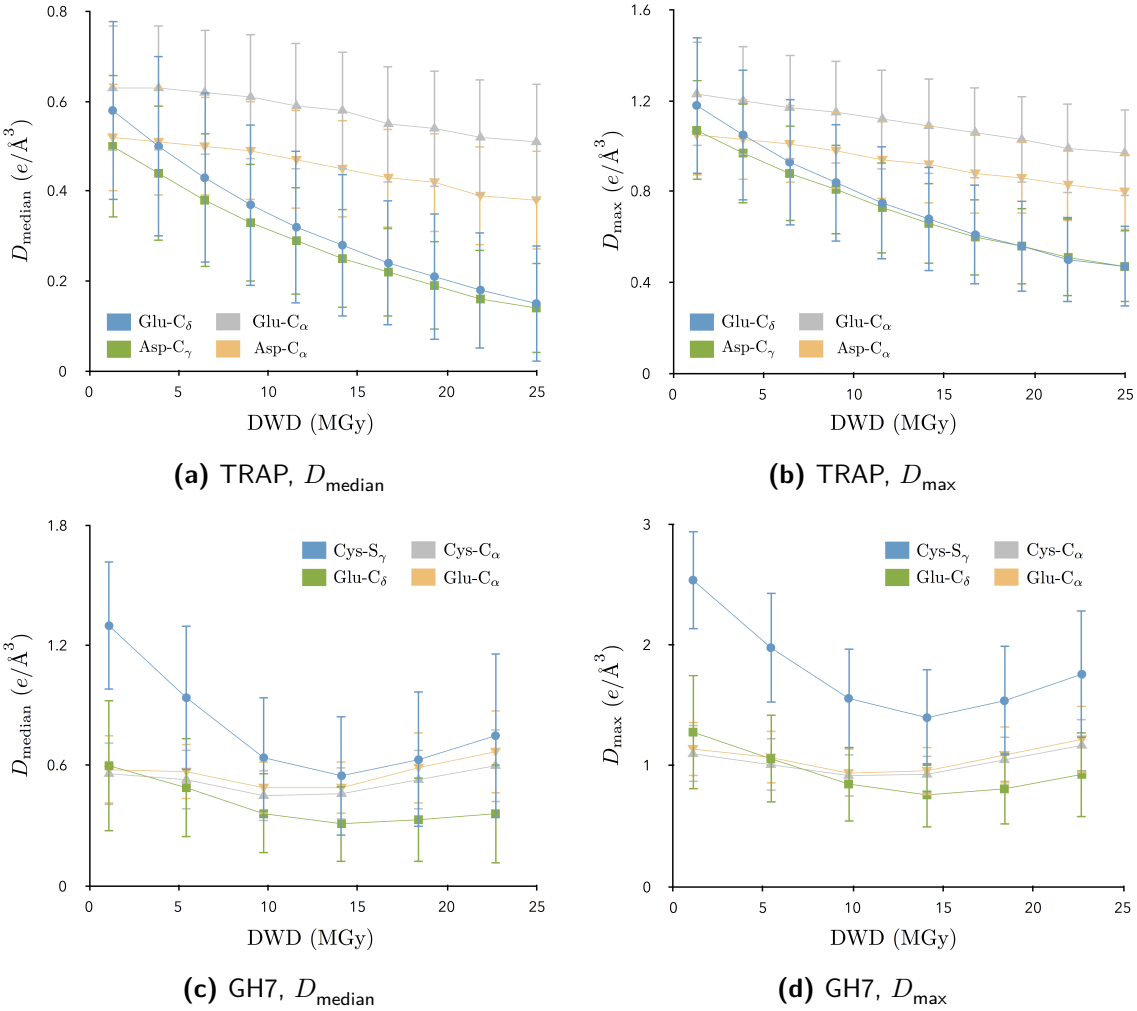
7.2.1.2 Discussion on suitability of $2mF_{obs} - DF_{calc}$ maps

A standard $2mF_{obs} - DF_{calc}$ map calculated by *FFT* is scaled such that the mean map voxel value is zero. This is due to the reflection corresponding to the $\mathbf{F}_{000}$ structure factor being unrecorded during data collection since it coincides directly with the beam stop centre. It can be demonstrated that the integral of the general electron density equation:

$$\rho(\mathbf{x}) = \frac{1}{V} \sum_{\mathbf{h}} \mathbf{F}(\mathbf{h}) e^{-2\pi i \mathbf{h}\cdot\mathbf{x}} \tag{7.2.3}$$

with respect to $\mathbf{x}$ over the unit cell dimensions is equal to $\mathbf{F}_{000}$[†], since all other trigonometric terms cancel. Therefore, in the absence of $\mathbf{F}_{000}$, the average value of the $\rho(\mathbf{x})$ map over the unit cell must be zero. Thus individual three-dimensional voxels in the $2mF_{obs} - DF_{calc}$ map can

[†]i.e. $\mathbf{F}_{000}$ is equal to the summed number of electrons in the unit cell.



(a) TRAP, D_{median} **(b)** TRAP, D_{max}

(c) GH7, D_{median} **(d)** GH7, D_{max}

Figure 7.1: The decay in $2mF_{obs} - DF_{calc}$ map electron density localised around specific atoms for the test cases of (a,b) TRAP and (c,d) the GH7 family protein. In (a) and (c) the local electron density in each $2mF_{obs} - DF_{calc}$ map has been quantified using the D_{median} metric. In (b) and (d), the D_{max} metric has been used to quantify local $2mF_{obs} - DF_{calc}$ map density around atoms. In each plot, points represent the average over multiple instances of an atom type, with error bars indicating 1 standard deviation.

take negative values[‡]. As a direct consequence, it is unclear from the plots in Figure 7.1 to what y-axis value the electron density should decay in the limit as dose $d \rightarrow \infty$. In the upcoming Section 7.2.1.5, a discussion on a physically reasonable y-axis limit will be presented in more detail. It is suggested here that the non-monotonic behaviour in Figure 7.1c-d could also be attributed to the enforced zero mean value for the $2mF_{obs} - DF_{calc}$ map at each dataset. In Section 7.2.1.3, an alternative map type (an END map) will be tested to attempt to circumvent this issue.

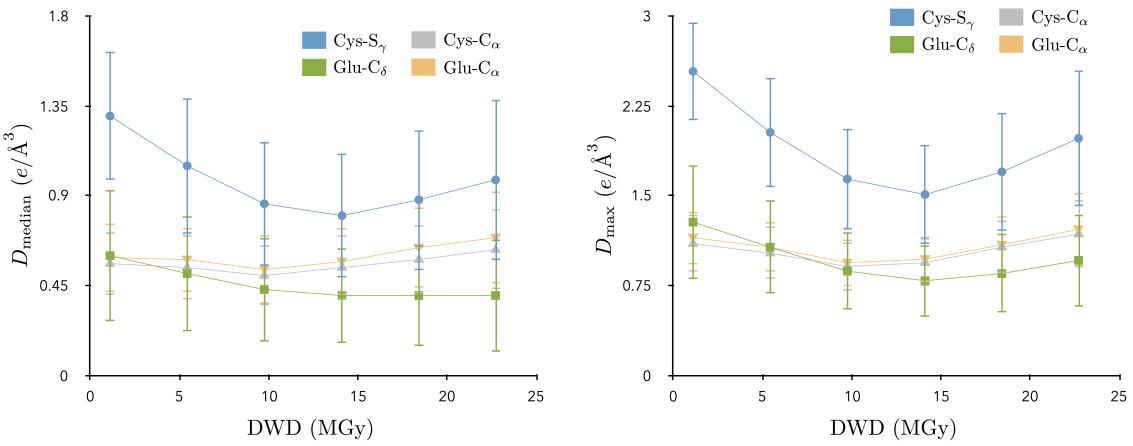
It should also be noted that in the current strategy employed here to generate $2mF_{obs} - DF_{calc}$ maps at each dose (see Section 7.2.1.1), a distinct set of $|\mathbf{F}_{calc,n}|$ and $\phi_{calc,n}$ terms has been used at each dataset n. In particular, $|\mathbf{F}_{obs,n}|$ has been scaled to $|\mathbf{F}_{calc,n}|$ independently for each dataset, yet never directly to $|\mathbf{F}_{calc,1}|$. To test whether the non-monotonic behaviour for the GH7 protein in Figure 7.1c-d was an artefact of the current strategy, $2mF_{obs} - DF_{calc}$ maps were generated at each dataset n, now instead using the fixed $|\mathbf{F}_{calc,1}|$ and $\phi_{calc,1}$ set corresponding to the first coordinate model throughout (with only $|\mathbf{F}_{obs,n}|$ being variable between datasets). Since non-monotonic $2mF_{obs} - DF_{calc}$ map density decay persisted when using these new maps (Figure 7.2), the observed behaviour did not appear to depend on the chosen map calculation strategy.

7.2.1.3 Testing the usefulness of END maps

It has previously demonstrated that the $\mathbf{F}_{000}$ structure factor value can be estimated when there is a $2mF_{obs} - DF_{calc}$ map and corresponding coordinate model (Lang *et al.*, 2014). An 'electron number density' (END) map can then be generated to translate and scale the $2mF_{obs} - DF_{calc}$ map such that each voxel attains a value strictly with units of absolute electrons per $Å^3$, and zero now approximately corresponds to the vacuum level of $0\,e/Å^3$. Figure 7.3 illustrates the protocol to generate an END map.

In the current investigation, the effect of monitoring electron density decay using END maps was tested. An END map could be generated from the $2mF_{obs} - DF_{calc}$ map at each dose, using a

[‡]A fortunate caveat of the $2mF_{obs} - DF_{calc}$ map attaining a zero mean is that in model building software such as *COOT*, only positive $2mF_{obs} - DF_{calc}$ electron density corresponding to electron-dense regions is visible by default.



(a) GH7, D_{median} **(b)** GH7, D_{max}

Figure 7.2: The decay in 2mF_{obs} − DF_{calc} map electron density localised around specific atoms for the test case of the GH7 family protein. Here, 2mF_{obs} − DF_{calc} maps have been calculated in an alternative strategy to the data presented in Figure 7.1, in which the fixed $|\mathbf{F}_{\text{calc},1}|$ and $\phi_{\text{calc},1}$ set corresponding to the first coordinate model have been used to generate all 2mF_{obs} − DF_{calc} maps for each dataset n. The local electron density in each 2mF_{obs} − DF_{calc} map has been quantified using the (a) D_{median} and (b) D_{max} metrics. In each plot, points represent the average over multiple instances of an atom type, with error bars indicating 1 standard deviation.

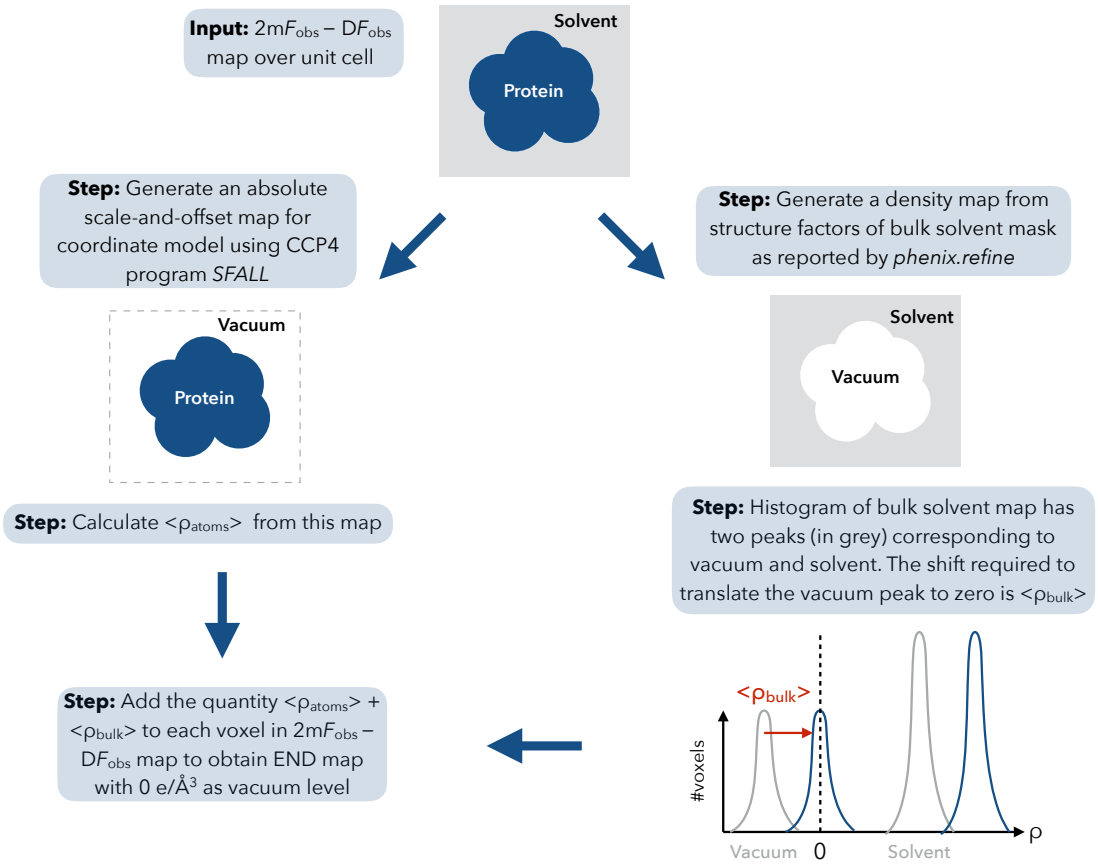
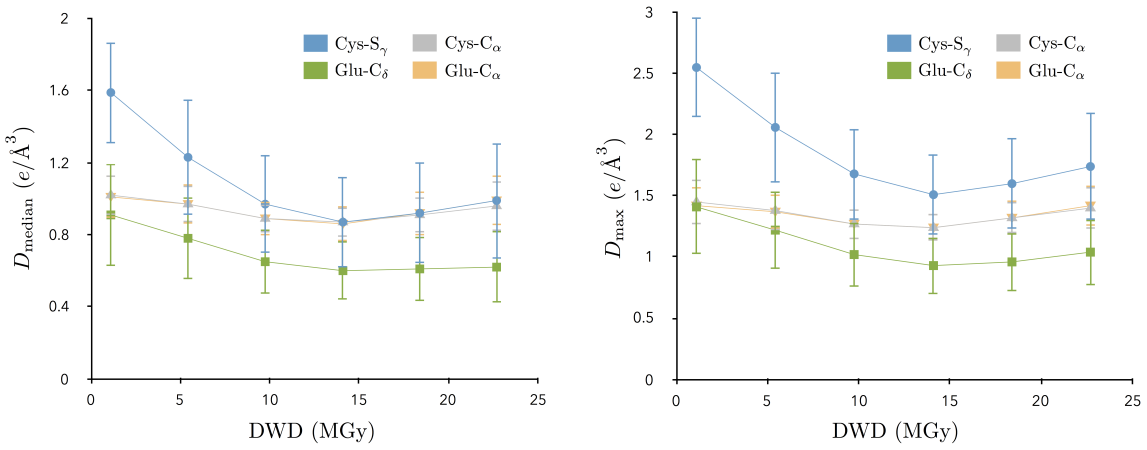


Figure 7.3: A schematic illustrating the steps performed to generate an END map from a 2mF_{obs} − DF_{calc} map (Lang *et al.*, 2014). The aim of the protocol is to calculate the average electron density values $\langle\rho_{atom}\rangle$ and $\langle\rho_{bulk}\rangle$ corresponding to the coordinate model and bulk solvent regions in the crystal, from which the total number of electrons, $\mathbf{F}_{000}$, can be derived. This protocol is performed by the *END_RAPID.com* script available at *http://bl831.als.lbl.gov/END/RAPID/*.



(a) GH7, D_{median} **(b)** GH7, D_{max}

Figure 7.4: The decay in END map electron density localised around specific atoms for the example case of the GH7 family protein. The local electron density in each END map has been quantified using the (a) D_{median} and (b) D_{max} metrics. In each plot, points represent the average over multiple instances of an atom type, with error bars indicating 1 standard deviation. Metric behaviour is qualitatively similar to Figure 7.1c-d.

custom script in *RIDL* to call the external END map shellscript written by Lang *et al.* (2014)[§]. Again due to the modular architecture of *RIDL*, the END map could be efficiently substituted for the $F_{\text{obs,n}} - F_{\text{obs,1}}$ map for the calculation of per-atom metric values. The runtime for the END map calculation was slow (approximately 30 mins per $2mF_{\text{obs}} - DF_{\text{calc}}$ map for the 96.8 kDa GH7 family protein asymmetric unit, running on a 2 GHz Intel Core i5 processor).

The effect of substituting each $2mF_{\text{obs}} - DF_{\text{calc}}$ map with an END map is demonstrated in Figure 7.4. For the GH7 family protein, non-monotonic behaviour for both the D_{median} and D_{max} metrics were still observed (similarly to Figure 7.1c-d), and an inter-map scaling issue appeared to persist. Whereas END maps aim to place all $2mF_{\text{obs}} - DF_{\text{calc}}$ maps on an absolute scale with 0 e/Å^3 corresponding to the vacuum density level, this provides only an approximation, and overall did not appear successful for all test cases.

Furthermore, it cannot be assumed that for a radiation-susceptible species, the end state of a damage event is a hard vacuum of 0 e/Å^3 (J. Holton, private communications). Indeed, clearly seen from Figure 7.4, electron density for atoms does not decay to 0 e/Å^3 with increasing dose. Thus the additional benefit of using END maps instead of $2mF_{\text{obs}} - DF_{\text{calc}}$ maps was concluded to be limited.

[§]The full script is available at *http://bl831.als.lbl.gov/END/RAPID/*. The *END_RAPID.com* main script required manual modification in order to correctly function on the computer operating system (macOS Sierra) of the current author, due to Linux specific commands that had been implemented in the original script.

7.2.1.4 The requirement for metric C_α-normalisation

In order to circumvent the above scale inconsistency between maps, metric values were next normalised to a radiation-insensitive standard, which is assumed not to undergo significant density decay with increasing dose. As presented previously in Section 2.5.8, the *RIDL* program already provides the functionality to perform per-atom metric normalisation to the metric values attained by the protein backbone C_α atoms:

$$D_{\text{max}}^{C_\alpha}(\text{atom,dose}) = \frac{D_{\text{max}}(\text{atom,dose}) - \langle D_{\text{max}}(\text{a,dose})\rangle_{a\in C_\alpha}}{\sigma_{a\in C_\alpha}[D_{\text{max}}(\text{a,dose})]} \tag{7.2.4}$$

where $\langle D_{\text{max}}(\text{a,dose})\rangle_{a\in C_\alpha}$ represents the mean D_{max} metric value calculated over all backbone C_α atoms. A similar equation holds for the case of the D_{median} metric defined above in Section 7.2.1.1.

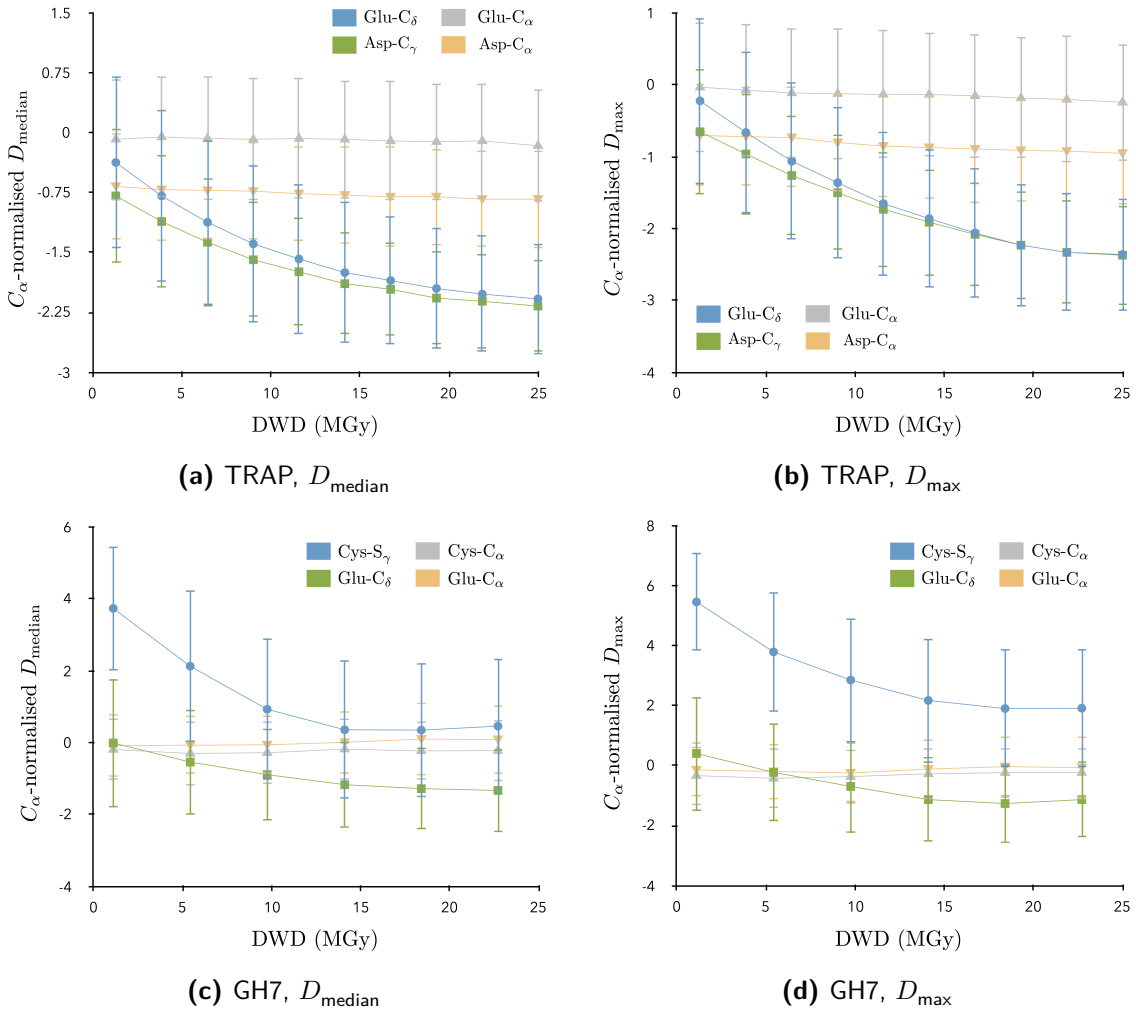
In effect, C_α-normalisation also removes any benefit from exploiting END maps to place each $2mF_{\text{obs}} - DF_{\text{calc}}$ map on an absolute scale with 0 e/Å^3 corresponding to the vaccuum density level. For a C_α-normalised metric, at any dose, zero now corresponds to the average metric value attained by the set of C_α atoms. In the following analysis, density $2mF_{\text{obs}} - DF_{\text{calc}}$ maps have therefore been used exclusively to monitor electron density decay, and all y-axis density decay metrics are C_α-normalised. Figure 7.5 illustrates the effect of C_α-normalisation for the example cases of the GH7 family protein and TRAP. For radiation-insusceptible species (Glu-C_α, Asp-C_α and Cys-C_α), negligible change in mean D_{median} and D_{max} metric values was observed with increasing dose.

7.2.1.5 A model for electron density decay

In this section, the magnitude of electron density at each atomic centre is assumed to follow a first order decay process:

$$\text{X}_{\text{undamaged}} \xrightarrow{k} \text{X}_{\text{damaged}} \tag{7.2.5}$$

where $X_{\text{undamaged}} \geq 0$ is the quantity of non-damaged instances of a given atom across all units cells that contribute to electron density, and $X_{\text{damaged}} \geq 0$ is the non-contributing number of damaged sites. k is the rate constant for the exponential decay of undamaged sites with dose, d.



(a) TRAP, D_{median} **(b)** TRAP, D_{max}

(c) GH7, D_{median} **(d)** GH7, D_{max}

Figure 7.5: The decay in $2mF_{obs} - DF_{calc}$ map electron density localised around specific atoms for the test cases of (a,b) TRAP and (c,d) the GH7 family protein. In (a) and (c) the local electron density in each $2mF_{obs} - DF_{calc}$ map has been quantified using a C_α-normalised version of the D_{median} metric. In (b) and (d), the C_α-normalised D_{max} metric has been used to quantify local $2mF_{obs} - DF_{calc}$ map density around atoms. In each plot, points represent the average over multiple instances of an atom type, with error bars indicating 1 standard deviation. For non-damaged atoms (Glu-C_α, Asp-C_α and Cys-C_α), no change in metric values was observed with increasing dose.

This leads to a generalised solution of the form:

$$X_{\text{undamaged}}(d) = a\exp(-k\cdot d) + b \tag{7.2.6}$$

where $a > 0$ and $b \geq 0$ are atom-specific constants such that for a given atom, the initial number of undamaged atom sites (at $d = 0\,\text{MGy}$) across all unit cell is $X^{\text{init}}_{\text{undamaged}} = a + b$.

Under the assumption that the magnitude of electron density, y, at an atom site, as measured by either the C_α-normalised D_{median} or D_{max} metrics, is a linear function[¶] of the number of undamaged sites, $X_{\text{undamaged}}$, y can also be expressed in the form:

$$y(d) = \alpha\exp(-k\cdot d) + \beta \tag{7.2.7}$$

where α and β are constants such that the initial magnitude of electron density is $y_0 = \alpha + \beta$. In contrast to $b \geq 0$ in Equation 7.2.6, it is noted that β in Equation 7.2.7 is not necessarily positive, since the C_α-normalised D_{median} and D_{max} metrics can both take negative values, as demonstrated later in Figure 7.6 and 7.7.

The decay model in Equation 7.2.7 was deemed to be a suitable description for radiation-sensitive atoms, based from the observations displayed in Figure 7.5. Such a first order model holds under the assumption that the rate of electron density decay for a given atom across all unit cells of the crystal depends only on the current state of that atom. Effects from other surrounding atoms have therefore been assumed to be negligible. In all subsequent plots, the model in Equation 7.2.7 has been fit to data using the *curve_fit* function (employing a non-linear least squares fitting approach) as part of the *SciPy* package in *Python*.

The magnitude of electron density, y_∞, in the asymptotic limit as the dose d tends to infinity, is β. As discussed in the sections above, a suitable value for β has been non-trivial to justify. In order to determine a suitable β, two strategies were tested:

Strategy 1

[¶]Since the electron density, y, at an atom site has been measured using C_α-normalised metrics generated by *RIDL*, which can take negative values, y is strictly a linear function of the number of undamaged sites, $X_{\text{undamaged}}$, as opposed to directly proportional to it. As such, it is assumed that constants $\tilde{a} > 0$ and $\tilde{b} \in \mathbb{R}$ such that $y = \tilde{a}X_{\text{undamaged}} + \tilde{b}$.

The first strategy assumes that a *global* end electron density state is eventually attained for any decaying atom ($= \beta_{\text{global}}$), which is set to the average bulk solvent density value. In principle this value could be determined by computing an additional solvent mask (using the CCP4 program *MAPMASK*) and then calculating the average density of voxels that had been assigned to solvent, $\langle\rho\rangle_{solvent}$. However, in order to avoid the additional computational time to do this, $\langle\rho\rangle_{solvent}$ could be calculated using the equation:

$$\langle\rho\rangle_{solvent} = \frac{\langle\rho\rangle_{overall} \times N_{overall} - \langle\rho\rangle_{protein} \times N_{protein}}{N_{overall} - N_{protein}} \tag{7.2.8}$$

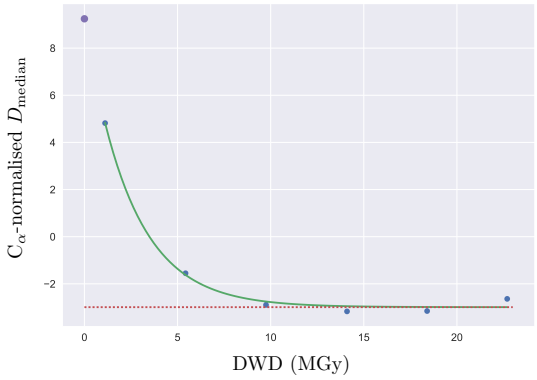
since the values $\langle\rho\rangle_{overall}$ and $\langle\rho\rangle_{protein}$, corresponding to the average density over the whole and protein components respectively, as well as the corresponding total number of voxels $N_{overall}$ and $N_{protein}$, were readily available already in *RIDL*.

Whereas the plot in Figure 7.6a demonstrates that this strategy did appear suitable for a subset of atoms, there were many cases where the density decayed to values other than this limit were observed (Figure 7.6 b and c). Consequently, assigning the β_{global} parameter as the average bulk solvent density value was deemed to be suboptimal.

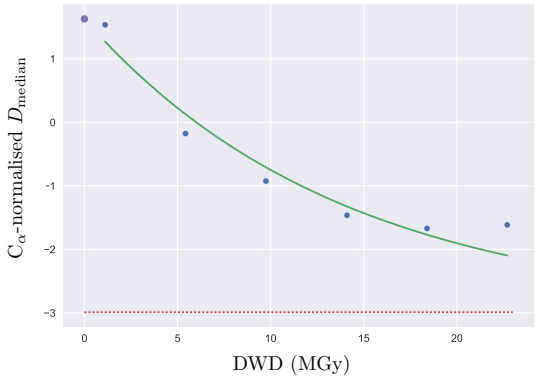
An alternative *global* end electron density state could instead be envisaged to be a '*protein crystal soup*', in which the electron density is entirely homogeneous throughout the unit cell at a value of $\beta_{\text{global}} = \frac{F_{000}}{V_{\text{cell}}}$, where F_{000} is the total number of electrons in the unit cell volume V_{cell}. However, similarly to the approach used above (and indeed true for any fixed global end density state), this limit will not be satisfied for all tested atoms (data not shown).

Strategy 2

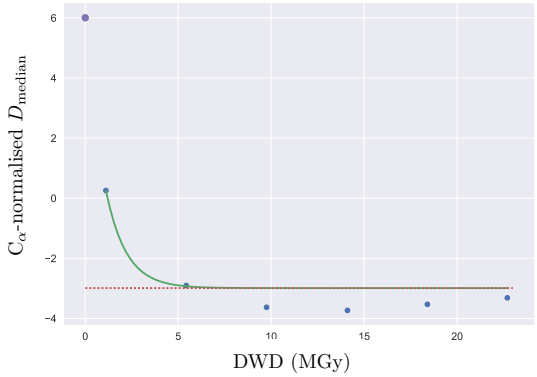
The second strategy entailed allowing the β parameter in Equation 7.2.6 to become an extra parameter to be fit for each atom ($= \beta_{\text{var}}$), so that the end density state for an atom is not fixed, but attains a value that best fits the data. To motivate an underlying physical justification for why β_{var} can vary between different atoms, the following extension of the decay model in



(a) A-Cys-364-S_γ



(b) B-Cys-162-S_γ



(c) B-Cys-364-S_γ

Figure 7.6: The decay in the C_α-normalised D_{median} metric as a function of dose for example atoms selected for the GH7 family protein. An exponential decay function of the form shown in Equation 7.2.7 has been fitted to data (blue points), and the fitted function has been shown in green. In all plots, the horizontal dotted line (in red) represents the average of C_α-normalised D_{median} values attained by the bulk solvent region of the unit cell, and it has been assumed here that this corresponds to the β constant in Equation 7.2.7. The purple point indicates the predicted initial density magnitude, predicted following fitting. Atoms have been chosen to specifically indicate cases where (a) a good fit was produced, and two examples where a poor fit was observed, where the data appeared to decay to above/below the bulk solvent level in (b) and (c) respectively. Clearly, in (c), since the β parameter in Equation 7.2.7 was constrained to the bulk solvent level (red horizontal line), only a poor fit to data was achievable.

Equation 7.2.5 has been necessary:

$$\begin{aligned} \mathrm{X_0} &\xrightarrow{\mathrm{k_1}} \mathrm{X_1} \\ \mathrm{X_0} &\xrightarrow{\mathrm{k_2}} \mathrm{X_2} \end{aligned} \tag{7.2.9}$$

Here, for a given atom, the undamaged instances of that atom across all unit cells, $X_0 \geq 0$, can undergo two competitive damage events: atomic disordering at rate $k_1 > 0$ and full cleavage/removal at rate $k_2 \geq 0$. These satisfy the rate equations:

$$\begin{aligned} \frac{dX_0}{dx} &= -(k_1 + k_2)X_0 \\ \frac{dX_1}{dx} &= k_1 X_0 \\ \frac{dX_2}{dx} &= k_2 X_0 \end{aligned} \tag{7.2.10}$$

with solutions:

$$X_0 = X_0^{\mathsf{init}} \exp(-(k_1 + k_2)x) \tag{7.2.11}$$

$$X_1 = \frac{k_1}{k_1 + k_2} X_0^{\mathsf{init}} (1 - \exp(-(k_1 + k_2)x)) + X_1^{\mathsf{init}} \tag{7.2.12}$$

$$X_2 = \frac{k_2}{k_1 + k_2} X_0^{\mathsf{init}} (1 - \exp(-(k_1 + k_2)x)) \tag{7.2.13}$$

Here the assumption has been made that the initial number of atoms in the undamaged state, X_0^{init}, and in the disordered state X_1^{init} are non-zero. No cleavage events are assumed to initially be present ($X_2^{\mathsf{init}} = 0$).

Equation 7.2.11 indicates that the undamaged fraction, X_0/X_0^{init}, will decay to zero as the dose $x \to \infty$. However, if it is also assumed that the disordered X_1 state can also contribute to the observable electron density at an atomic site, albeit at a reduced weighting $w < 1$ relative to the X_0 state, then the number of atoms, X_{contr}, contributing to the observable electron density, at a

dose d is:

$$X_{\text{contr}}(d) = X_0(d) + w \cdot X_1(d) \tag{7.2.14}$$

$$= X_0^{\text{init}} \exp(-(k_1+k_2)d) + w \cdot \frac{k_1}{k_1+k_2} X_0^{\text{init}}(1-\exp(-(k_1+k_2)d)) + w \cdot X_1^{\text{init}} \tag{7.2.15}$$

$$= w \cdot X_1^{\text{init}} + w\frac{k_1}{k_1+k_2} X_0^{\text{init}} + (X_0^{\text{init}} - w \cdot \frac{k_1}{k_1+k_2} X_0^{\text{init}}) \exp(-(k_1+k_2)d)) \tag{7.2.16}$$

$$= b + a\exp(-kd) \tag{7.2.17}$$

with $b = X_1^{\text{init}} + w \cdot \frac{k_1}{k_1+k_2} X_0^{\text{init}}$, $a = (X_0^{\text{init}} - w \cdot \frac{k_1}{k_1+k_2} X_0^{\text{init}})$, and $k = k_1+k_2$. As in Equation 7.2.6, a and b must again both clearly be positive here.

Again under the assumption that the observable electron density, y, as measured by either the C_α-normalised D_{median} or D_{max} metrics, is a linear function of the number of contributing atoms X_{contr}, there also exist constants $\tilde{a} > 0$ and $\tilde{b} \in \mathbb{R}$ such that $y(d) = \tilde{a} \cdot X_{\text{contr}}(d) + \tilde{b}$. Thus:

$$y(d) = \beta_{\text{var}} + \alpha\exp(-kd) \tag{7.2.18}$$

holds with $\alpha = a \cdot \tilde{a}$ and $\beta_{\text{var}} = \tilde{a} \cdot b + \tilde{b}$. This is an identical form to Equation 7.2.7 above, as required. The expression for $\beta_{\text{var}} \in \mathbb{R}$ here is dependent on the initial states X_0^{init} and X_1^{init} as well as the atom-specific decay constants k_1 and k_2; it is therefore justifiable that β_{var} can attain different values for each atom.

The data in Figure 7.7 illustrate the effect of fitting the β_{var} parameter for the example cases of the GH7 family protein and TRAP. Note that whereas explicit expressions for α and β_{var} can be derived above in terms of the unknowns $(w, X_0^{\text{init}}, X_1^{\text{init}}, k_1, k_2)$, only a three parameter fitting involving $(\alpha, \beta_{\text{var}}, k)$ has been performed for the data shown in Figure 7.7 to avoid overfitting.

For the majority of tested Cys-S_γ, Glu-C_δ and Glu-$O_{\epsilon 1}$ atoms, reasonable exponential fits could be produced (Figure 7.7a-d). Figure 7.7d provides an example of a radiation-insensitive Glu-$O_{\epsilon 1}$ atom in TRAP that undergoes only minor electron density decay. However, instances were also present when only poor fits were obtained (Figure 7.7e-f). In Figure 7.7f, the negative y-axis scale indicates that negligible $2mF_{\text{obs}} - DF_{\text{calc}}$ density was originally detectable at the B-Glu-84-C_δ site,

and it was thus unsurprising that only a poor fit was observed.

Objective criteria for assessing fit quality has been presented in the next section.

7.2.1.6 Half dose determination

For the decay model presented in Equation 7.2.10, even though the $2mF_{obs} - DF_{calc}$ map density at an atom site includes contributions from both X_0 and X_1 (as shown in Equation 7.2.14), it is possible to easily extract a half-dose value for the number of atoms in the ordered state $X_0 \geq 0$ to decay to halfway between its initial value, X_0^{init} and zero:

$$d_{1/2} = \frac{\ln(2)}{k} = \frac{\ln(2)}{k_1 + k_2} \tag{7.2.19}$$

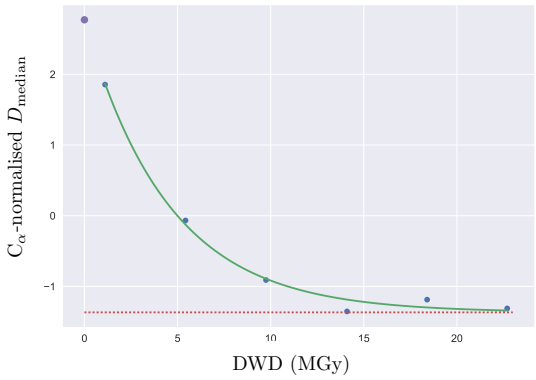
This is the standard half-dose for a first order decay process and is independent of the parameter β_{var}.

In terms of the observed electron density in Figure 7.7, this is also the dose required to reduce each reported metric (C_α-normalised D_{median} or D_{max}) to halfway between the initial metric value at zero dose, $y_0 = \alpha + \beta_{var}$, and the high dose limit $y_\infty = \beta_{var}$. It is important to note that this is not the dose required to reduce the overall number of atoms contributing to the electron density, $X_{contr} \geq 0$ to 50 % of its original value, which would strictly be calculated as:

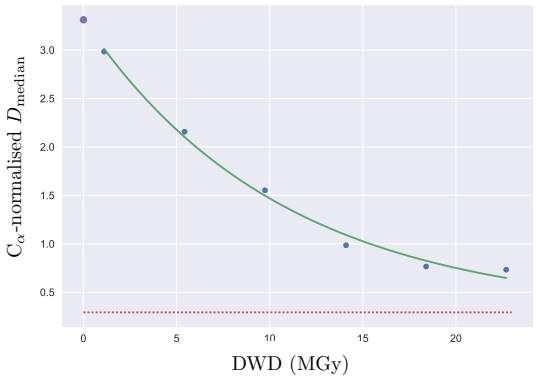
$$d_{1/2}^{X_0 + w \cdot X_1} = \frac{1}{k} \ln(\frac{2a}{a-b}) \tag{7.2.20}$$

where $a > 0$ and $b \geq 0$ are constants as described previously in Equation 7.2.17. The half dose in Equation 7.2.20 does not provide a suitable indication of the radiation sensitivity of each atom, and indeed is undefined when $b > a$ (which can occur if the initial number of disordered sites X_1^{init} is large enough). The $d_{1/2}$ in Equation 7.2.19 will subsequently referred to as the *half-dose* value for an atom in the following analysis.

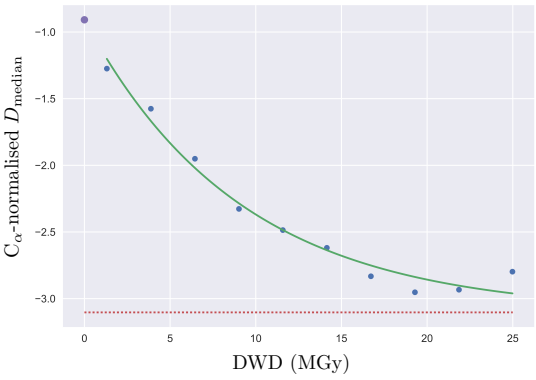
For each susceptible atom type in each test protein, half dose values could be calculated following the exponential fitting illustrated in Figure 7.7. Since it was observed that a subset of atoms exhibited poor fits (with examples provided in Figure 7.7 e and f), half-dose values have been



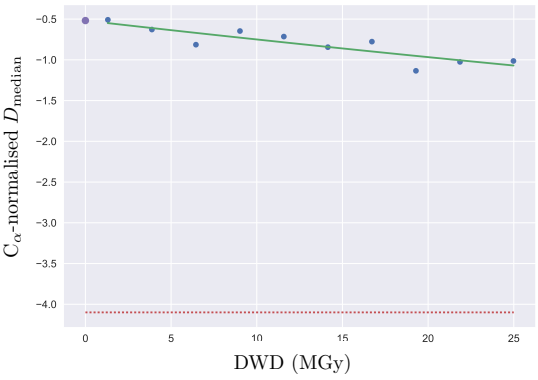
(a) GH7, A-Cys-364-S_γ



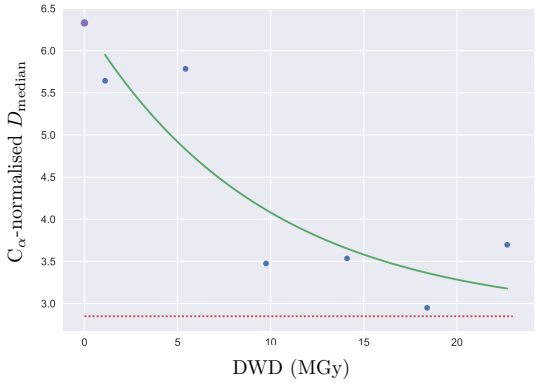
(b) GH7, B-Glu-236-C_δ



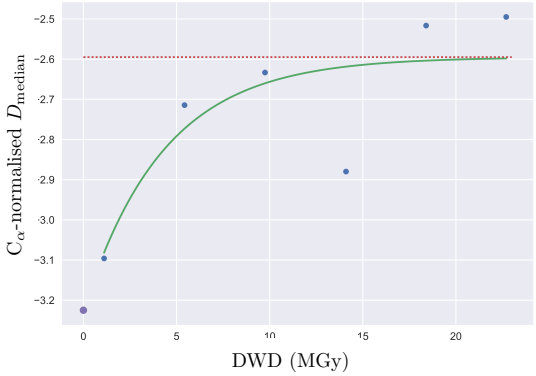
(c) TRAP, B-Glu-36-$O_{\epsilon 1}$



(d) TRAP, R-Glu-36-$O_{\epsilon 1}$



(e) GH7, B-Cys-234-S_γ



(f) GH7, B-Glu-84-C_δ

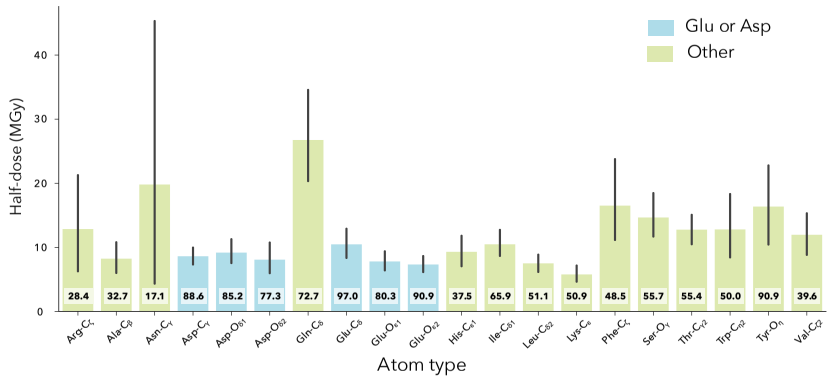
Figure 7.7: The decay in the C_α-normalised D_{median} metric as a function of dose for example atoms selected for the GH7 family protein and TRAP. An exponential decay function of the form shown in Equation 7.2.17 has been fit to data (blue points), and the fitted function has been shown in green. The end density state (β in Equation 7.2.17) has been fit for each atom, as indicated by the horizontal dotted line (in red). The purple point indicates the predicted initial density magnitude, predicted following fitting. Atoms have been chosen to specifically indicate cases where (a-d) a good fit was produced, and (e-f) two examples where a poor fit was observed.

computed only atoms for which reliable fits were produced. The assessment of fitting quality was performed objectively using the criteria:

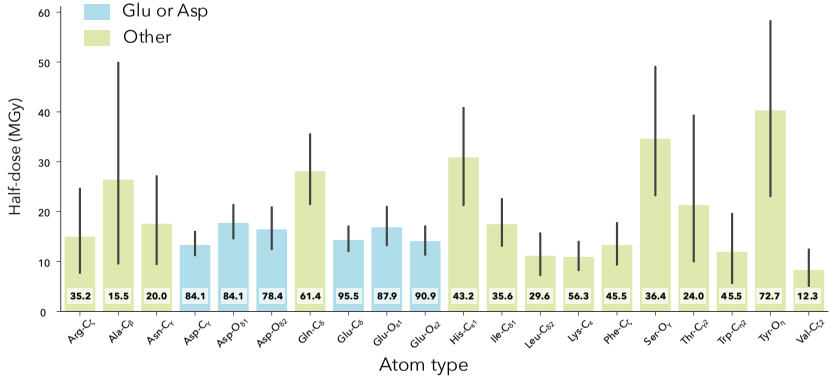
- The α parameter must exceed 0
- The rate constant k must exceed 0
- The mean of the *relative standard deviation* (RSD) of each fitted parameter (α, β, k) must be below 1. For each parameter, RSD was defined as the standard deviation for the fitted parameter divided by the fitted parameter value. The threshold of 1 was chosen arbitrarily to exclude cases where parameters had been fit with high error. Visual inspection of a subset of plots indicated that this threshold was suitable for the current analysis.

The data shown in Figure 7.8 report the average half dose values calculated for selected atoms types across TRAP and the GH7 family protein, as calculated using both the D_{median} and D_{max} metrics. Terminal side-group carbon atoms for each residue type have been included, in addition to the Glu-$O_{\epsilon 1}$, Glu-$O_{\epsilon 2}$, Asp-O_{δ}, Asp-O_{δ}, Met-S_{δ} and Cys-S_{γ} atoms, which are known to be radiation-sensitive. Values have only been included for individual atoms that pass the selection criteria above: for each atom type, the percentage of selected atoms out of all these atoms in each structure has also been included in Figure 7.8. As expected, consistently a larger percentage of Asp, Glu, Met and Cys atoms exhibit density decay, relative to other atom types.

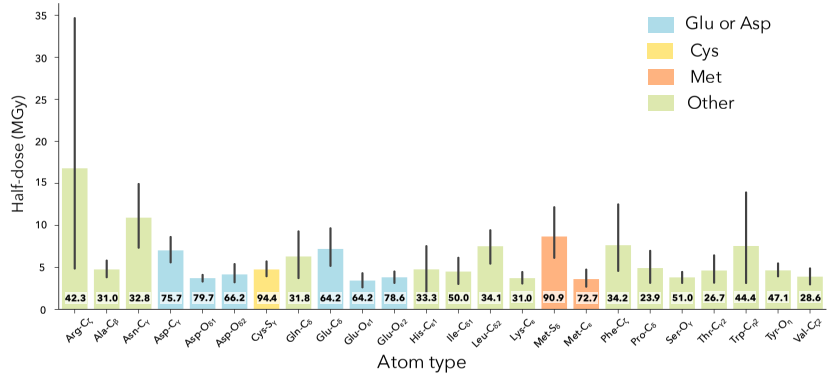
Overall the data exhibited in Figure 7.8 provide evidence that for the TRAP and GH7 proteins, a diffraction-weighted dose of $\approx$ 10 MGy (as a conservative estimate) is sufficient, on average, reduce the electron density around Glu, Asp and Met side-chain termini atoms, as well as Cys disulphides, to halfway between the initial density state (at 0 MGy) and high dose limit (as dose $d \rightarrow \infty$). It must be note that the current section has focused on the development of a protocol to calculate half-dose values across proteins, and has only demonstrated preliminary results on two test protein systems. However, the extension of the current protocol to additional proteins would be necessary to determine whether this half-dose value persists to a wider range of protein crystals at 100 K.



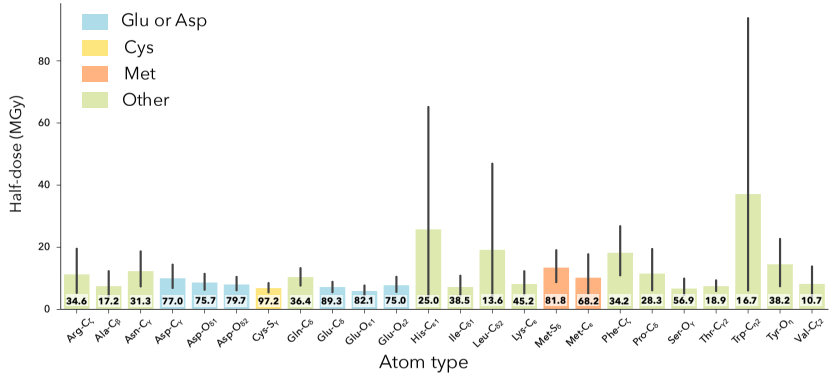
(a) TRAP, D_{median}



(b) TRAP, D_{max}



(c) GH7, D_{median}



(d) GH7, D_{max}

Figure 7.8: Half-dose values for $2mF_{obs} - DF_{calc}$ map electron density to decay to halfway between the initial density state (at 0 MGy) and high dose limit (as dose $d \to \infty$). Values have been reported for a selection of side-chain termini atoms in (a,b) TRAP and (c,d) the GH7 family protein, for both the D_{median} and D_{max} metrics. Each bar represents the half-dose calculated over only atoms satisfying the selection criteria presented in Section 7.2.1.6, in order to exclude atoms for which a reliable exponential decay function (Equation 7.2.19) fit to data could not be achieved. For each atom type, the percentage of the total number of atoms deemed suitable have been specified (shown for each bar). Error bars represent 1 standard deviation in half-dose values for each atom type. Half-dose values were calculated using Equation 7.2.19.

7.2.1.7 Do half dose values infer the susceptibility of atoms?

Interestingly, whereas the values in Figure 7.8 indicate that residues that are known to be radiation-susceptible (Asp, Glu, Cys, Met) on average exhibit low half-dose values (typically $\lesssim 10\,\mathrm{MGy}$), terminal side-groups of multiple other residues also exhibit low half-dose values, albeit calculated for consistently smaller fractions of atoms for which reliable fittings could be produced (typically $\lesssim 60\,\%$).

It appears that half-dose values, on their own, do not distinguish the most prominent sites of radiation damage in a protein, as have been previously characterised in the radiation damage literature (e.g. Weik *et al.* (2000); Fioravanti *et al.* (2007); Garman (2010)) and in other chapters of this thesis. In particular, as illustrated in Figure 7.9, two atoms A and B can have the same half-dose value, yet exhibit a significantly different overall decay in electron density with increasing dose. In the current example, the different y-axis intercepts for atoms A and B have been a result of allowing the initial ratio of ordered to disordered instances of the atom throughout the crystal, $X_0^{\mathrm{init}}/X_1^{\mathrm{init}}$, to differ for A and B. The ratio $X_0^{\mathrm{init}}/X_1^{\mathrm{init}}$ directly affects the α parameter in Equation 7.2.18[‖], but not the resulting half-dose in Equation 7.2.19.

In Figure 7.9, intuitively atom A should be considered to be a more significant damage site since it has lost a greater magnitude of electron density than atom B at each dose value. The initial rate of $2\mathrm{m}F_{\mathrm{obs}} - \mathrm{D}F_{\mathrm{calc}}$ density decay:

$$\left.\frac{dy}{dx}\right|_{x=0} = -\alpha \cdot k \tag{7.2.21}$$

provides a measure of the rate at which electron density is lost by an atom, however unlike $d_{1/2}$, it also accounts for each atom not being in a fully ordered state prior to irradiation (i.e. a variable $X_0^{\mathrm{init}}/X_1^{\mathrm{init}}$ ratio between atoms). As shown in Figure 7.10, the initial decay rate is consistently higher for Asp, Glu, Cys and Met groups.

In light of this, it is suggested that the reported half-doses in Figure 7.8 provide evidence that, beside Asp, Glu, Cys and Met, other side-groups (e.g. Lys-C_ϵ in TRAP and Ile-$C_{\delta 1}$ in the GH7

[‖] This can be shown explicitly from Strategy 2 in Section 7.2.1.5 previously, since α in Equation 7.2.18 is proportional to a in Equation 7.2.17, which itself takes the explicit form: $a = (X_0^{\mathrm{init}} - w \cdot \frac{k_1}{k_1+k_2} X_0^{\mathrm{init}})$. The value X_1^{init} is implicit in the parametrisation of a, since the total number of sites $X_{\mathrm{total}} = X_0^{\mathrm{init}} + X_1^{\mathrm{init}}$ is fixed.

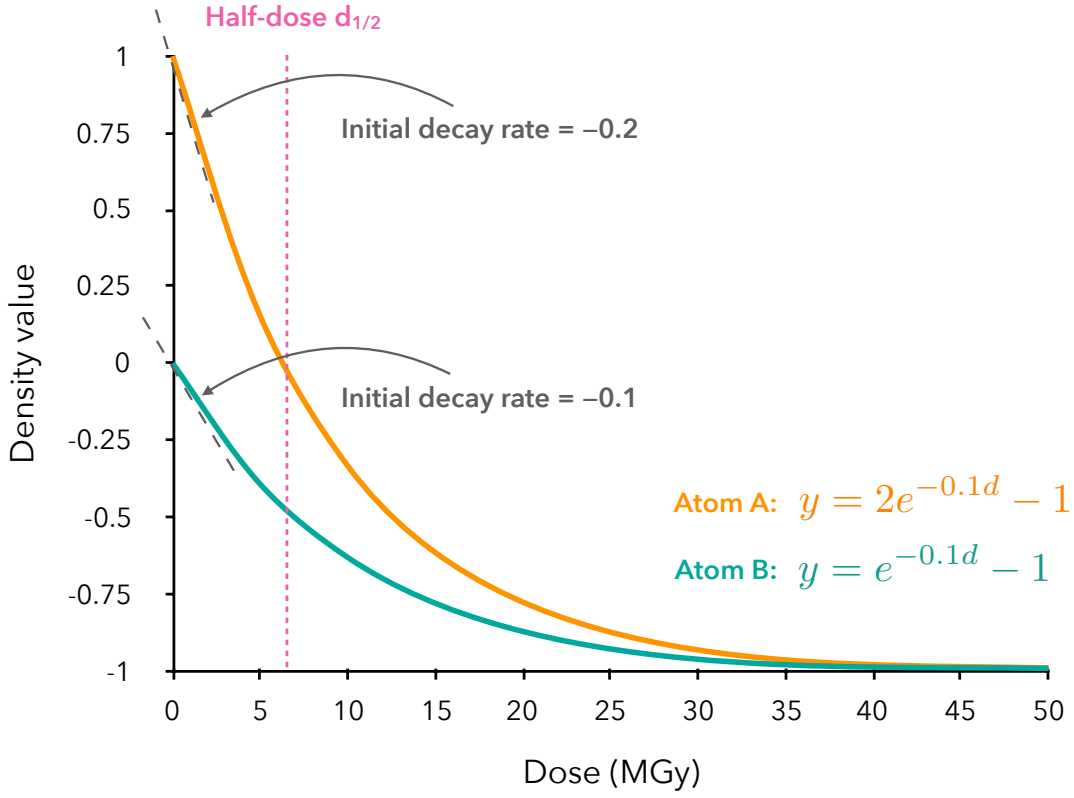
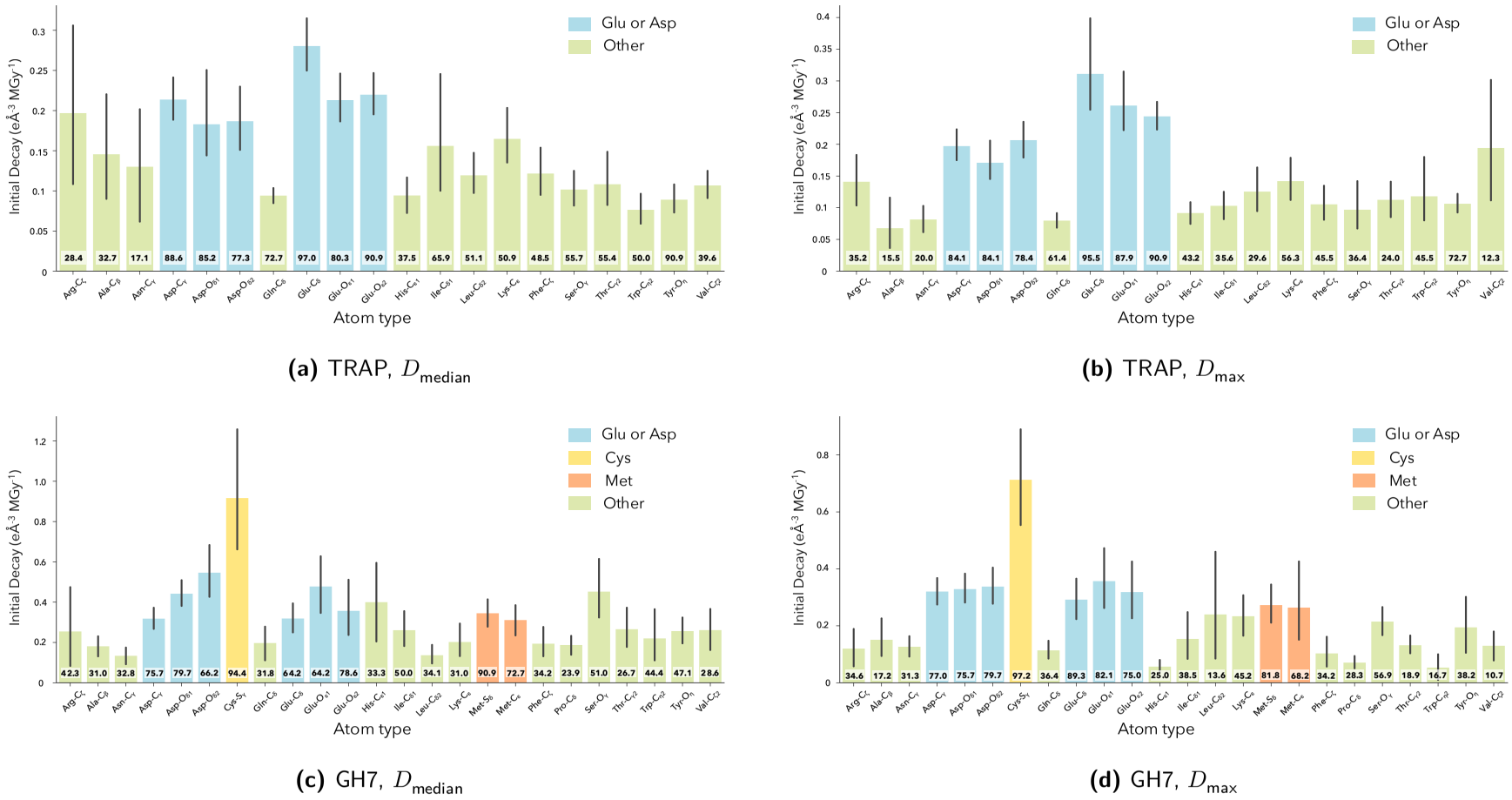


Figure 7.9: A simulation of electron density decay (y-axis) against dose (x-axis) for two hypothetical atoms A (in orange) and B (in teal). The α, β and k parameters in the decay equation $y(d) = \alpha e^{-kd} + \beta$ for each atom have been arbitrarily chosen and do not relate to any specific known atoms in either TRAP or the GH7 protein. Despite the different decay equations for A and B, both atoms attain the same half-dose value (in pink). However, the initial rate of density decay for atom A is shown to be two-fold that of atom B.



(a) TRAP, D_{median} **(b)** TRAP, D_{max}

(c) GH7, D_{median} **(d)** GH7, D_{max}

Figure 7.10: Initial decay rate, $\frac{dy}{dx}\big|_{x=0}$ (Equation 7.2.21), values for $2mF_{obs} - DF_{calc}$ map electron density to decay to halfway between the initial density state (at 0 MGy) and high dose limit (as dose $d \to \infty$). Values have been reported for a selection of side-chain termini atoms in (a,b) TRAP and (c,d) the GH7 family protein, for both the D_{median} and D_{max} metrics. Each bar represents the half-dose calculated over only atoms satisfying the selection criteria presented in Section 7.2.1.6, in order to exclude atoms for which a reliable exponential decay function (Equation 7.2.19) fit to data could not be achieved. For each atom type, the percentage of the total number of atoms deemed suitable have been specified (shown for each bar). Error bars represent 1 standard deviation in half-dose values for each atom type. Initial decay rate values were calculated using Equation 7.2.21.

protein) also exhibit rapid electron density loss due to radiation-induced disordering. But also that these effects are less pronounced in $2mF_{obs} - DF_{calc}$ maps (or indeed in $F_{obs,n} - F_{obs,1}$ maps), since the overall the magnitude of loss (as quantified by the initial rate of decay) is much lower than for Asp, Glu, Cys and Met groups.

7.3 How much diffraction data are required to detect site-specific damage?

The work presented in the following section aims to determine how much MX diffraction data are required for site-specific radiation damage events to become detectable in electron density maps. This section will provide a proof of principle that $F_{obs,n} - F_{obs,1}$ maps can be used in conjunction with the *RIDL* pipeline (originally presented in Chapter 2) to screen for site-specific changes to local electron density at atoms throughout the collection of a single MX dataset of diffraction images. These images would typically all be merged to generate a set of structure factor amplitudes, to be used in subsequent phasing, model building and refinement. The method developed here could ultimately provide a tool for MX experimenters to validate whether their MX datasets are devoid from all radiation-induced structural alterations that could mislead the experimenter during model building into electron density.

7.3.1 The test proteins

The protocol described in this section will be demonstrated for two distinct MX damage series, however only images corresponding to a single complete MX dataset will be required for each:

- **Bovine insulin**. This MX damage series was collected on cubic bovine insulin crystals (5.8 kDa), and was supplied by Dr. J. C. Brooks-Bartlett in the form of refined coordinate models and corresponding merged structure factor amplitudes. In total, ten disjoint diffraction datasets had been collected at 100 K on the PETRA III Hamburg beamline P14 using an X-ray energy of 12.7 keV, and a collimated 140 µm × 140 µm beam with 'top hat' beam profile. Thus the $\sim$ 100 µm × 100 µm × 100 µm cubic crystal was continuously and uniformly

bathed in the beam throughout data collection. Each dataset consisted of 90° of rotation, with an oscillation step of 0.1° per image and exposure time of 0.5 s per image. The crystal initially diffracted to a high resolution (1.38 Å) with a reported DWD of 0.89 MGy for the first dataset, as calculated by *RADDOSE-3D*. Full data processing statistics were reported in the D.Phil. thesis of J. C. Brooks-Bartlett, Chapter 2, and have been omitted from this section. See Appendix A of the current thesis for details.

- **GH7 family protein**. The full MX data collection information for this dataset is presented in Section 2.3. Relative to insulin, GH7 is a large (96.8 kDa) protein case study containing a large number of radiation susceptible residues (Cys: 36, Glu: 28, Asp: 74, Met: 22). The first dataset corresponds to 180° of rotation, with an angular step of 0.2° per image and exposure time of 0.04 s per image. In contrast to insulin, the $75 \times 75 \times 10\,\mu m^3$ crystal was non-homogeneously irradiated in an approximately Gaussian profile beam (FWHM: $17.7 \times 104.8\,\mu m^2$, vertical $\times$ horizontal).

The insulin dataset was chosen for the current analysis primarily since the aforementioned data collection strategy had ensured a homogeneous dose distribution throughout the crystal during X-irradiation. Consequently, any effect of variation of exposed crystal volume on diffraction data quality throughout the collection of a single dataset was minimised. Furthermore, the insulin crystal formed in space group $I2_13$, such that significantly less than 90° of diffraction images were required for high ($> 90\,\%$) diffraction data completeness. The current analysis will aim to determine whether there is any significant benefit in restricting to less than 90° to evade site-specific damage effects.

In contrast to insulin, the GH7 family protein dataset has been presented to provide a more challenging case study, in which the crystal has been heterogeneously irradiated in a Gaussian-shaped beam profile. It will be demonstrated that this data collection strategy significantly hinders detection of site-specific damage, although detection is still achievable.

7.3.2 A protocol to screen for damage in a single MX dataset

This section will develop an explicit computational protocol to screen for the onset of site-specific damage in a single set of MX diffraction images.

7.3.2.1 Definition of sub-datasets D_1 and D_2

Given a set of images ($i = 1, ...n$) forming a single complete MX dataset, define the sub-datasets as D_1 and D_2 as:

$$D_1 : \ i = 1, ..., p, \ \text{ such that } 1 \leq p < n \tag{7.3.1}$$

$$D_2 : \ i = 1, ..., q, \ \text{ such that } 1 < q \leq n \tag{7.3.2}$$

with $p < q$.

Reflection intensities integrated from image sets D_1 and D_2 could be scaled and merged separately in *AIMLESS* and converted to structure factor amplitudes with *CTRUNCATE*, as part of the data reduction pipeline of the CCP4i GUI. In an identical protocol to that described previously in Section 2.5.4, a Fourier difference map between image set D_1 and D_2, $F_{obs,D_2} - F_{obs,D_1}$, could then be generated. Here, a phase set was used corresponding to the coordinate model refined against the observed structure factor amplitudes derived using all n images.

Figure 7.11 provides an example $F_{obs,D_2} - F_{obs,D_1}$ map for insulin with (a) $p = 240$, $q = 900$ and (b) $p = 240$, $q = 440$. It is demonstrated that sites of disulphide bond cleavage were present, even when sub-datasets D_1 and D_2 overlap. The accumulation of positive density in the solvent region in close proximity to each sulphur provided convincing evidence that these were true damage events, with each cysteine sulphur having reorientated following bond rupture.

7.3.2.2 Systematic variation of D_1 and D_2

In principle, by varying images pertaining to sub-datasets D_1 and D_2, and then analysing each corresponding $F_{obs,D_2} - F_{obs,D_1}$ map, the opportunity to detect any site-specific damage occurring

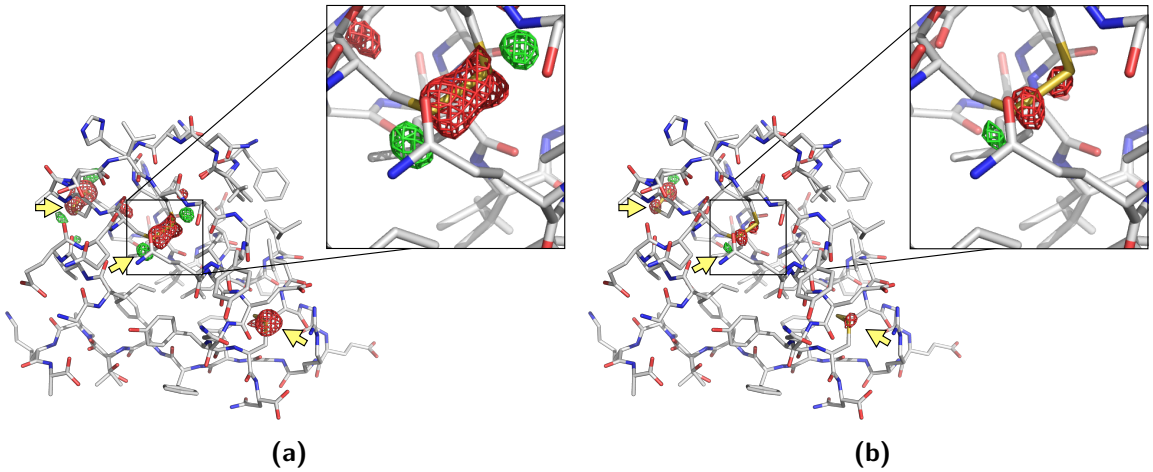


Figure 7.11: $F_{\text{obs,D}_2} - F_{\text{obs,D}_1}$ Fourier difference maps generated from a single MX dataset for insulin. In (a), $p = 240$ and $q = 900$, such that D_1 corresponds to images $1, ..., 240$ and D_2 corresponds to images $1, ..., 900$. In (b), $p = 240$ and $q = 440$. All maps are contoured at $\pm 5\sigma$ in green/red. Arrows (in yellow) indicate the three sites of disulphide bond cleavage in insulin, as detected in each difference map.

across the total image set will be maximised. A custom Python script was written to systematically vary the number of images corresponding to sub-datasets D_1 and D_2, and for each selected sub-dataset pair (D_1,D_2), independently perform data merging using *AIMLESS* and *CTRUNCATE*. Figure 7.12 provides a schematic to illustrate this process.

The end images p and q for D_1 and D_2 were varied between 1 and n, whilst retaining the $p < q$ criterion. This procedure ensures that sub-dataset D_2 always contains later images than D_1. This procedure will be referred to as a '*sliding image window*' in the subsequent analysis. By fixing the sliding image window for both D_1 and D_2 to start from the first image ($i = 1$), the maximum number of possible (D_1,D_2) image pairings becomes $\sum_{i=1}^{n-1} n - i$. It is noted that other image division schemes are possible. However, the current strategy was chosen since (a) it minimises the number of *AIMLESS* processing jobs that must be performed to a maximum of n**, and (b) it reduces the analysis to two dimensions (see section below††).

In practice, since it is considerably time-intensive to generate merged `.mtz` files for all possible (D_1,D_2) pairings, a coarseness parameter c has been specified, such that the above sliding window

**Since all sub-datasets, whether D_1 or D_2, all start from image number 1, a maximum of n unique `.mtz` files (containing merged structure factor amplitudes) need to be generated.

††The heat map analysis presented in Figure 7.13 is performed by assigning p and q as the y and x axes, respectively. A more complex division of images to generate D_1 and D_2 may require analysis in more than two-dimensions.

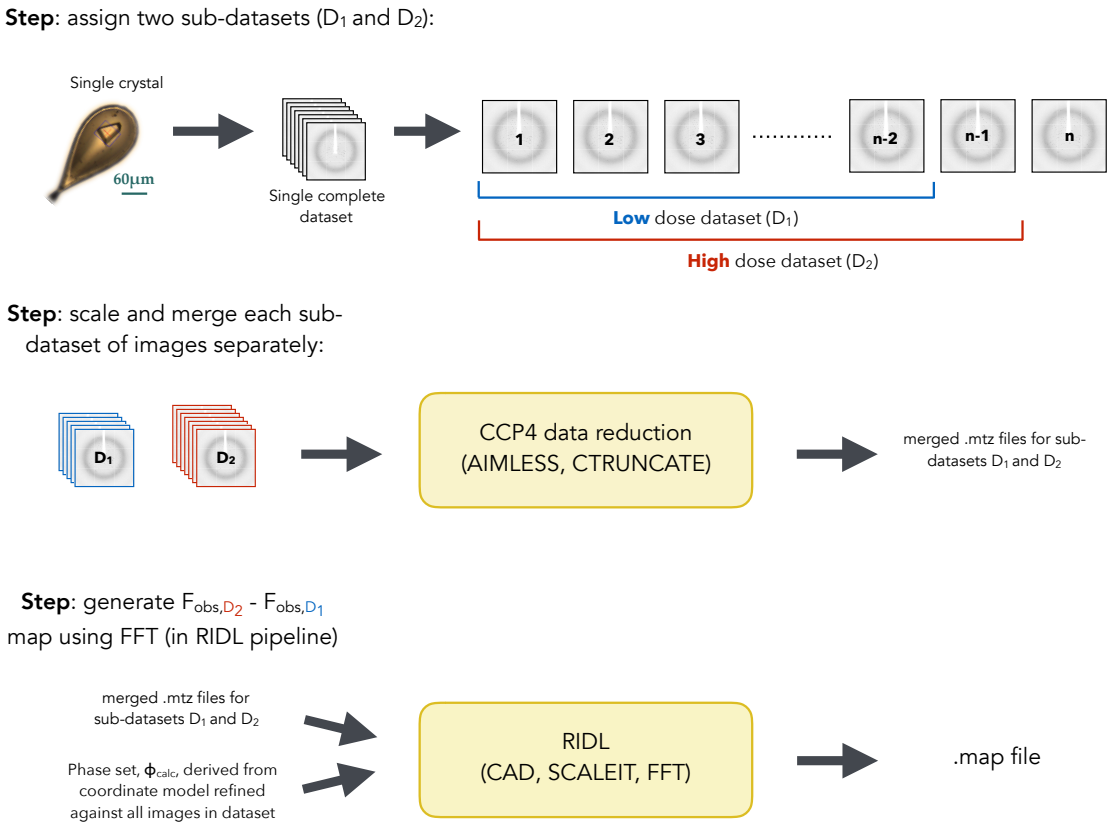


Figure 7.12: A schematic illustrating a protocol to systematically generate Fourier difference maps, $F_{\mathrm{obs,D_2}} - F_{\mathrm{obs,D_1}}$, from a single MX dataset. Necessary CCP4 suite programs for each step have been given in brackets.

scheme includes only a subset of possible (D_1,D_2) image pairings:

$$D_1: \ i = 1, ..., p, \ \text{ such that } 1 \leq p < n \text{ and } \text{mod}(p, c) = 0 \tag{7.3.3}$$

$$D_2: \ i = 1, ..., q, \ \text{ such that } 1 < q \leq n \text{ and } \text{mod}(q, c) = 0 \tag{7.3.4}$$

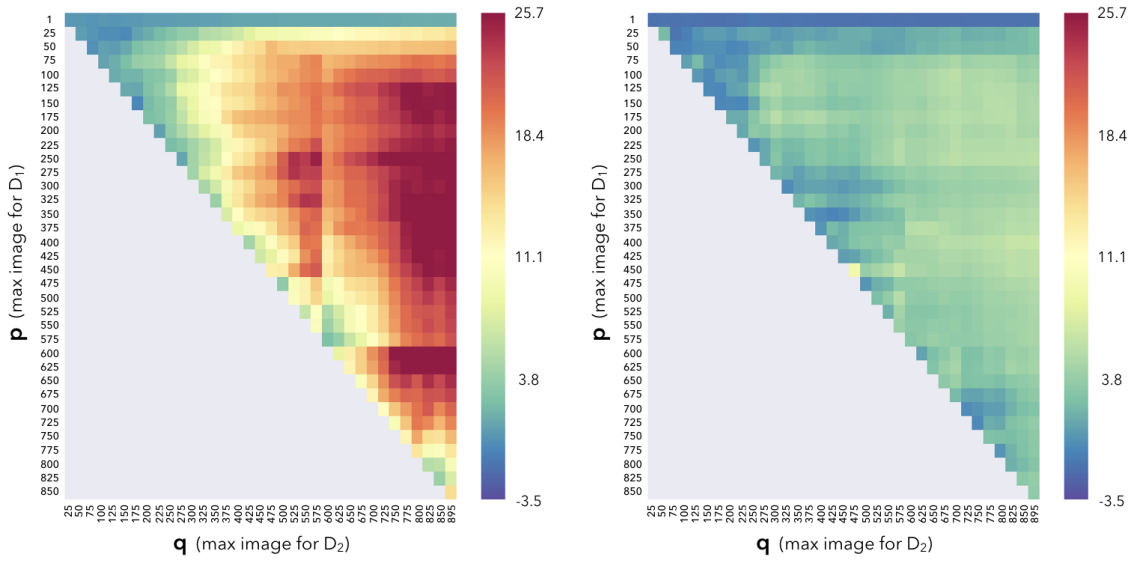
where $\text{mod}(x, y)$ is the modulus function, such that $\text{mod}(x, y) = 0$ if x is divisible by y. For each generated (D_1,D_2) pairing, the custom Python script then proceeds to generate a $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ map using the *RIDL* pipeline.

7.3.3 Damage detection to individual atoms

The per-atom summary metrics computed by *RIDL* (e.g. D_{loss}) provide suitable objective metrics to screen for radiation-induced electron density changes across the large number of generated $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ maps. The data shown in Figure 7.13 illustrate the variation in the C_α-normalised D_{loss} metric between each generated $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ map, for selected atoms in insulin. For each heat map, the vertical/horizontal axis denotes the images included in the D_1/D_2 image set. In the following analysis, the values of p and q will be used to identify specific sub-datasets, D_1 and D_2 (using the notation in the previous section).

Across Cys-S_γ atoms, a differential susceptibility of damage was observed, with B-Cys-7-S_γ exhibiting the highest D_{loss} metrics relative to other atom types (Figure 7.13a). In general, for each Cys-S_γ atom in insulin, the largest D_{loss} values coincide with a region of (D_1,D_2)-space such that $p < 500$ for D_1, and $q > 700$ for D_2 (corresponding DWD values of < 0.49 MGy and > 0.69 MGy respectively). This observation is consistent with there being a relatively large cumulative dose difference between D_1 and D_2 in this region (ΔDWD > 0.2 MGy), such that radiation damage effects became more pronounced. For insulin, the D_{loss} metric diminished as p decreases below image number ~ 125 (< 0.12 MGy), and is suggested to be due to the low completeness of sub-dataset D_1 in this region of (D_1,D_2)-space (below 50 % overall completeness), as illustrated in Figure 7.14a.

Figure 7.15 presents the corresponding analysis for the GH7 family protein, again for example Cys-S_γ atoms. For the case of D_{loss} metric, differential behaviour between sulphur atoms is again



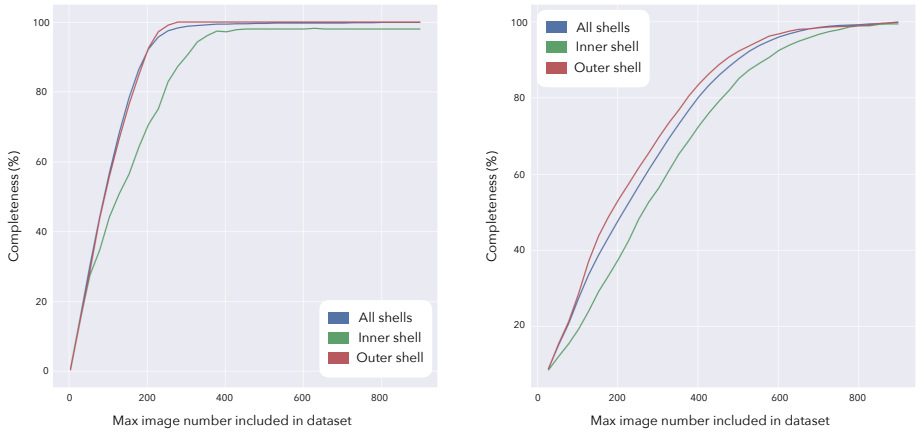
(a) B-Cys-7-S_γ **(b)** B-Cys-19-S_γ

Figure 7.13: Heat map analysis of the radiation-susceptibility of individual atoms in insulin. The C_α-normalised D_{loss} metric has been plotted for varying sub-datasets D_1 and D_2 for (a) B-Cys-7-S_γ and (b) B-Cys-19-S_γ. Metric values have been derived from $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ maps using *RIDL*, for each (D_1,D_2) pairing. The maximum image ($= p$) assigned to each lower dose dataset, D_1, has been displayed on each y-axis, with the maximum image ($= q$) for each higher dose dataset, D_2, displayed on the x-axis. Colour scheme is scaled across all refined atoms in the insulin coordinate model, and is not specific to each plot.

observed. However, for atoms exhibiting high D_{loss} values (e.g. A-Cys-364-S_γ in Figure 7.15a), the colour gradient distribution is qualitatively dissimilar to that of insulin in Figure 7.13. Most notably, the region of (D_1,D_2)-space with $p < 200$ (for sub-dataset D_1 on the y-axis) exhibits high D_{loss} values. Furthermore, this did not appear to be an artefact of the D_{loss} metric, since as shown in Figure 7.15c, heat map analysis instead using the density-weighted D^{-}_{mean} metric[‡‡] indicated that whereas density-weighted D^{-}_{mean} metric values appear diminished (relative to the observations for D_{loss}) for D_1 with $p < 200$, high metric values were still nonetheless reported. For this metric, map voxels that are closer to the atomic centre of the A-Cys-364-S_γ atom have been up-weighted to contribute more to the metric value relative to distal voxels.

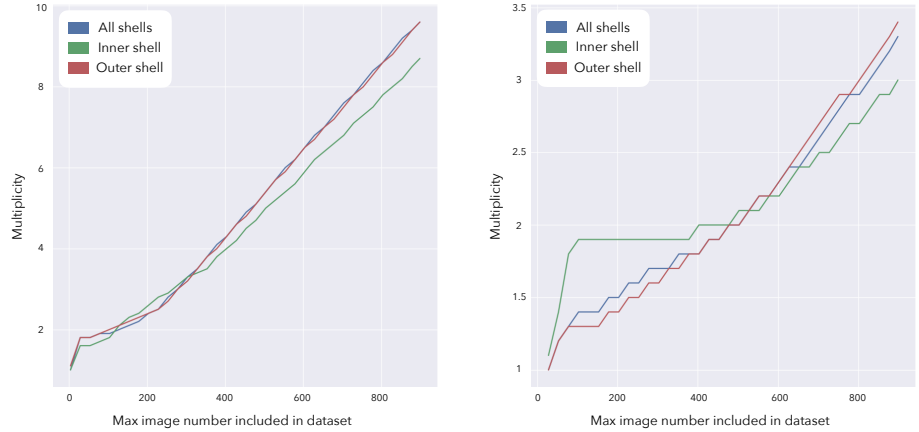
To elucidate whether this region ($p < 200$) represents the detection of true site-specific damage, Figure 7.16a provides a snapshot of the $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ map generated for the particular case of $p = 75$ and $q = 900$ (rendered in *COOT*). Pronounced negative density is present at the A-Cys-364-S_γ atom, but is poorly aligned with the sulphur atom. This $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ map was noted to be noisy, with long parallel *striations* of negative and positive map density dominating across the unit cell below a $\pm 3\sigma$ threshold. It is highly suspected that these map striations

[‡‡] see Section 2.5.7 for full metric construction



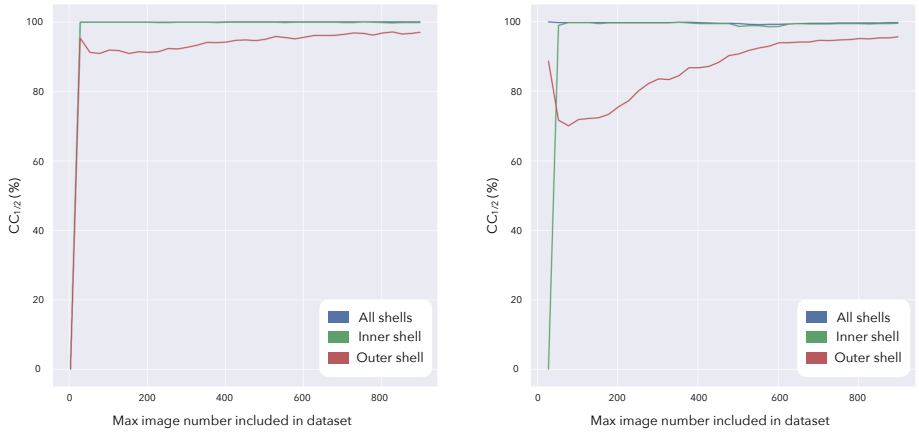
(a) Completeness (%), insulin

(b) Completeness (%), GH7



(c) Multiplicity, insulin

(d) Multiplicity, GH7



(e) $CC_{1/2}$ (%), insulin

(f) $CC_{1/2}$ (%), GH7

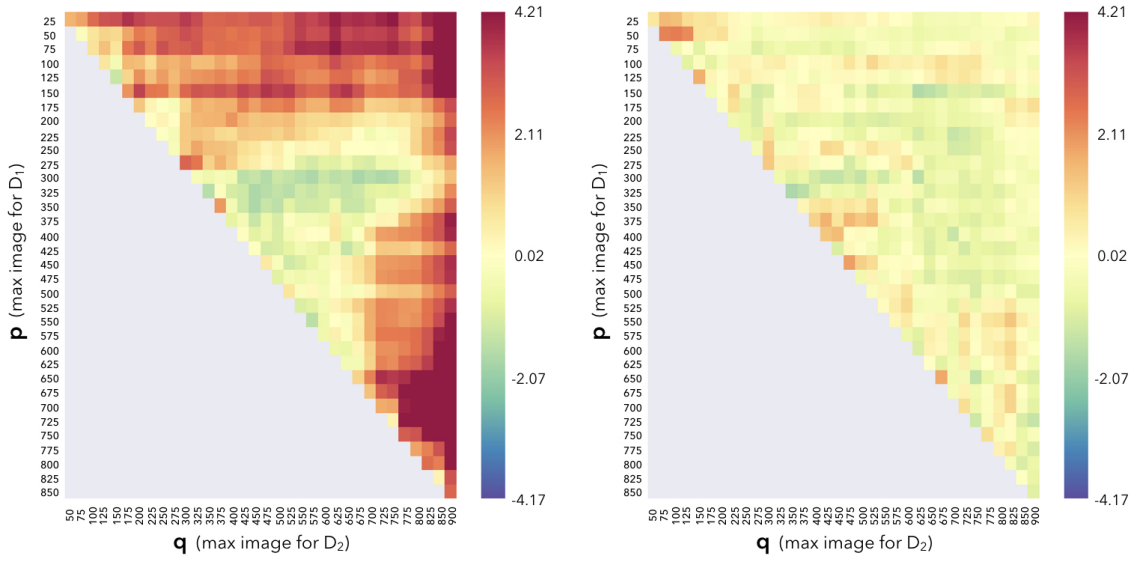
Figure 7.14: The data reduction statistics output by *AIMLESS* for the scaling and merging of each sub-dataset (D_1 or D_2) for insulin and the GH7 family protein. Values of (a,b) the data completeness, (c,d) reflection multiplicity, and (e,f) $CC_{1/2}$ have been reported for inner, outer and over all resolution shells. On each x-axis, the largest image number included in a sub-dataset has been quoted. For example, a value of $x = 100$ corresponds to images 1 to 100 inclusive.

were a manifestation of the low data completeness for sub-dataset D_1 at $p = 75$, as shown in Figure 7.14b (with overall completeness of $\sim 20\,\%$). However, this observation alone was insufficient to conclude anything further. Below, a comparison will next be made with additional $F_{obs,n} - F_{obs,1}$ maps in which true damage to the A-Cys-364-S_γ site is known to have manifested.

It is noted in Figure 7.15a that high D_{loss} values were also attained by the A-Cys-364-S_γ atom for D_1 with $p > 550$. The $F_{obs,D_2} - F_{obs,D_1}$ map generated for the example case of $p = 750$ and $q = 900$ is shown in Figure 7.16b. The map density is poorly defined at the sulphur atom, but is consistent with reorientation of the A-Cys-364-S_γ atom following bond cleavage. In Figure 7.16c, for reference, a $F_{obs,n} - F_{obs,1}$ Fourier difference map is shown that has been independently calculated from the first and second *full* diffraction datasets (corresponding to disjoint sets of images $i = 1, ..., 900$ and $j = 901, ...1800$.) The difference density is qualitatively similar to that shown in Figure 7.17b, including radiation-induced reorientation of the Met side-group to the left of the A-Cys-364 residue. Interestingly, in Figure 7.16a, the positive $F_{obs,n} - F_{obs,1}$ map peak to the right of the disulphide bridge is also conserved, providing additional evidence that the region of (D_1,D_2)-space with $p < 200$ may also correspond to the detection of a true damage site.

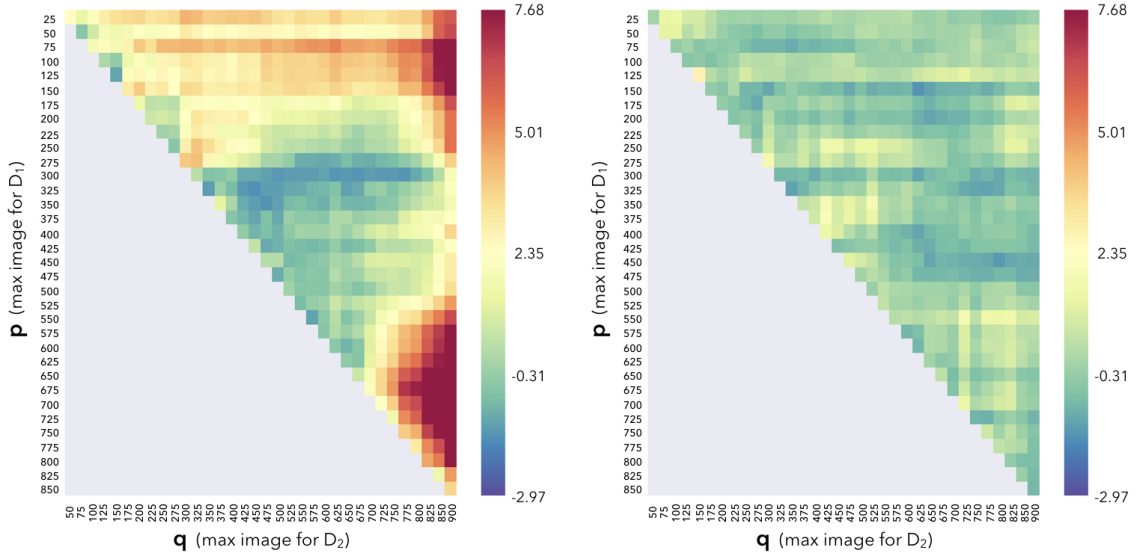
Overall, from these observations it has been concluded that site-specific damage to individual cysteine sulphur atoms can be detected in a single set of MX diffraction images. The evidence in Figure 7.16a may also suggest that such observations can be made in the absence of high completeness diffraction data (for D_1 with $p < 200$).

With reference to Figure 7.15a, the lack of high D_{loss} metric values in the intermediate region ($200 < p < 500$) may be explained due to the data collection strategy for GH7. For the plot shown in Figure 7.17, *RADDOSE-3D* has been used to calculate dose metrics (max dose, average dose, DWD) for increasing rotation ranges between 10° and 180° (all starting at 0°). The data indicate that there is a decrease in the rate at which the DWD increases as the crystal is rotated between $70 - 120°$. The difference in dose between sub-datasets D_1 and D_2 is therefore reduced in this region, relative to other regions of each heat map shown in Figure 7.15. It is highly likely that the Gaussian beam profile is responsible for this non-uniform rate of dose accumulation as the GH7 crystal was rotated.



(a) A-Cys-364-S$_\gamma$, D_{loss}

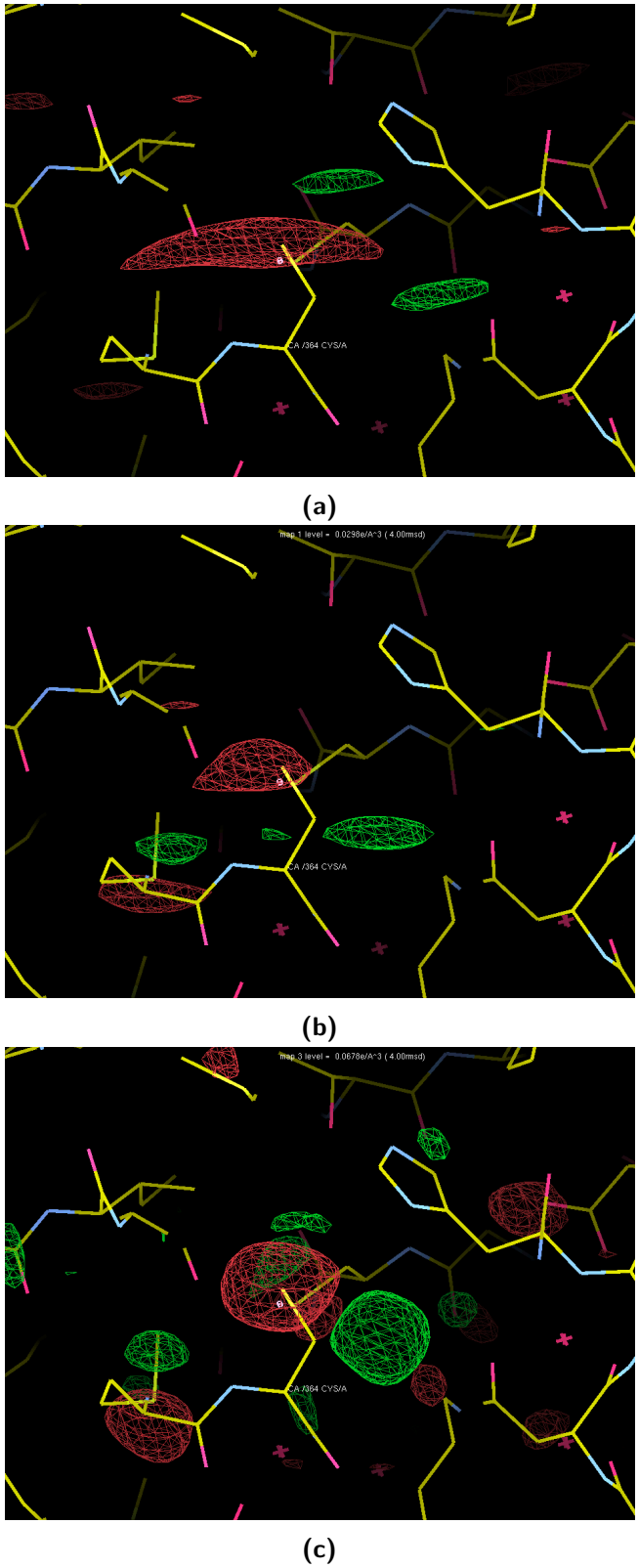
(b) B-Cys-254-S$_\gamma$, D_{loss}



(c) A-Cys-364-S$_\gamma$, density-weighted D^{-}_{mean}

(d) B-Cys-254-S$_\gamma$ density-weighted D^{-}_{mean}

Figure 7.15: Heat map analysis of the radiation-susceptibility of individual atoms in the GH7 family protein. The (a,b) D_{loss} metric and (c,d) density-weighted D^{-}_{mean} metrics (both C$_\alpha$-normalised) have been plotted for varying sub-datasets D$_1$ and D$_2$ for (a) B-Cys-7-S$_\gamma$ and (b) B-Cys-19-S$_\gamma$. Metric values have been derived from $F_{\text{obs,D}_2} - F_{\text{obs,D}_1}$ maps using *RIDL*, for each (D$_1$,D$_2$) pairing. The definition of the density-weighted D^{-}_{mean} metric is provided in Section 2.5.7. The maximum image ($= p$) assigned to each lower dose dataset, D$_1$, has been displayed on each y-axis, with the maximum image ($= q$) for each higher dose dataset, D$_2$, displayed on the x-axis. Colour scheme is scaled across all refined atoms in the GH7 protein coordinate model, and is not specific to each plot.



(a)

(b)

(c)

Figure 7.16: A visualisation of Fourier difference maps generated from a single MX dataset for the GH7 family protein. (a,b) The $F_{\text{obs,D}_2} - F_{\text{obs,D}_1}$ map has been generated in *RIDL* for D_1 and D_2 corresponding to (a) $p = 75$ and $q = 900$, and (b) $p = 750$ and $q = 900$. In (c) a $F_{\text{obs,n}} - F_{\text{obs,1}}$ map has been calculated from the first and second *full* diffraction datasets (corresponding to disjoint sets of images $i = 1, ..., 900$ and $j = 901, ..., 1800$ for comparison with (a) and (b). All maps have been centred at the A-Cys-364-S_γ atom in the GH7 protein, and have been contoured at $\pm 4\sigma$ in green/red. Screenshots have been taken directly from *COOT*.

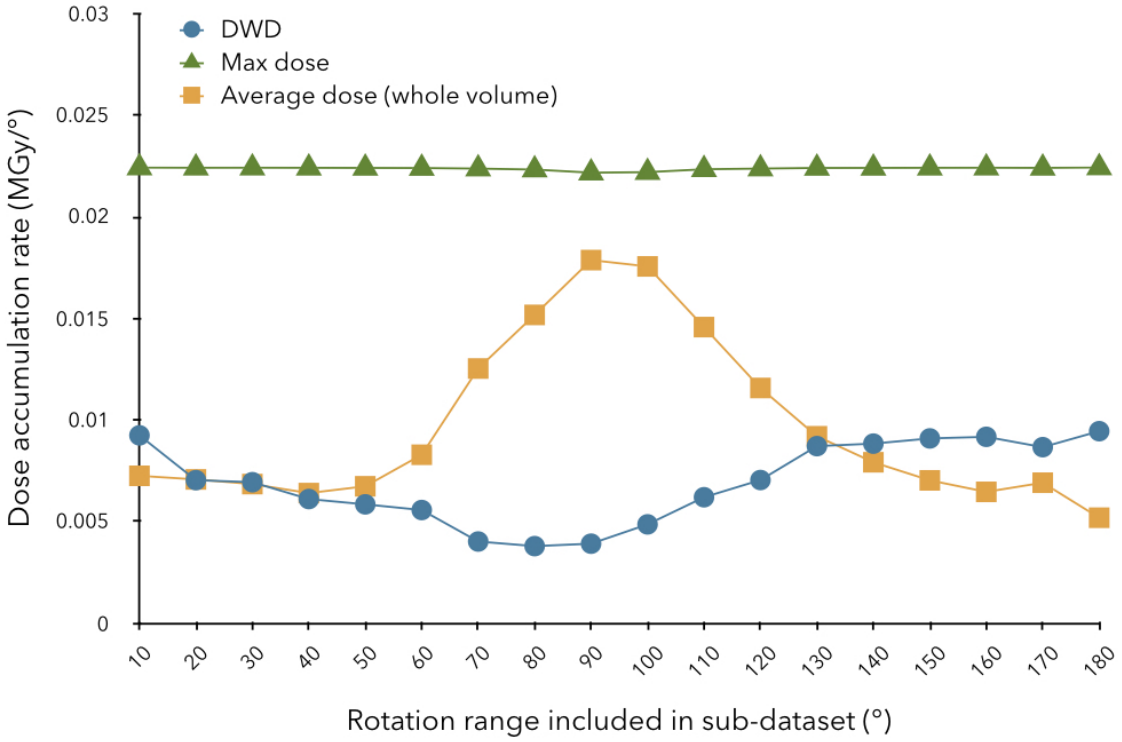


Figure 7.17: The rate of dose accumulation (in units of MGy/$^\circ$) for the case of the GH7 family protein data collection. The maximum dose, average dose and DWD were calculated by *RADDOSE-3D* for increasing angular ranges between 10° and 180° (all starting at 0°). For an angular range $0 - i^\circ$, the angular DWD rate (y-axis) was approximated by $\frac{\text{DWD}(i) - \text{DWD}(i-10)}{10^\circ}$, where DWD$(i)$ and DWD$(i-10)$ are the DWD values corresponding to angular ranges $0 - i^\circ$ and $0 - (i-10)^\circ$. The maximum and average dose rates were defined similarly.

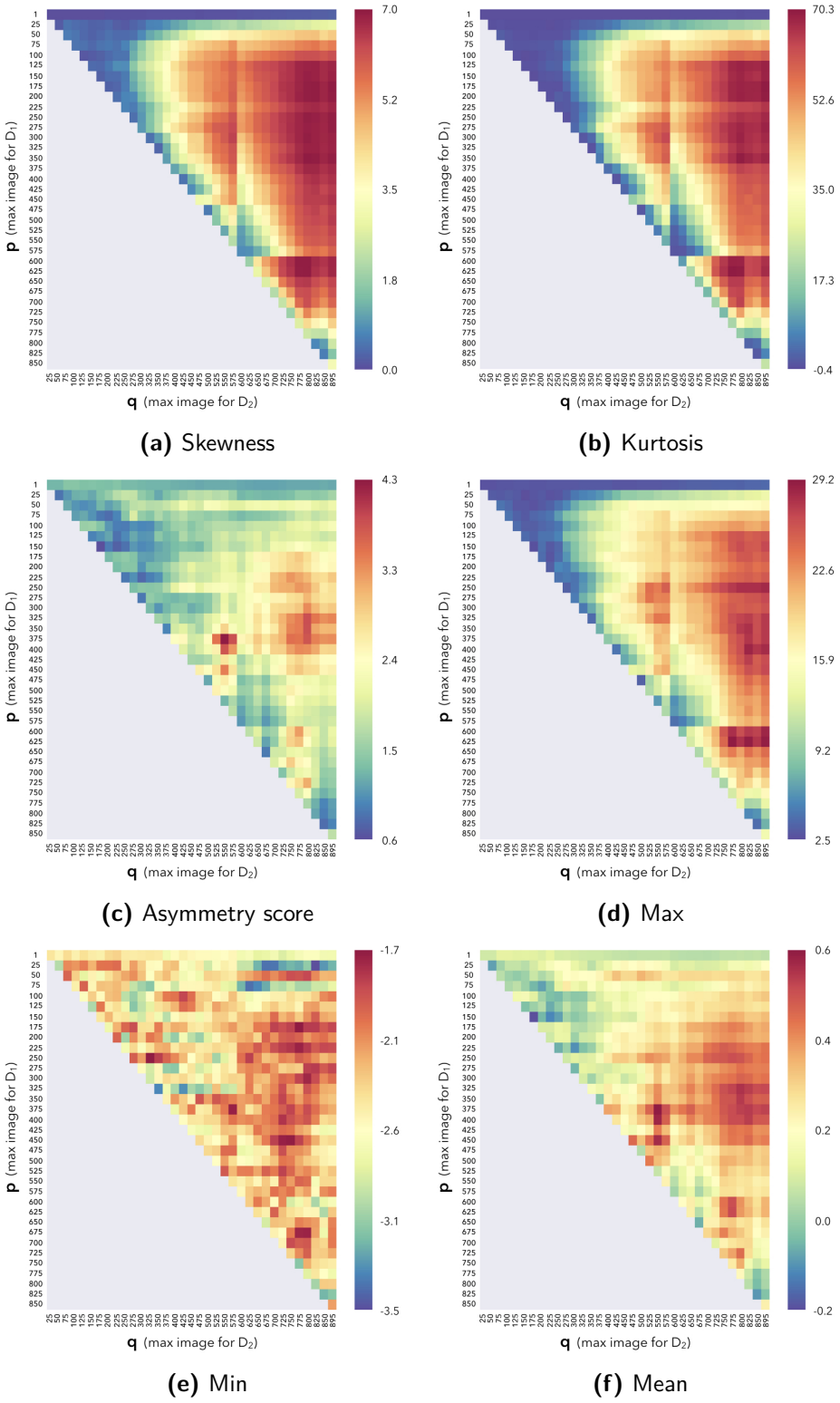
7.3.4 Assessment of damage to overall structure

To extend the observations above for individual atoms, this section will demonstrate that it was also possible to assess the damage to each overall structure, using metrics output by *RIDL*. As has previously been presented in Section 3.2.6.3, the asymmetry of the distribution of the C_α-normalised D_{loss} metric ($D_{\text{loss}}^{C_\alpha}$) values over all atoms in a structure has been proposed as a method to quantify the overall damage to a structure. Following C_α-normalisation, the D_{loss} metric attains the value of zero for radiation-insensitive atoms, with a positive skewed tail indicating the prevalence of damaged amino acids. Here, three statistics have been calculated by *RIDL* to assess the asymmetry in $D_{\text{loss}}^{C_\alpha}$ metric distribution about $D_{\text{loss}}^{C_\alpha} = 0$: skewness, kurtosis, and '*asymmetry score*'*. In addition, the maximum, mean and minimum $D_{\text{loss}}^{C_\alpha}$ values were calculated over all refined atoms in each structure, as additional indicators of overall damage.

Figure 7.18 indicates variation of each metric over (D_1,D_2)-space. For the case of insulin, the majority of the metrics appear to indicate the increasing prevalence of site-specific damage events when the dose separation between D_1 and D_2 is large. However, the sensitivities of these metrics differ: for example, the *asymmetry score* metric appears less sensitive than either the kurtosis or skewness. The minimum of $D_{\text{loss}}^{C_\alpha}$ values exhibit no such trend, which is expected since the minimum $D_{\text{loss}}^{C_\alpha}$ values correspond to atoms with the least electron density loss in each $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ map.

For the GH7 protein, the data shown in Figure 7.19 illustrate qualitatively different behaviour than for that for insulin. In particular, the skewness and kurtosis are consistently low (attaining maximum values of 1.1 and 3.65, relative to 7 and 70.3 for insulin) across the sampled (D_1,D_2)-space. Overall, the prevalence of site-specific damage across all atoms in the GH7 protein was low, compared to insulin. Indeed, besides Cys-S_γ density loss (as presented in Figure 7.15), no clear damage events were detectable for any Glu, Asp or Met groups using heat map analysis. It could not be ruled out that for GH7, the detection of site-specific damage in Figure 7.19 has been limited by the data completeness and intensity multiplicity for each sub-dataset (as shown in Figure 7.14b and d).

*As in Section 3.2.6.3, the asymmetry score has been calculated as the ratio $\Sigma_\uparrow/\Sigma_\downarrow$ between the sum of $D_{\text{loss}}^{C_\alpha}$ values attained above ($\Sigma_\uparrow$) and below ($\Sigma_\downarrow$) $D_{\text{loss}}^{C_\alpha} = 0$

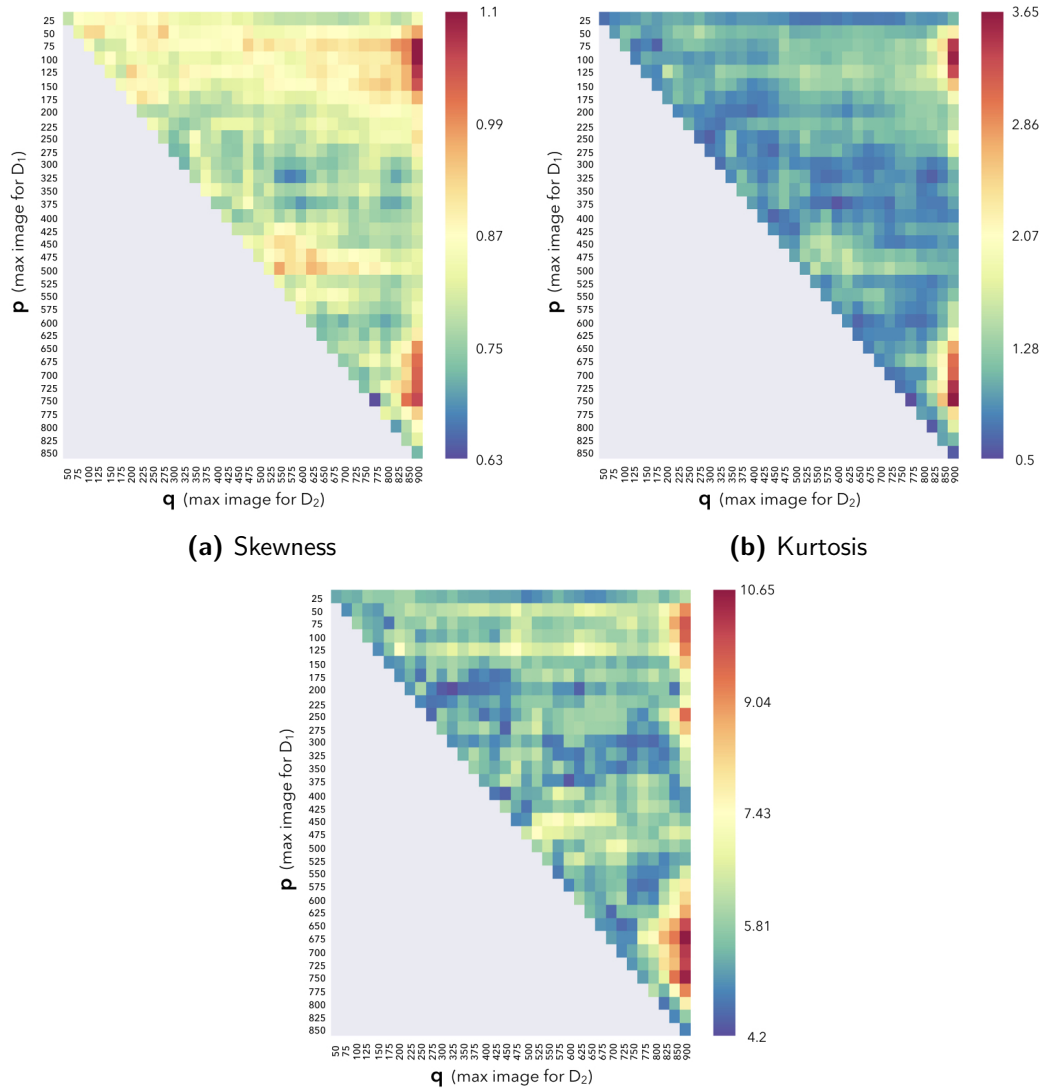


(a) Skewness **(b)** Kurtosis

(c) Asymmetry score **(d)** Max

(e) Min **(f)** Mean

Figure 7.18: Heat map analysis of the radiation-susceptibility over all atoms in insulin. For varying sub-datasets D_1 and D_2, the statistical properties of the distribution of C_α-normalised D_{loss} metric values over all refined atoms in the insulin coordinate model have been plotted. Metric values have been derived from $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ maps using *RIDL*, for each (D_1,D_2) pairing. The maximum image ($= p$) assigned to each lower dose dataset, D_1, has been displayed on each y-axis, with the maximum image ($= q$) for each higher dose dataset, D_2, displayed on the x-axis. The colour scheme is independent for each plot and should not be compared between statistics.



(a) Skewness **(b)** Kurtosis

(c) Max

Figure 7.19: Heat map analysis of the radiation-susceptibility over all atoms in the GH7 family protein. For varying sub-datasets D_1 and D_2, the statistical properties of the distribution of C_α-normalised D_{loss} metric values over all refined atoms in the GH7 protein coordinate model have been plotted. Metric values have been derived from $F_{\text{obs},D_2} - F_{\text{obs},D_1}$ maps using *RIDL*, for each (D_1,D_2) pairing. The maximum image ($= p$) assigned to each lower dose dataset, D_1, has been displayed on each y-axis, with the maximum image ($= q$) for each higher dose dataset, D_2, displayed on the x-axis. The colour scheme is independent for each plot and should not be compared between statistics.

7.3.5 Extension to other protein datasets

The analysis in this section has focused on two case studies (bovine insulin and the GH7 protein), in order demonstrate the proposed protocol. Comparable analysis has also been performed on other MX damage series for proteins that have been presented in detail elsewhere in this thesis (TRAP and the C.Esp1396I-DNA complex). However, these studies yielded null results, which have not been discussed in detail here. Neither protein contains disulphide bonds, and it is suggested that the current protocol may be too insensitive to detect site-specific damage to other groups besides cysteine sulphur atoms. Indeed, a further investigation has since been conducted on a single MX dataset obtained from a lysozyme crystal with data collected to 1.4 Å (supplied by D. Barrett, visiting Master's student to the Garman group), consisting of 90° of images[†] with an angular step of 0.1°. Similar to insulin and GH7 protein, heat map analysis revealing electron density loss consistent with disulphide bond cleavage was detectable for several disulphide bonds in lysozyme during the 90° data collection.

It is clear that important parameters to determine whether the current sliding window strategy is successful or fails are the data multiplicity and completeness for partial subsets of images that are then combined to generate Fourier difference maps. As shown in Figure 7.14, the multiplicity is consistently lower for the GH7 protein compared to the insulin case study, and the data completeness for the GH7 protein only approaches 100 % when almost all images (>600/900) are included in a sub dataset. However, despite this variability, it has been demonstrated in this chapter that the current strategy has been successful in the detection of disulphide bridge damage for both test systems. A systematic study on the relationship of data quality and success over a wider range of protein systems would be required to elucidate this dependency.

7.4 Discussion

This chapter has presented two extensions of electron density analysis using *RIDL*. Both of these applications have aimed to improve the current understanding of site-specific damage processes in cryocooled MX data collection, and to provide MX experimenters with additional information and

[†]The lysozyme crystallised in space group $P4_32_12$, and it was observed that 90° of diffraction data were required for high data completeness (near 100 %) calculated over all measured resolution shells.

tools in order to ensure the collection of the best possible quality diffraction data in the inevitable presence of radiation damage

The first section has addressed the calculation of half-doses, to identify the dose required to reduce the $2mF_{obs} - DF_{calc}$ density surrounding individual atoms by $50\,\%$. In order to estimate half-dose values susceptible groups in the chosen protein case studies (TRAP and the GH7 family protein) a simplistic three state model was proposed in which the ordered population of a given atom across all unit cells, X_0, will diminish either by disordering to state X_1, or dissociation (by covalent bond cleavage) to a third state, X_2. By assuming this model, the $2mF_{obs} - DF_{calc}$ density decay could be approximated as an exponential decay function, and an explicit half-dose value calculated by fitting the decay function to data at each dose. Here, the magnitude of $2mF_{obs} - DF_{calc}$ density per-atom was quantified using variants of the per-atom damage metrics introduced in Section 2.5.6, as generated by the *RIDL* pipeline.

Two main observations were made through the analysis for TRAP and the GH7 protein. Firstly, rapid $2mF_{obs} - DF_{calc}$ map density decay was widely reported for protein side-groups at low DWD values (< 5 MGy), and not restricted to Cys, Met, Glu and Asp groups. A conservative DWD limit of 10 MGy has been proposed from the current findings, however significant variation in half-dose was reported between atoms, as anticipated. Secondly, for individual atoms, the local $2mF_{obs} - DF_{calc}$ map density decayed to a variable end state in the limit as dose d became large. It therefore could not be assumed that for radiation-susceptible groups (e.g. Glu carboxylates), that all originally ordered atoms prior to X-irradiation had eventually undergone bond cleavage at high doses. Instead, a disordered (yet bound) fraction is assumed to persist and to continue to contribute to the electron density at that site.

The current analysis has been limited by the availability of MX protein damage series collected over a sufficient number of datasets to perform reliable exponential fitting without the risk of overfitting to too few data points. In addition, accurate estimations of the cumulative dose for each dataset must be available in order to calculate meaningful half-dose values. Unfortunately, these criteria lead to damage series previously deposited in the PDB (as utilised in Section 6.3) being unsuitable. Further work on multiple independent protein damage series would be required to validate the proposed specific damage DWD half-dose of 10 MGy above.

The second section of this chapter has aimed to develop a high-throughput protocol to screen for site-specific damage in a single complete MX damage series. For standard MX data collection for structure solution, typically an insufficient quantity of diffraction data is collected to generate $F_{\text{obs,n}} - F_{\text{obs,1}}$ Fourier difference maps between successive disjoint datasets of images. In this section, it has been demonstrated that a limited quantity of diffraction damage can still be exploited to generate Fourier difference maps, from which site-specific damage signal can be detected. A protocol has been presented in which a series of images $i = 1, ..., n$ is divided into subsets of images D_1 and D_2, and a $F_{\text{obs,D}_2} - F_{\text{obs,D}_1}$ Fourier difference map is then generated between each sub-datasets. For example case studies of bovine insulin and the GH7 family protein, it has been shown that by varying the number of images corresponding to D_1 and D_2, disulphide bond cleavage can be detected in electron density derived from a single diffraction dataset. For the GH7 protein, which crystallises in space group C121, the full 180° of data were required for a dataset of high completeness. Despite this limitation, electron density loss to Cys-S_γ atoms was still detectable.

These results reinforce the observations made earlier for the half-dose calculations, that specific damage can cause detectable alterations to electron density rapidly, using data collection at doses below 1 MGy. Overall, this section has provided a proof of principle of a protocol, but further investigation on other independent MX datasets would be required to establish the reliability (and detection rate) of this protocol. In particular, whereas the heat map analysis conducted in Section 7.3.3 and 7.3.4 provides a high-throughput approach to screen for damage events, it is likely to be sensitive to noisy artefacts in each $F_{\text{obs,D}_2} - F_{\text{obs,D}_1}$ maps. However, it does provide an efficient means to identify particular $F_{\text{obs,D}_2} - F_{\text{obs,D}_1}$ maps that may require manual inspection. The further development of such a damage screening tool could be highly advantageous for validation of structural models prior to publication or associated data deposition in openly-available online databases (e.g. the wwPDB, *https://www.wwpdb.org*).

CHAPTER 8

Conclusion

In macromolecular X-ray crystallography (MX), radiation damage is still a major limiting factor to successful data collection from crystals held at both room and cryo-cooled temperatures (100 K). Even for cases where a sufficient quantity of diffraction data can be collected from a crystal before the onset of significant global radiation damage, radiation-induced chemical and conformation changes to particular amino acid groups have been widely reported and at doses of just several MGy. When undetected, such artefacts of X-ray irradiation lead to detrimental errors in model building; unfortunately sites of functional importance in proteins (ligand binding sites, inter-molecular interfaces) often appear particularly prone to these effects. It is thus vital to (a) correctly characterise possible damage events in a wide range of macromolecular structures that experimenters may come across, and to (b) develop reliable tools to help experimenters to detect and correct for such damage events in their own models.

Visual inspection of $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ Fourier difference maps is a widely utilised tool for detecting damage events in protein models, giving the change in electron density between at least two successive dose states. The direct dependence of the map on diffraction intensities, as opposed to the refined coordinate model, reduces the effects of model bias and poor refinement relative to other damage detection strategies. However, to date, visual interpretation of such maps has been hindered by the inherent subjective bias of the experimenter, and furthermore becomes increasingly intractable to perform for large structures.

Two tools have been developed to optimise the use of $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ difference maps. The program *PEAKS* was first written to facilitate higher-throughput manual inspection of multiple $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ maps comprising an MX damage series. It operates by filtering the large number of $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map peaks based on a set of objective criteria, in order to minimise the number of noisy background map peaks that must be inspected by the experimenter. *PEAKS* then guides the experimenter through manual inspection of the remaining peaks. Nonetheless, *PEAKS* is still limited by the subjectivity of the human inspector.

A further computational method, *RIDL*, has been developed to enable objective quantification of local changes in $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map density for each individual refined atom in an input coordinate model, in the form of per-atom damage metrics. Any subjective opinion of the investigator is thus eliminated entirely. The suitability of different per-atom metrics has been tested, and the $D_{\mathrm{loss}}(\mathrm{atom})$ metric has ultimately been presented as a simple, yet powerful, indicator of radiation

damage. D_{loss} (in units of e/Å^3) provides a worst case scenario for the electron density lost by an atom within a specified search radius, thus allowing the investigator to objectively screen for damage hotspots in proteins. Metrics requiring integration of electron density map values in the vicinity of an atom were deemed highly inadvisable, due to the heavy dependence on the allocated search radius around an atom. To overcome this, the use of a ρ_{calc} density-weighting scheme to weight the contribution of each $F_{\text{obs,n}} - F_{\text{obs,1}}$ map voxel by its distance from an atom during metric calculation has been presented. The resulting density-weighted D^{-}_{mean} metric has been proposed as a further useful metric, giving information in addition to the upper bound that D_{loss} provides. Density-weighted metrics are, however, far more dependent on the quality of the coordinate model from which the ρ_{calc} map is derived. In addition, metric C_α-normalisation schemes have been discussed as a method to permit each per-atom metric to be directly compared between independent protein systems, thus improving the general utility of *RIDL*. Indeed, in Chapter 6, C_α-normalisation proved to be a vital component enabling independent MX damage series downloaded from the *PDB_REDO* online server to be directly compared using *RIDL*.

The *RIDL* program is now openly available for use by the crystallographic community (available at *https://github.com/GarmanGroup/RIDL*), and has been accessed by several international research groups. Besides studies on radiation damage, it is also believed that *RIDL* will prove to be a valuable tool enabling the quantitative tracking of electron density changes to localised regions of proteins in time-resolved crystallographic experiments, using data possibly merged from a large number of isomorphous crystals. Similarly to site-specific damage analysis to date, such time-resolved changes are regularly only qualitatively probed by visual inspection of electron density maps, and are often hindered by the weak density changes in maps being misinterpreted as noise (and vice versa). However, the development of *RIDL* in the current thesis has focused exclusively on its application to radiation damage research; as has been discussed below, it has enabled the systematic investigation of damage to be performed on a wide range of macromolecular systems.

Whereas there exist numerous studies detailing site-specific effects to proteins at 100 K, the wider class of macromolecules including nucleic acids and their complexes with proteins have remained uncharacterised. As the MX structural biology community now strives to solve structures for increasingly larger and complex macromolecular assemblies, it essential to understand how such structures are affected by the X-ray radiation used to solve them. To this end, the work in this

thesis has performed systematic analyses on multiple DNA and RNA -containing crystal systems.

Investigation of pre-formed nucleoprotein complexes provides a tractable basis to compare relative damage profiles between the nucleic acid and protein under the same controlled experimental conditions. Chapter 3 has reported investigations on two well-characterised bacterial C-protein dsDNA complexes (C.Esp1396I and C.Csp231I). To the author's knowledge these were the first MX damage investigations to be carried out on nucleoprotein complexes. Whereas classical protein site-specific damage was detectable in $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps at doses of several MGy, the bound DNA components remained relatively unaffected by X-irradiation up to high doses (> 20 MGy). At high doses, electron density loss localised at the O3' and O5' atoms of the phosphate backbone was present for both the C.Esp1396I and C.Csp231I system. However, due to the limited resolution (2.8 Å and 2.5 Å, respectively), the electron density loss could not be assigned specifically to either the 3' or 5' C−O phosphate-sugar bonds that would indicate a backbone cleavage event. Analysis of the C.Csp231I system was also limited by the high twinning fraction of 46 % reported for this data.

To determine whether the strong helical structure of the 35 and 21 bp DNA in C.Esp1396I and C.Csp231I was responsible for this observed radiation-insensitivity of DNA, Chapter 4 presented a further case study on the large (204 kDa) *Trp* RNA-binding attenuation protein (TRAP) bound to a 53 bp repeating $(\text{GAGUU})_{10}$GAG RNA sequence. Similarly to DNA, the flexible RNA component again was determined to be highly resistant to damage effects up to high doses (> 20 MGy). The TRAP system was chosen since it crystallises in a 1:1 ratio of RNA-bound and non-bound protein in the asymmetric unit, permitting a controlled investigation on the effect of RNA-binding on protein damage susceptibility. Per-atom quantification of electron density loss (D_{loss}) using *RIDL* revealed short scale (< 7 Å) stabilisation of protein residues by RNA. This effect was most significant for residues (Glu-36, Lys-37 and Phe-32) that were essential for high-affinity RNA-binding in each of the eleven symmetrically-dispersed binding interfaces around TRAP. The data were consistent with a model in which RNA bases physically stabilised amino acid side-groups either through direct hydrogen bonding or non-polar stacking interactions. The preservation of these interactions ensured the RNA-complex remained intact during X-ray irradiation of the crystal.

For the previous case studies, it could not be distinguished whether nucleic acid is intrinsically

robust to site-specific damage effects or whether bound protein may offer a protective effect, to divert mobile radical species (migrating low energy electrons) to the protein component. Indeed, such scavenging capabilities have been observed by small molecule radioprotectants soaked into protein crystals (De la Mora *et al.*, 2011). To elucidate this, Chapter 5 presented further investigations on three distinct DNA crystal forms: TCGGCGCCGA, CGCAAATTTGCG and CCCT. These crystals were chosen to represent a range of DNA conditions, including variable sequence length, binding configuration, and heavy-metal content. Consistent with the results detailed in Chapter 3 and 4, electron density loss localised to the DNA backbone was observable across the tested systems, with a detectable onset between $11.2 - 29.5\,\mathrm{MGy}$ across the three systems. There was insufficient evidence to suggest that the absence of bound protein had accelerated the loss of electron density. Despite the high resolution of these studies (between 1.4 Å and 2.0 Å), negative density in $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ maps could not convincingly be attributed to cleavage of any individual bond of the phosphate backbone. It is thus suggested here that such electron density loss therefore may correspond to merely atomic disordering, as opposed to a well-defined cleavage event. However it cannot be ruled out that the P−O3' and P−O5' bonds are broken with the same frequency at a phosphate site across the crystal and the resulting $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map density is an averaging over both events. To elucidate this, it is suggested that future work should now focus on isolating the cleavage of a particular bond at the DNA backbone of crystalline DNA, with the aid of complementary spectroscopic techniques to probe any radiation-induced changes at either the P−O3' and P−O5' bonds.

In addition, although not a focus of the current research, the investigated DNA crystals have also been concluded to be highly resistant to global radiation damage effects. In particular, the global half-dose values, $D_{1/2}$ for the mean intensity of diffraction intensities to reduce by $50\,\%$, appeared to slightly surpass the upper bound reported for protein crystals (Owen *et al.*, 2006). Due to the limited statistics on three DNA crystals, the current work cannot predict an upper half-dose threshold applicable to DNA, and it would therefore be highly beneficial to the DNA crystallographic community to identify whether such a threshold exists over a much wider range of distinct DNA crystals.

For protein crystals irradiated at 100 K, a differential susceptibility of residues has been widely reported in the previous radiation damage literature, with disulphide bond cleavage, decarboxyla-

tion of Glu and Asp, and Met $C_\gamma - S_\delta$ bond cleavage documented as the most prevalent events at doses of several MGy. However, a previous suggestion of hydroxyl group cleavage for tyrosine in myrosinase (Burmeister, 2000) has rapidly disseminated throughout the literature, despite being generally unobserved in other protein systems. Moreover, in contrast to Burmeister (2000), the phenolic C−O bond is known to be highly robust throughout a wide range of literature from the radiation chemistry community. In Chapter 6, a critical reanalysis was reported of the original myrosinase data deposited by Burmeister (2000), followed by a systematic analysis using *RIDL* of all suitable MX damage series available to the current author through the *PDB_REDO* online server. It was concluded that the original observations of $F_{\mathrm{obs,n}} - F_{\mathrm{obs,1}}$ map electron density localised to the Tyr −OH group are not a widespread feature of protein damage series. Furthermore, the combined data support a model in which the full aromatic ring of Tyr is disordered by irradiation, with no need to invoke an energetically unfavourable bond cleavage event. The results described in Chapter 6 conclude that Tyr should not be considered amongst the most radiation-sensitive groups in proteins.

Whereas advisory half-dose limits (Owen *et al.*, 2006; Henderson, 1990) have been defined, at which point any protein crystals at 100 K will have experienced significant global radiation damage effects, no such counterpart exists for protein site-specific damage events. This issue is compounded further since the literature often misquotes the above limits as *safe-dose* limits to evade specific damage. The work presented in Chapter 7 has aimed to estimate half-dose limits for protein specific damage events for test protein case studies (TRAP and a GH7 family protein, chosen due to their large molecular weights). The program *RIDL* has been modified to quantify and track the magnitude of $2mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$ map density as a function of dose for individual atoms. Using a simplistic first order reaction model, the decay in electron density at each atom has been approximated as an exponential function, from which numerical half-dose values could be extracted. For Cys-S_γ and Glu and Asp $-CO_2$ groups, half-doses varied considerably, however a conservative mean half-dose of 10 MGy was computed for the two test proteins. For each individual atom, local $2mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$ map density was observed to decay and plateau with dose to a fixed end state as dose increased, but this end state differed between atoms. This provided evidence that an equilibrium state is reached for each atom at high doses, with only a sub-unity fraction of each group undergoing full dissociation across all unit cells. Further work would be essential to extend the current half-dose calculation protocol to a wider range of protein systems,

in order to establish a statistically-significant half dose for each cleavage event.

For the work presented in Chapters 3-5, the *RIDL* pipeline has proved a vital tool for performing objective analysis of MX radiation damage series. However, the majority of diffraction datasets collected by the MX community for structural characterisation do not typically contain enough diffraction images to be divided into multiple disjoint and complete datasets in order to generate $F_{\text{obs,n}} - F_{\text{obs,1}}$ maps. Chapter 7 described an investigation as to whether it is still possible to divide a limited quantity of images into two sub-datasets, D_1 and D_2, such that a $F_{\text{obs,D}_2} - F_{\text{obs,D}_1}$ map still yields true damage signal at atomic sites. For the cases of an insulin and GH7 protein dataset, it was demonstrated that *RIDL* can still detect site-specific damage to disulphide bonds using only a single set of images that would all typically be merged for use in refinement. Unfortunately, the current protocol did not appear sensitive enough to detect decarboxylation events, and was additionally susceptible to map noise due to the reduced completeness and multiplicity associated with the dividing images into subsets. Further work would be required to extend this protocol to additional protein test cases, with the ultimate aim in developing a rigorous high-throughput screen for site-specific damage events in diffraction data prior to its deposition in openly available online data reserves.

Taken together, the research for this thesis has provided a tried and tested tool that is applicable to both future radiation damage research, and also the wider MX community.

Bibliography

Abrahams, S. C. International Union of Crystallography Commission on Crystallographic Apparatus Single-Crystal Radiation-Damage Survey. *Acta Crystallographica Section A*, **A 29**(1):111–6, 1973. doi:10.1107/S056773947300032x.

Adam, V., Royant, A., Nivière, V., Molina-Heredia, F. P., and Bourgeois, D. Structure of superoxide reductase bound to ferrocyanide and active site expansion upon X-ray-induced photo-reduction. *Structure*, **12**(9):1729–40, 2004. doi:10.1016/j.str.2004.07.013.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 2):213–21, 2010. doi:10.1107/S0907444909052925.

Alizadeh, E., Orlando, T. M., and Sanche, L. Biomolecular damage induced by ionizing radiation: the direct and indirect effects of low-energy electrons on DNA. *Annu Rev Phys Chem*, **66**:379–98, 2015. doi:10.1146/annurev-physchem-040513-103605.

Alizadeh, E. and Sanche, L. Low-energy-electron interactions with DNA: approaching cellular conditions with atmospheric experiments. *European Physical Journal D*, **68**(4):1–13, 2014. doi:10.1140/epjd/e2014-40775-5.

Allan, E. G., Kander, M. C., Carmichael, I., and Garman, E. F. To scavenge or not to scavenge, that is STILL the question. *J Synchrotron Radiat*, **20**(Pt 1):23–36, 2013. doi:10.1107/S0909049512046237.

Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B*, **58**(Pt 3 Pt 1):380–8, 2002. doi:10.1107/S0108768102003890.

Alphey, M. S., Gabrielsen, M., Micossi, E., Leonard, G. A., McSweeney, S. M., Ravelli, R. B., Tetaud, E., Fairlamb, A. H., Bond, C. S., and Hunter, W. N. Tryparedoxins from Crithidia fasciculata and Trypanosoma brucei: photoreduction of the redox disulfide using synchrotron radiation and evidence for a conformational switch implicated in function. *J Biol Chem*, **278**(28):25919–25, 2003. doi:10.1074/jbc.M301526200.

Antson, A. A., Dodson, E. J., Dodson, G., Greaves, R. B., Chen, X., and Gollnick, P. Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature*, **401**(6750):235–42, 1999. doi:10.1038/45730.

Arzt, S., Beteva, A., Cipriani, F., Delageniere, S., Felisaz, F., Förstner, G., Gordon, E., Launer, L., Lavault, B., Leonard, G., Mairs, T., McCarthy, A., McCarthy, J., McSweeney, S., Meyer, J., Mitchell, E., Monaco, S., Nurizzo, D., Ravelli, R., Rey, V., Shepard, W., Spruce, D., Svensson, O., and Theveneau, P. Automation of macromolecular crystallography beamlines. *Prog Biophys Mol Biol*, **89**(2):124–52, 2005. doi:10.1016/j.pbiomolbio.2004.09.003.

Assi, H. A., Harkness, R. W. t., Martin-Pintado, N., Wilds, C. J., Campos-Olivas, R., Mittermaier, A. K., Gonzalez, C., and Damha, M. J. Stabilization of i-motif structures by 2'-beta-fluorination of DNA. *Nucleic Acids Res*, **44**(11):4998–5009, 2016. doi:10.1093/nar/gkw402.

Axford, D., Foadi, J., Hu, N. J., Choudhury, H. G., Iwata, S., Beis, K., Evans, G., and Alguel, Y. Structure determination of an integral membrane protein at room temperature from crystals in situ. *Acta Crystallogr D Biol Crystallogr*, **71**(Pt 6):1228–37, 2015a. doi:10.1107/S139900471500423X.

Axford, D., Foadi, J., Hu, N. J., Choudhury, H. G., Iwata, S., Beis, K., Evans, G., and Alguel, Y. Structure determination of an integral membrane protein at room temperature from crystals in situ. *Acta Crystallogr D Biol Crystallogr*, **71**(Pt 6):1228–37, 2015b. doi:10.1107/S139900471500423X.

Ball, N., Streeter, S. D., Kneale, G. G., and McGeehan, J. E. Structure of the restriction-modification controller protein C.Esp1396I. *Acta Crystallogr D Biol Crystallogr*, **65**(Pt 9):900–5, 2009. doi:10.1107/S0907444909020514.

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. *J Synchrotron Radiat*, **16**(Pt 2):205–16, 2009. doi:10.1107/S0909049509003343.

Barkley, M. D. and Bourgeois, S. Repressor recognition of operator and effectors, page 177âĂŞ220. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York., 1978. doi:10.1101/087969133.7.177.

Barrios, R., Skurski, P., and Simons, J. Mechanism for damage to DNA by low-energy electrons. *Journal of Physical Chemistry B*, **106**(33):7991–4, 2002. doi:10.1021/jp013861i.

Baxter, R. H., Seagle, B. L., Ponomarenko, N., and Norris, J. R. Specific radiation damage illustrates light-induced structural changes in the photosynthetic reaction center. *J Am Chem Soc*, **126**(51):16728–9, 2004. doi:10.1021/ja0448115.

Begusova, M., Eon, S., Sy, D., Culard, F., Charlier, M., and Spotheim-Maurizot, M. Radiosensitivity of DNA in a specific protein-DNA complex: the lac repressor-lac operator complex. *Int J Radiat Biol*, **77**(6):645–54, 2001. doi:10.1080/09553000110041876.

Beitlich, T., Kühnel, K., Schulze-Briese, C., Shoeman, R. L., and Schlichting, I. Cryoradiolytic reduction of crystalline heme proteins: analysis by UV-Vis spectroscopy and X-ray crystallography. *J Synchrotron Radiat*, **14**(Pt 1):11–23, 2007. doi:10.1107/S0909049506049806.

Benson, S. W. Thermochemistry and Kinetics of Sulfur-Containing Molecules and Radicals. *Chemical Reviews*, **78**(1):23–35, 1978. doi:10.1021/cr60311a003.

Berdys, J., Skurski, P., and Simons, J. Damage to Model DNA Fragments by 0.25âĹŠ1.0 eV Electrons Attached to a Thymine pi* Orbital. *The Journal of Physical Chemistry B*, **108**(18):5800–5, 2004. doi:10.1021/jp049728i.

Berglund, G. I., Carlsson, G. H., Smith, A. T., Szöke, H., Henriksen, A., and Hajdu, J. The catalytic pathway of horseradish peroxidase at high resolution. *Nature*, **417**(6887):463–8, 2002. doi:10.1038/417463a.

Blake, C. and Phillips, D. C. pages 183–91. Vienna: International Atomic Energy Agency, 1962.

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M., and Bricogne, G. Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr D Biol Crystallogr*, **60**(Pt 12 Pt 1):2210–21, 2004. doi:10.1107/S0907444904016427.

Blow, D. M. and Crick, F. H. C. The Treatment of Errors in the Isomorphous Replacement Method. *Acta Crystallographica*, **12**(10):794–802, 1959. doi:10.1107/S0365110x59002274.

Borek, D., Ginell, S. L., Cymborowski, M., Minor, W., and Otwinowski, Z. The many faces of radiation-induced changes. *J Synchrotron Radiat*, **14**(Pt 1):24–33, 2007. doi:10.1107/S0909049506046589.

Borshchevskiy, V., Round, E., Erofeev, I., Weik, M., Ishchenko, A., Gushchin, I., Mishin, A., Willbold, D., Büldt, G., and Gordeliy, V. Low-dose X-ray radiation induces structural alterations in proteins. *Acta Crystallogr D Biol Crystallogr*, **70**(Pt 10):2675–85, 2014. doi:10.1107/S1399004714017295.

Boudaïffa, B., Cloutier, P., Hunting, D., Huels, M. A., and Sanche, L. Resonant formation of DNA strand breaks by low-energy (3 to 20 eV) electrons. *Science*, **287**(5458):1658–60, 2000. doi:10.1126/science.287.5458.1658.

Bragg, W. H. and Bragg, W. L. The Reflection of X-rays by Crystals. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **88**(605):428–38, 1913. ISSN 0950-1207. doi:10.1098/rspa.1913.0040.

Brockhauser, S., Ravelli, R. B., and McCarthy, A. A. The use of a mini-kappa goniometer head in macromolecular crystallography diffraction experiments. *Acta Crystallogr D Biol Crystallogr*, **69**(Pt 7):1241–51, 2013. doi:10.1107/S0907444913003880.

Burmeister, W. P. Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Crystallogr D Biol Crystallogr*, **56**(Pt 3):328–41, 2000. doi:10.1107/S0907444999016261.

Burmeister, W. P., Cottaz, S., Driguez, H., Iori, R., Palmieri, S., and Henrissat, B. The crystal structures of Sinapis alba myrosinase and a covalent glycosyl-enzyme intermediate provide insights into the substrate recognition and active-site machinery of an S-glycosidase. *Structure*, **5**(5):663–675, 1997. doi:10.1016/S0969-2126(97)00221-9.

Burmeister, W. P., Cottaz, S., Rollin, P., Vasella, A., and Henrissat, B. High resolution X-ray crystallography shows that ascorbate is a cofactor for myrosinase and substitutes for the function of the catalytic base. *J Biol Chem*, **275**(50):39385–93, 2000. doi:10.1074/jbc.M006796200.

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. *J Synchrotron Radiat*, **22**(2):213–24, 2015. doi:10.1107/S1600577514026289.

Bury, C. S., Carmichael, I., and Garman, E. F. OH cleavage from tyrosine: debunking a myth. *J Synchrotron Radiat*, **24**(Pt 1):7–18, 2017. doi:10.1107/S1600577516016775.

Bury, C. S., Carmichael, I., McGeehan, J. E., and Garman, E. F. Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. *Radiation Physics and Chemistry*, **128**:118–25, 2016a. doi:10.1016/j.radphyschem.2016.05.023.

Bury, C. S., McGeehan, J. E., Antson, A. A., Carmichael, I., Gerstel, M., Shevtsov, M. B., and Garman, E. F. RNA protects a nucleoprotein complex against radiation damage. *Acta Crystallogr D Struct Biol*, **72**(Pt 5):648–57, 2016b. doi:10.1107/S2059798316003351.

Cadet, J., Delatour, T., Douki, T., Gasparutto, D., Pouget, J. P., Ravanat, J. L., and Sauvaigo, S. Hydroxyl radicals and DNA base damage. *Mutat Res*, **424**(1-2):9–21, 1999.

Campbell, J. L., Boyd, N. I., Grassi, N., Bonnick, P., and Maxwell, J. A. The Guelph PIXE software package IV. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, **268**(20):3356–63, 2010. doi:10.1016/j.nimb.2010.07.012.

Carugo, O. and Djinovic Carugo, K. When X-rays modify the protein structure: radiation damage at work. *Trends Biochem Sci*, **30**(4):213–9, 2005. doi:10.1016/j.tibs.2005.02.009.

Castelvecchi, D. Next-generation X-ray source fires up. *Nature*, **525**(7567):15–6, 2015. doi:10.1038/nature.2015.18253.

Chance, M. R., Sclavi, B., Woodson, S. A., and Brenowitz, M. Examining the conformational dynamics of macromolecules with time-resolved synchrotron X-ray 'footprinting'. *Structure*, **5**(7):865–9, 1997. doi:10.1016/S0969-2126(97)00241-4.

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., 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., Homke, A., Reich, C., Pietschner, D., Struder, L., Hauser, G., Gorke, H., Ullrich, J., Herrmann, S., Schaller, G., Schopper, F., Soltau, H., Kuhnel, 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., Jonsson, O., Svenda, M., Stern, S., Nass, K., Andritschke, R., Schroter, C. D., Krasniqi, F., Bott, M., Schmidt, K. E., Wang, X., Grotjohann, I., Holton,

J. M., Barends, T. R., 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. Femtosecond X-ray protein nanocrystallography. *Nature*, **470**(7332):73–7, 2011. doi:10.1038/nature09750.

Charlier, M., Eon, S., Seche, E., Bouffard, S., Culard, F., and Spotheim-Maurizot, M. Radiolysis of lac repressor by gamma-rays and heavy ions: a two-hit model for protein inactivation. *Biophys J*, **82**(5):2373–82, 2002. doi:10.1016/S0006-3495(02)75582-3.

Chayen, N. E. and Saridakis, E. Protein crystallization: from purified protein to diffraction-quality crystal. *Nat Methods*, **5**(2):147–53, 2008. doi:10.1038/nmeth.f.203.

Chen, V. B., Arendall, r., 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 Crystallogr D Biol Crystallogr*, **66**(Pt 1):12–21, 2010. doi:10.1107/S0907444909042073.

Chen, X., Antson, A. A., Yang, M., Li, P., Baumann, C., Dodson, E. J., Dodson, G. G., and Gollnick, P. Regulatory features of the trp operon and the crystal structure of the trp RNA-binding attenuation protein from Bacillus stearothermophilus. *J Mol Biol*, **289**(4):1003–16, 1999. doi:10.1006/jmbi.1999.2834.

Chrysochoos, J. Pulse radiolysis of phenylalanine and tyrosine. *Radiat Res*, **33**(3):465–79, 1968.

Clavel, D., Gotthard, G., von Stetten, D., De Sanctis, D., Pasquier, H., Lambert, G. G., Shaner, N. C., and Royant, A. Structural analysis of the bright monomeric yellow-green fluorescent protein mNeonGreen obtained by directed evolution. *Acta Crystallogr D Struct Biol*, **72**(Pt 12):1298–307, 2016. doi:10.1107/S2059798316018623.

Corbett, M. C., Latimer, M. J., Poulos, T. L., Sevrioukova, I. F., Hodgson, K. O., and Hedman, B. Photoreduction of the active site of the metalloprotein putidaredoxin by synchrotron radiation. *Acta Crystallogr D Biol Crystallogr*, **63**(Pt 9):951–60, 2007. doi:10.1107/S0907444907035160.

Cosier, J. and Glazer, A. M. A Nitrogen-Gas-Stream Cryostat for General X-Ray-Diffraction Studies. *Journal of Applied Crystallography*, **19**(2):105–7, 1986. doi:10.1107/S0021889886089835.

Cragg, S. M., Beckham, G. T., Bruce, N. C., Bugg, T. D., Distel, D. L., Dupree, P., Etxabe, A. G., Goodell, B. S., Jellison, J., McGeehan, J. E., McQueen-Mason, S. J., Schnorr, K., Walton, P. H., Watts, J. E., and Zimmer, M. Lignocellulose degradation mechanisms across the Tree of Life. *Curr Opin Chem Biol*, **29**:108–19, 2015. doi:10.1016/j.cbpa.2015.10.018.

Crick, F. H. C. and Magdoff, B. S. The theory of the method of isomorphous replacement for protein crystals. I. *Acta Crystallographica*, **9**(11):901–8, 1956. doi:10.1107/s0365110x56002552.

Czapla-Masztafiak, J., Szlachetko, J., Milne, C. J., Lipiec, E., Sá, J., Penfold, T. J., Huthwelker, T., Borca, C., Abela, R., and Kwiatek, W. M. Investigating DNA Radiation Damage Using X-Ray Absorption Spectroscopy. *Biophys J*, **110**(6):1304–11, 2016. doi:10.1016/j.bpj.2016.01.031.

Day, H. A., Pavlou, P., and Waller, Z. A. i-Motif DNA: structure, stability and targeting with ligands. *Bioorg Med Chem*, **22**(16):4407–18, 2014. doi:10.1016/j.bmc.2014.05.047.

De la Mora, E., Carmichael, I., and Garman, E. F. Effective scavenging at cryotemperatures: further increasing the dose tolerance of protein crystals. *J Synchrotron Radiat*, **18**(Pt 3):346–57, 2011. doi:10.1107/S0909049511007163.

de Sanctis, D., Oscarsson, M., Popov, A., Svensson, O., and Leonard, G. Facilitating best practices in collecting anomalous scattering data for de novo structure solution at the ESRF Structural Biology Beamlines. *Acta Crystallogr D Struct Biol*, **72**(Pt 3):413–20, 2016. doi:10.1107/S2059798316001042.

Dickerson, R. E. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res*, **26**(8):1906–26, 1998. doi:10.1093/nar/26.8.1906.

Diederichs, K. Some aspects of quantitative analysis and correction of radiation damage. *Acta Crystallogr D Biol Crystallogr*, **62**(Pt 1):96–101, 2006. doi:10.1107/S0907444905031537.

Doutch, J., Hough, M. A., Hasnain, S. S., and Strange, R. W. Challenges of sulfur SAD phasing as a routine method in macromolecular crystallography. *J Synchrotron Radiat*, **19**(Pt 1):19–29, 2012. doi:10.1107/S0909049511049004.

Du, H. and Babitzke, P. trp RNA-binding attenuation protein-mediated long distance RNA refolding regulates translation of trpE in Bacillus subtilis. *J Biol Chem*, **273**(32):20494–503, 1998.

Dubnovitsky, A. P., Ravelli, R. B., Popov, A. N., and Papageorgiou, A. C. Strain relief at the active site of phosphoserine aminotransferase induced by radiation damage. *Protein Sci*, **14**(6):1498–507, 2005. doi:10.1110/ps.051397905.

Edwards, K. J., Brown, D. G., Spink, N., Skelly, J. V., and Neidlet, S. Molecular structure of the B-DNA dodecamer d(CGCAAATTTGCG)2. An examination of propeller twist and minor-groove water structure at 2.2 A resolution. *J. Mol. Biol.*, **226**(4):1161–73, 1992.

Elliott, M. B., Gottlieb, P. A., and Gollnick, P. Probing the TRAP-RNA interaction with nucleoside analogs. *RNA*, **5**(10):1277–89, 1999. doi:10.1017/S1355838299991057.

Elliott, M. B., Gottlieb, P. A., and Gollnick, P. The mechanism of RNA binding to TRAP: initiation and cooperative interactions. *RNA*, **7**(1):85–93, 2001. doi:10.1017/S135583820100173X.

Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 4):486–501, 2010. doi:10.1107/S0907444910007493.

Engh, R. A. and Huber, R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A Foundations of Crystallography*, **47**(4):392–400, 1991. doi:10.1107/s0108767391001071.

Eon, S., Culard, F., Sy, D., Charlier, M., and Spotheim-Maurizot, M. Radiation disrupts protein-DNA complexes through damage to the protein. The lac repressor-operator system. *Radiation Research*, **156**(1):110–7, 2001. doi:10.1667/0033-7587(2001)156[0110:Rdpdct]2.0.Co;2.

Evans, P. Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr*, **62**(Pt 1):72–82, 2006. doi:10.1107/S0907444905036693.

Evans, P. and McCoy, A. An introduction to molecular replacement. *Acta Crystallogr D Biol Crystallogr*, **64**(Pt 1):1–10, 2008. doi:10.1107/S0907444907051554.

Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr*, **67**(Pt 4):282–92, 2011. doi:10.1107/S090744491003982X.

Evans, P. R. and Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr*, **69**(Pt 7):1204–14, 2013. doi:10.1107/S0907444913000061.

Fasanella, E. L. and Gordy, W. Electron Spin Resonance of an Irradiated Single Crystal of L-Tyrosine-Hcl. *Proceedings of the National Academy of Sciences*, **62**(2):299–304, 1969.

Ferradini, C. and Jay-Gerin, J. P. The effect of pH on water radiolysis: A still open question - A minireview. *Research on Chemical Intermediates*, **26**(6):549–65, 2000. doi:10.1163/156856700x00525.

Fioravanti, E., Vellieux, F. M., Amara, P., Madern, D., and Weik, M. Specific radiation damage to acidic residues and its relation to their chemical and structural environment. *J Synchrotron Radiat*, **14**(Pt 1):84–91, 2007. doi:10.1107/S0909049506038623.

Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N., and Alber, T. Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc Natl Acad Sci U S A*, **108**(39):16247–52, 2011. doi:10.1073/pnas.1111325108.

French, S. and Wilson, K. On the treatment of negative intensity observations. *Acta Crystallographica Section A*, **34**(4):517–25, 1978. doi:10.1107/s0567739478001114.

Galli, L., Son, S. K., Klinge, M., Bajt, S., Barty, A., Bean, R., Betzel, C., Beyerlein, K. R., Caleman, C., Doak, R. B., Duszenko, M., Fleckenstein, H., Gati, C., Hunt, B., Kirian, R. A., Liang, M., Nanao, M. H., Nass, K., Oberthur, D., Redecke, L., Shoeman, R., Stellato, F., Yoon, C. H., White, T. A., Yefanov, O., Spence, J., and Chapman, H. N. Electronic damage in S atoms in a native protein crystal induced by an intense X-ray free-electron laser pulse. *Struct Dyn*, **2**(4):041703, 2015. doi:10.1063/1.4919398.

Garman, E. Leaving no element of doubt: analysis of proteins using microPIXE. *Structure*, **7**(12):R291–R299, 1999. doi:10.1016/S0969-2126(00)88335-5.

Garman, E. F. Crystal handling for cryogenic data collection, pages 181–96. International University Line, La Jolla, U.S.A., 2009. ISBN 0-9636817-5-3.

Garman, E. F. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 4):339–51, 2010. doi:10.1107/S0907444910008656.

Garman, E. F. and Grime, G. W. Elemental analysis of proteins by microPIXE. *Prog Biophys Mol Biol*, **89**(2):173–205, 2005. doi:10.1016/j.pbiomolbio.2004.09.005.

Garman, E. F. and Owen, R. L. Cryocooling and radiation damage in macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, **62**(Pt 1):32–47, 2006. doi:10.1107/S0907444905034207.

Garman, E. F. and Schneider, T. R. Macromolecular cryocrystallography. *Journal of Applied Crystallography*, **30**(3):211–37, 1997. doi:10.1107/S0021889897002677.

Garman, E. F. and Weik, M. Radiation Damage in Macromolecular Crystallography. *Methods Mol Biol*, **1607**:467–89, 2017a. doi:10.1007/978-1-4939-7000-1{_}20.

Garman, E. F. and Weik, M. X-ray radiation damage to biological macromolecules: further insights. *J Synchrotron Radiat*, **24**(Pt 1):1–6, 2017b. doi:10.1107/S160057751602018X.

Garrison, W. M. Reaction-Mechanisms in the Radiolysis of Peptides, Polypeptides, and Proteins. *Chemical Reviews*, **87**(2):381–98, 1987. doi:10.1021/cr00078a006.

Gerstel, M., Deane, C. M., and Garman, E. F. Identifying and quantifying radiation damage at the atomic level. *J Synchrotron Radiat*, **22**(2):201–12, 2015. doi:10.1107/S1600577515002131.

Gillard, N., Begusova, M., Castaing, B., and Spotheim-Maurizot, M. Radiation affects binding of Fpg repair protein to an abasic site containing DNA. *Radiat Res*, **162**(5):566–71, 2004. doi:10.1667/RR3247.

Gillard, N., Goffinont, S., Buré, C., Davídková, M., Maurizot, J. C., Cadene, M., and Spotheim-Maurizot, M. Radiation-induced oxidative damage to the DNA-binding domain of the lactose repressor. *Biochem J*, **403**(3):463–72, 2007. doi:10.1042/BJ20061466.

Gonzalez, A., Thompson, A. W., and Nave, C. CryoâĂŘprotection of protein crystals in intense xâĂŘray beams. *Review of Scientific Instruments*, **63**(1):1177–80, 1992. doi:10.1063/1.1143127.

Grime, G. W. The "Q factor" method: Quantitative microPIXE analysis using RBS normalisation. *Nuclear Instruments & Methods in Physics Research Section B-Beam Interactions with Materials and Atoms*, **109**:170–4, 1996. doi:10.1016/0168-583x(95)00901-9.

Gu, B., Smyth, M., and Kohanoff, J. Protection of DNA against low-energy electrons by amino acids: a first-principles molecular dynamics study. *Phys Chem Chem Phys*, **16**(44):24350–8, 2014. doi:10.1039/c4cp03906h.

Haas, D. J. and Rossmann, M. G. Crystallographic Studies on Lactate Dehydrogenase at-75 Degrees C. *Acta Crystallographica Section B-Structural Crystallography and Crystal Chemistry*, **B 26**:998–1004, 1970. doi:10.1107/S0567740870003485.

Hall, J. P., O'Sullivan, K., Naseer, A., Smith, J. A., Kelly, J. M., and Cardin, C. J. Structure determination of an intercalating ruthenium dipyridophenazine complex which kinks DNA by semiintercalation of a tetraazaphenanthrene ligand. *Proc Natl Acad Sci U S A*, **108**(43):17610–4, 2011. doi:10.1073/pnas.1108685108.

Hall, J. P., Poynton, F. E., Keane, P. M., Gurung, S. P., Brazier, J. A., Cardin, D. J., Winter, G., Gunnlaugsson, T., Sazanovich, I. V., Towrie, M., Cardin, C. J., Kelly, J. M., and Quinn, S. J. Monitoring one-electron photo-oxidation of guanine in DNA crystals using ultrafast infrared spectroscopy. *Nat Chem*, **7**(12):961–7, 2015. doi:10.1038/nchem.2369.

Hart, E. J. The Hydrated Electron: Properties and reactions of this most reactive and elementary of aqueous negative ions are discussed. *Science*, **146**(3640):19–25, 1964. doi:10.1126/science.146.3640.19.

Helliwell, J. R. Protein Crystal Perfection and the Nature of Radiation-Damage. *Journal of Crystal Growth*, **90**(1-3):259–72, 1988. doi:10.1016/0022-0248(88)90322-3.

Henderson, R. Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction. *Proceedings: Biological Sciences*, **241**(1300):6–8, 1990.

Hendrickson, W. A., Love, W. E., and Karle, J. Crystal-Structure Analysis of Sea Lamprey Hemoglobin at 2 angstrom Resolution. *J Mol Biol*, **74**(3):331–61, 1973.

Hiroki, A., Pimblott, S. M., and LaVerne, J. A. Hydrogen Peroxide Production in the Radiolysis of Water with High Radical Scavenger Concentrations. *The Journal of Physical Chemistry A*, **106**(40):9352–8, 2002. doi:10.1021/jp0207578.

Holton, J. M. XANES measurements of the rate of radiation damage to selenomethionine side chains. *J Synchrotron Radiat*, **14**(Pt 1):51–72, 2007. doi:10.1107/S0909049506048898.

Holton, J. M. A beginner's guide to radiation damage. *J Synchrotron Radiat*, **16**(Pt 2):133–42, 2009. doi:10.1107/S0909049509004361.

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 J*, **281**(18):4046–60, 2014. doi:10.1111/febs.12922.

Homer, C., Cooper, L., and Gonzalez, A. Energy dependence of site-specific radiation damage in protein crystals. *J Synchrotron Radiat*, **18**(Pt 3):338–45, 2011. doi:10.1107/S0909049511005504.

Hopcroft, N. H., Manfredo, A., Wendt, A. L., Brzozowski, A. M., Gollnick, P., and Antson, A. A. The interaction of RNA with TRAP: the role of triplet repeats and separating spacer nucleotides. *J Mol Biol*, **338**(1):43–53, 2004. doi:10.1016/j.jmb.2004.02.038.

Hopcroft, N. H., Wendt, A. L., Gollnick, P., and Antson, A. A. Specificity of TRAP-RNA interactions: crystal structures of two complexes with different RNA sequences. *Acta Crystallogr D Biol Crystallogr*, **58**(Pt 4):615–21, 2002. doi:10.1107/S0907444902003189.

Hope, H. Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallogr B*, **44 (Pt 1)**(1):22–6, 1988. doi:10.1107/S0108768187008632.

Horrell, S., Antonyuk, S. V., Eady, R. R., Hasnain, S. S., Hough, M. A., and Strange, R. W. Serial crystallography captures enzyme catalysis in copper nitrite reductase at atomic resolution from one crystal. *IUCrJ*, **3**(Pt 4):271–81, 2016. doi:10.1107/S205225251600823X.

Houée-Levin, C., Bobrowski, K., Horakova, L., Karademir, B., Schoneich, C., Davies, M. J., and Spickett, C. M. Exploring oxidative modifications of tyrosine: an update on mechanisms of formation, advances in analysis and biological consequences. *Free Radic Res*, **49**(4):347–73, 2015. doi:10.3109/10715762.2015.1007968.

Hough, M. A., Antonyuk, S. V., Strange, R. W., Eady, R. R., and Hasnain, S. S. Crystallography with online optical and X-ray absorption spectroscopies demonstrates an ordered mechanism in copper nitrite reductase. *J Mol Biol*, **378**(2):353–61, 2008. doi:10.1016/j.jmb.2008.01.097.

Housecroft, C. E. and Sharpe, A. G. Inorganic Chemistry. v. 1. Pearson Prentice Hall, 2008. ISBN 9780131755536.

Howell, P. L. and Smith, G. D. Identification of Heavy-Atom Derivatives by Normal Probability Methods. *Journal of Applied Crystallography*, **25**(pt 1):81–6, 1992. doi:10.1107/S0021889891010385.

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., and Starodub, D. An assessment of the resolution limitation due to radiation-damage in x-ray diffraction microscopy. *J Electron Spectros Relat Phenomena*, **170**(1-3):4–12, 2009. doi:10.1016/j.elspec.2008.10.008.

Huang, C. Y., Olieric, V., Ma, P., Panepucci, E., Diederichs, K., Wang, M., and Caffrey, M. In meso in situ serial X-ray crystallography of soluble and membrane proteins. *Acta Crystallogr D Biol Crystallogr*, **71**(Pt 6):1238–56, 2015. doi:10.1107/S1399004715005210.

Jones, G. D., Lea, J. S., Symons, M. C., and Taiwo, F. A. Structure and mobility of electron gain and loss centres in proteins. *Nature*, **330**(6150):772–3, 1987. doi:10.1038/330772a0.

Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard, M. Improved Methods for Building Protein Models in Electron-Density Maps and the Location of Errors in These Models. *Acta Crystallographica Section A*, **47**(2):110–9, 1991. doi:10.1107/S0108767390010224.

Joosten, R. P., Long, F., Murshudov, G. N., and Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ*, **1**(Pt 4):213–20, 2014. doi:10.1107/S2052252514009324.

Juers, D. H. and Weik, M. Similarities and differences in radiation damage at 100 K versus 160 K in a crystal of thermolysin. *J Synchrotron Radiat*, **18**(Pt 3):329–37, 2011. doi:10.1107/S0909049511007631.

Kabsch, W. Xds. *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 2):125–32, 2010. doi:10.1107/S0907444909047337.

Kaila, V. R., Verkhovsky, M. I., and Wikstrom, M. Proton-coupled electron transfer in cytochrome oxidase. *Chem Rev*, **110**(12):7062–81, 2010. doi:10.1021/cr1002003.

Kang, C., Berger, I., Lockshin, C., Ratliff, R., Moyzisl, R., and Rich, A. Crystal structure of intercalated four-stranded d(C3T) at 1.4 A resolution. *Proc Natl Acad Sci U S A*, **91**(24):11636–40, 1994.

Keedy, D. A., Kenner, L. R., Warkentin, M., Woldeyes, R. A., Hopkins, J. B., Thompson, M. C., Brewster, A. S., Van Benschoten, A. H., Baxter, E. L., Uervirojnangkoorn, M., McPhillips, S. E., Song, J., Alonso-Mori, R., Holton, J. M., Weis, W. I., Brunger, A. T., Soltis, S. M., Lemke, H., Gonzalez, A., Sauter, N. K., Cohen, A. E., van den Bedem, H., Thorne, R. E., and Fraser, J. S. Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife*, **4**, 2015. doi:10.7554/eLife.07574.

Kern, M., McGeehan, J. E., Streeter, S. D., Martin, R. N., Besser, K., Elias, L., Eborall, W., Malyon, G. P., Payne, C. M., Himmel, M. E., Schnorr, K., Beckham, G. T., Cragg, S. M., Bruce, N. C., and McQueen-Mason, S. J. Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc Natl Acad Sci U S A*, **110**(25):10189–94, 2013. doi:10.1073/pnas.1301502110.

Kleywegt, G. J. and Jones, T. A. xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Crystallogr D Biol Crystallogr*, **52**(Pt 4):826–8, 1996. doi:10.1107/S0907444995014983.

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 Crystallogr D Biol Crystallogr*, **62**(Pt 9):1030–8, 2006. doi:10.1107/S0907444906023869.

Kmetko, J., Warkentin, M., Englich, U., and Thorne, R. E. Can radiation damage to protein crystals be reduced using small-molecule compounds? *Acta Crystallogr D Biol Crystallogr*, **67**(Pt 10):881–93, 2011. doi:10.1107/S0907444911032835.

Kopoldová, J., Liebster, J., and Gross, E. Radiation chemical reactions in aqueous solutions of methionine and its peptides. *Radiat Res*, **30**(2):261–74, 1967.

Koukourakis, M. I. Radiation damage and radioprotectants: new concepts in the era of molecular medicine. *Br J Radiol*, **85**(1012):313–30, 2012. doi:10.1259/bjr/16386034.

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. *J Synchrotron Radiat*, **18**(Pt 3):387–97, 2011. doi:10.1107/S0909049511008235.

Lang, P. T., Holton, J. M., Fraser, J. S., and Alber, T. Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proc Natl Acad Sci U S A*, **111**(1):237–42, 2014. doi:10.1073/pnas.1302823110.

Lang, P. T., Ng, H. L., Fraser, J. S., Corn, J. E., Echols, N., Sales, M., Holton, J. M., and Alber, T. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Sci*, **19**(7):1420–31, 2010. doi:10.1002/pro.423.

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**(November):283–91, 1993. doi:10.1107/S0021889892009944.

Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D., and Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res*, **37**(17):5917–29, 2009. doi:10.1093/nar/gkp608.

Leal, R. M., 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. *J Synchrotron Radiat*, **20**(Pt 1):14–22, 2013. doi:10.1107/S0909049512049114.

Lebedev, A. A., Young, P., Isupov, M. N., Moroz, O. V., Vagin, A. A., and Murshudov, G. N. JLigand: a graphical tool for the CCP4 template-restraint library. *Acta Crystallogr D Biol Crystallogr*, **68**(Pt 4):431–40, 2012. doi:10.1107/S090744491200251X.

Leiros, H. K., Timmins, J., Ravelli, R. B., and McSweeney, S. M. Is radiation damage dependent on the dose rate used during macromolecular crystallography data collection? *Acta Crystallogr D Biol Crystallogr*, **62**(Pt 2):125–32, 2006. doi:10.1107/S0907444905033627.

Leiros, H. K. S., McSweeney, S. M., and Smalås, A. O. Atomic resolution structures of trypsin provide insight into structural radiation damage. *Acta Crystallographica Section D-Biological Crystallography*, **57**(4):488–97, 2001. doi:10.1107/S0907444901000646.

Leslie, A. G. W. and Powell, H. R. Processing diffraction data with MOSFLM. *Evolving Methods for Macromolecular Crystallography*, **245**:41–51, 2007. doi:10.1007/978-1-4020-6316-9{_}4.

Liu, H., Xu, Y., Li, F., Yang, Y., Wang, W., Song, Y., and Liu, D. Light-driven conformational switch of i-motif DNA. *Angew Chem Int Ed Engl*, **46**(14):2515–7, 2007. doi:10.1002/anie.200604589.

Lomb, L., Barends, T. R., Kassemeyer, S., Aquila, A., Epp, S. W., Erk, B., Foucar, L., Hartmann, R., Rudek, B., Rolles, D., Rudenko, A., Shoeman, R. L., Andreasson, J., Bajt, S., Barthelmess, M., Barty, A., Bogan, M. J., Bostedt, C., Bozek, J. D., Caleman, C., Coffee, R., Coppola, N., Deponte, D. P., Doak, R. B., Ekeberg, T., Fleckenstein, H., Fromme, P., Gebhardt, M., Graafsma, H., Gumprecht, L., Hampton, C. Y., Hartmann, A., Hauser, G., Hirsemann, H., Holl, P., Holton, J. M., Hunter, M. S., Kabsch, W., Kimmel, N., Kirian, R. A., Liang, M., Maia, F. R., Meinhart, A., Marchesini, S., Martin, A. V., Nass, K., Reich, C., Schulz, J., Seibert, M. M., Sierra, R., Soltau, H., Spence, J. C., Steinbrener, J., Stellato, F., Stern, S., Timneanu, N., Wang, X., Weidenspointner, G., Weierstall, U., White, T. A., Wunderer, C., Chapman, H. N., Ullrich, J., Struder, L., and Schlichting, I. Radiation damage in protein serial femtosecond crystallography using an x-ray free-electron laser. *Phys Rev B Condens Matter Mater Phys*, **84**(21):214111, 2011. doi:10.1103/PhysRevB.84.214111.

Long, F., Nicholls, R. A., Emsley, P., Graaeulis, S., Merkys, A., Vaitkus, A., and Murshudov, G. N. AceDRG: a stereochemical description generator for ligands. *Acta Crystallogr D Struct Biol*, **73**(Pt 2):112–22, 2017. doi:10.1107/S2059798317000067.

Lu, X. J. and Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, **31**(17):5108–21, 2003. doi:10.1093/nar/gkg680.

Marcélis, L., Moucheron, C., and Kirsch-De Mesmaeker, A. Ru-TAP complexes and DNA: from photo-induced electron transfer to gene photo-silencing in living cells. *Philos Trans A Math Phys Eng Sci*, **371**(1995):20120131, 2013. doi:10.1098/rsta.2012.0131.

Matlab. version 7.10.0 (R2010a). The MathWorks Inc., 2010.

Matsui, Y., Sakai, K., Murakami, M., Shiro, Y., Adachi, S., Okumura, H., and Kouyama, T. Specific damage induced by X-ray radiation and structural changes in the primary photoreaction of bacteriorhodopsin. *J Mol Biol*, **324**(3):469–81, 2002. doi:10.1016/S0022-2836(02)01110-5.

Matthews, B. W. Solvent content of protein crystals. *J Mol Biol*, **33**(2):491–7, 1968. doi:10.1016/0022-2836(68)90205-2.

McCarthy, A. A., Brockhauser, S., Nurizzo, D., Theveneau, P., Mairs, T., Spruce, D., Guijarro, M., Lesourd, M., Ravelli, R. B., and McSweeney, S. A decade of user operation on the macromolecular crystallography MAD beamline ID14-4 at the ESRF. *J Synchrotron Radiat*, **16**(Pt 6):803–12, 2009. doi:10.1107/S0909049509035377.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. Phaser crystallographic software. *J Appl Crystallogr*, **40**(Pt 4):658–74, 2007. doi:10.1107/S0021889807021206.

McGaughey, G. B., Gagne, M., and Rappe, A. K. pi-Stacking interactions. Alive and well in proteins. *J Biol Chem*, **273**(25):15458–63, 1998.

McGeehan, J., Ravelli, R. B., Murray, J. W., Owen, R. L., Cipriani, F., McSweeney, S., Weik, M., and Garman, E. F. Colouring cryo-cooled crystals: online microspectrophotometry. *J Synchrotron Radiat*, **16**(Pt 2):163–72, 2009. doi:10.1107/S0909049509001629.

McGeehan, J. E., Carpentier, P., Royant, A., Bourgeois, D., and Ravelli, R. B. X-ray radiation-induced damage in DNA monitored by online Raman. *J Synchrotron Radiat*, **14**(Pt 1):99–108, 2007. doi:10.1107/S0909049506043251.

McGeehan, J. E., Streeter, S. D., Thresh, S. J., Ball, N., Ravelli, R. B., and Kneale, G. G. Structural analysis of the genetic switch that regulates the expression of restriction-modification genes. *Nucleic Acids Res*, **36**(14):4778–87, 2008. doi:10.1093/nar/gkn448.

Meents, A., Dittrich, B., and Gutmann, S. A new aspect of specific radiation damage: hydrogen abstraction from organic molecules. *J Synchrotron Radiat*, **16**(Pt 2):183–90, 2009. doi:10.1107/S0909049509002192.

Meents, A., Gutmann, S., Wagner, A., and Schulze-Briese, C. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proc Natl Acad Sci U S A*, **107**(3):1094–9, 2010. doi:10.1073/pnas.0905481107.

Meharenna, Y. T., Doukov, T., Li, H., Soltis, S. M., and Poulos, T. L. Crystallographic and single-crystal spectral analysis of the peroxidase ferryl intermediate. *Biochemistry*, **49**(14):2984–6, 2010. doi:10.1021/bi100238r.

Merritt, E. A. To {\it B} or not to {\it B}: a question of resolution? *Acta Crystallogr D Biol Crystallogr*, **68**(Pt 4):468–77, 2012. doi:10.1107/S0907444911028320.

Michael, B. D. and O'Neill, P. Molecular biology. A sting in the tail of electron tracks. *Science*, **287**(5458):1603–4, 2000. doi:10.1126/science.287.5458.1603.

Michaud, M., Bazin, M., and Sanche, L. Measurement of inelastic cross sections for low-energy electron scattering from DNA bases. *Int J Radiat Biol*, **88**(1-2):15–21, 2012. doi:10.3109/09553002.2011.577505.

Mueller-Dieckmann, C., Panjikar, S., Tucker, P. A., and Weiss, M. S. On the routine use of soft X-rays in macromolecular crystallography. Part III. The optimal data-collection wavelength. *Acta Crystallogr D Biol Crystallogr*, **61**(Pt 9):1263–72, 2005. doi:10.1107/S0907444905021475.

Murray, J. and Garman, E. Investigation of possible free-radical scavengers and metrics for radiation damage in protein cryocrystallography. *J Synchrotron Radiat*, **9**(Pt 6):347–54, 2002. doi:10.1107/S0909049502014632.

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):513–22, 2004. doi:10.1107/S0021889804010660.

Murray, J. W., Rudiño-Piñera, E., Owen, R. L., Grininger, M., Ravelli, R. B., and Garman, E. F. Parameters affecting the X-ray dose absorbed by macromolecular crystals. *J Synchrotron Radiat*, **12**(Pt 3):268–75, 2005. doi:10.1107/S0909049505003262.

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 Crystallogr D Biol Crystallogr*, **67**(Pt 4):355–67, 2011. doi:10.1107/S0907444911001314.

Nanao, M. H., Sheldrick, G. M., and Ravelli, R. B. Improving radiation-damage substructures for RIP. *Acta Crystallogr D Biol Crystallogr*, **61**(Pt 9):1227–37, 2005. doi:10.1107/S0907444905019360.

Nass, K., Foucar, L., Barends, T. R., Hartmann, E., Botha, S., Shoeman, R. L., Doak, R. B., Alonso-Mori, R., Aquila, A., Bajt, S., Barty, A., Bean, R., Beyerlein, K. R., Bublitz, M., Drachmann, N., Gregersen, J., Jonsson, H. O., Kabsch, W., Kassemeyer, S., Koglin, J. E., Krumrey, M., Mattle, D., Messerschmidt, 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. *J Synchrotron Radiat*, **22**(2):225–38, 2015. doi:10.1107/S1600577515002349.

Nave, C. Radiation-Damage in Protein Crystallography. *Radiation Physics and Chemistry*, **45**(3):483–90, 1995. doi:10.1016/0969-806x(95)92800-E.

Nave, C. and Garman, E. F. Towards an understanding of radiation damage in cryocooled macromolecular crystals. *J Synchrotron Radiat*, **12**(Pt 3):257–60, 2005. doi:10.1107/S0909049505007132.

Nave, C. and Hill, M. A. Will reduced radiation damage occur with very small crystals? *J Synchrotron Radiat*, **12**(Pt 3):299–303, 2005. doi:10.1107/S0909049505003274.

Nave, C., Sutton, G., Evans, G., Owen, R., Rau, C., Robinson, I., and Stuart, D. I. Imperfection and radiation damage in protein crystals studied with coherent radiation. *J Synchrotron Radiat*, **23**(1):228–37, 2016. doi:10.1107/S1600577515019700.

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E., and Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, **406**(6797):752–7, 2000. doi:10.1038/35021099.

Nogly, P., James, D., Wang, D., White, T. A., Zatsepin, N., Shilova, A., Nelson, G., Liu, H., Johansson, L., Heymann, M., Jaeger, K., Metz, M., Wickstrand, C., Wu, W., Båth, P., Berntsen, P., Oberthuer, D., Panneels, V., Cherezov, V., Chapman, H., Schertler, G., Neutze, R., Spence, J., Moraes, I., Burghammer, M., Standfuss, J., and Weierstall, U. Lipidic cubic phase serial millisecond crystallography using synchrotron radiation. *IUCrJ*, **2**(Pt 2):168–76, 2015. doi:10.1107/S2052252514026487.

Nowak, E., Brzuszkiewicz, A., Dauter, M., Dauter, Z., and Rosenbaum, G. To scavenge or not to scavenge: that is the question. *Acta Crystallogr D Biol Crystallogr*, **65**(Pt 9):1004–6, 2009. doi:10.1107/S0907444909026821.

Nukaga, M., Mayama, K., Hujer, A. M., Bonomo, R. A., and Knox, J. R. Ultrahigh resolution structure of a class A beta-lactamase: on the mechanism and specificity of the extended-spectrum SHV-2 enzyme. *J Mol Biol*, **328**(1):289–301, 2003. doi:10.1016/S0022-2836(03)00210-9.

Oghbaey, S., Sarracini, A., Ginn, H. M., Pare-Labrosse, O., Kuo, A., Marx, A., Epp, S. W., Sherrell, D. A., Eger, B. T., Zhong, Y., Loch, R., Mariani, V., Alonso-Mori, R., Nelson, S., Lemke, H. T., Owen, R. L., Pearson, A. R., Stuart, D. I., Ernst, O. P., Mueller-Werkmeister, H. M., and Miller, R. J. Fixed target combined with spectral mapping: approaching 100rates for serial crystallography. *Acta Crystallogr D Struct Biol*, **72**(Pt 8):944–55, 2016. doi:10.1107/S2059798316010834.

O'Neill, P., Stevens, D. L., and Garman, E. F. Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective. *J Synchrotron Radiat*, **9**(Pt 6):329–32, 2002. doi:10.1107/S0909049502014553.

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., Ren, J., Stuart, D. I., and Evans, G. Outrunning free radicals in room-temperature macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, **68**(Pt 7):810–8, 2012. doi:10.1107/S0907444912012553.

Owen, R. L., Axford, D., Sherrell, D. A., Kuo, A., Ernst, O. P., Schulz, E. C., Miller, R. J., and Mueller-Werkmeister, H. M. Low-dose fixed-target serial synchrotron crystallography. *Acta Crystallogr D Struct Biol*, **73**(Pt 4):373–8, 2017. doi:10.1107/S2059798317002996.

Owen, R. L., Holton, J. M., Schulze-Briese, C., and Garman, E. F. Determination of X-ray flux using silicon pin diodes. *J Synchrotron Radiat*, **16**(Pt 2):143–51, 2009. doi:10.1107/S0909049508040429.

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 Crystallogr D Biol Crystallogr*, **70**(Pt 5):1248–56, 2014. doi:10.1107/S1399004714005379.

Owen, R. L., Rudiño-Piñera, E., and Garman, E. F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proc Natl Acad Sci U S A*, **103**(13):4912–7, 2006. doi:10.1073/pnas.0600973103.

Owen, R. L. and Sherrell, D. A. Radiation damage and derivatization in macromolecular crystallography: a structure factor's perspective. *Acta Crystallogr D Struct Biol*, **72**(Pt 3):388–94, 2016. doi:10.1107/S2059798315021555.

Owen, R. L., Yorke, B. A., Gowdy, J. A., and Pearson, A. R. Revealing low-dose radiation damage using single-crystal spectroscopy. *J Synchrotron Radiat*, **18**(Pt 3):367–73, 2011. doi:10.1107/S0909049511004250.

Paithankar, K. S. and Garman, E. F. Know your dose: RADDOSE. *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 4):381–8, 2010. doi:10.1107/S0907444910006724.

Paithankar, K. S., Owen, R. L., and Garman, E. F. Absorbed dose calculations for macromolecular crystals: improvements to RADDOSE. *J Synchrotron Radiat*, **16**(Pt 2):152–62, 2009. doi:10.1107/S0909049508040430.

Pearson, A. R., Pahl, R., Kovaleva, E. G., Davidson, V. L., and Wilmot, C. M. Tracking X-ray-derived redox changes in crystals of a methylamine dehydrogenase/amicyanin complex using single-crystal UV/Vis microspectrophotometry. *J Synchrotron Radiat*, **14**(Pt 1):92–8, 2007. doi:10.1107/S0909049506051259.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**:2825–30, 2011.

Penner-Hahn, J. E., Murata, M., Hodgson, K. O., and Freeman, H. C. Low-temperature x-ray absorption spectroscopy of plastocyanin: evidence for copper-site photoreduction at cryogenic temperatures. *Inorganic Chemistry*, **28**(10):1826–32, 1989. doi:10.1021/ic00309a015.

Perutz, M. F. Isomorphous Replacement and Phase Determination in Non-Centrosymmetric Space Groups. *Acta Crystallographica*, **9**(10):867–73, 1956. doi:10.1107/S0365110x56002485.

Petrova, T., Ginell, S., Mitschler, A., Kim, Y., Lunin, V. Y., Joachimiak, G., Cousido-Siah, A., Hazemann, I., Podjarny, A., Lazarski, K., and Joachimiak, A. X-ray-induced deterioration of disulfide bridges at atomic resolution. *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 10):1075–91, 2010. doi:10.1107/S0907444910033986.

Pike, A. C., Garman, E. F., Krojer, T., von Delft, F., and Carpenter, E. P. An overview of heavy-atom derivatization of protein crystals. *Acta Crystallogr D Struct Biol*, **72**(Pt 3):303–18, 2016. doi:10.1107/S2059798316000401.

Ptasińska, S. and Sanche, L. Dissociative electron attachment to hydrated single DNA strands. *Phys Rev E Stat Nonlin Soft Matter Phys*, **75**(3 Pt 1):031915, 2007. doi:10.1103/PhysRevE.75.031915.

Ramachandran, G. N. and Venkatachalam, C. M. Stereochemical Criteria for Polypeptides and Proteins .4. Standard Dimensions for Cis-Peptide Unit and Conformation of Cis-Polypeptides. *Biopolymers*, **6**(9):1255–62, 1968. doi:10.1002/bip.1968.360060903.

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 Crystallogr D Biol Crystallogr*, **61**(Pt 9):1289–98, 2005. doi:10.1107/S0907444905022316.

Rappaport, F. and Diner, B. A. Primary photochemistry and energetics leading to the oxidation of the (Mn)4Ca cluster and to the evolution of molecular oxygen in Photosystem II. *Coordination Chemistry Reviews*, **252**(3-4):259–72, 2008. doi:10.1016/j.ccr.2007.07.016.

Ravelli, R. B. and Garman, E. F. Radiation damage in macromolecular cryocrystallography. *Curr Opin Struct Biol*, **16**(5):624–9, 2006. doi:10.1016/j.sbi.2006.08.001.

Ravelli, R. B. and McSweeney, S. M. The 'fingerprint' that X-rays can leave on structures. *Structure*, **8**(3):315–28, 2000. doi:10.1016/S0969-2126(00)00109-X.

Ravelli, R. B. G., Leiros, H. K. S., Pan, B. C., Caffrey, M., and McSweeney, S. Specific radiation damage can be used to solve macromolecular crystal structures. *Structure*, **11**(2):217–24, 2003. doi:10.1016/S0969-2126(03)00006-6.

Ravelli, R. B. G., Theveneau, P., McSweeney, S., and Caffrey, M. Unit-cell volume change as a metric of radiation damage in crystals of macromolecules. *J Synchrotron Radiat*, **9**(6):355–60, 2002. doi:10.1107/S0909049502014541.

Read, R. J. Improved Fourier Coefficients for Maps Using Phases from Partial Structures with Errors. *Acta Crystallographica Section A*, **42**(3):140–9, 1986. doi:10.1107/S0108767386099622.

Read, R. J., Adams, P. D., Arendall, r., W. B., Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lutteke, T., Otwinowski, Z., Perrakis, A., Richardson, J. S., Sheffler, W. H., Smith, J. L., Tickle, I. J., Vriend, G., and Zwart, P. H. A new generation of crystallographic validation tools for the protein data bank. *Structure*, **19**(10):1395–412, 2011. doi:10.1016/j.str.2011.08.006.

Rice, L. M., Earnest, T. N., and Brunger, A. T. Single-wavelength anomalous diffraction phasing revisited. *Acta Crystallogr D Biol Crystallogr*, **56**(Pt 11):1413–20, 2000. doi:10.1107/S0907444900010039.

Roberts, B. R., Wood, Z. A., Jönsson, T. J., Poole, L. B., and Karplus, P. A. Oxidized and synchrotron cleaved structures of the disulfide redox center in the N-terminal domain of Salmonella typhimurium AhpF. *Protein Sci*, **14**(9):2414–20, 2005. doi:10.1110/ps.051459705.

Roedig, P., Duman, R., Sanchez-Weatherby, J., Vartiainen, I., Burkhardt, A., Warmer, M., David, C., Wagner, A., and Meents, A. Room-temperature macromolecular crystallography using a micro-patterned silicon chip with minimal background scattering. *J Appl Crystallogr*, **49**(Pt 3):968–75, 2016. doi:10.1107/S1600576716006348.

Sanche, L. Low energy electron-driven damage in biomolecules. *European Physical Journal D*, **35**(2):367–90, 2005. doi:10.1140/epjd/e2005-00206-6.

Sanishvili, R., Yoder, D. W., Pothineni, S. B., Rosenbaum, G., Xu, S., Vogt, S., Stepanov, S., Makarov, O. A., Corcoran, S., Benn, R., Nagarajan, V., Smith, J. L., and Fischetti, R. F. Radiation damage in protein crystals is reduced with a micron-sized X-ray beam. *Proc Natl Acad Sci U S A*, **108**(15):6127–32, 2011. doi:10.1073/pnas.1017701108.

Sarsero, J. P., Merino, E., and Yanofsky, C. A Bacillus subtilis gene of previously unknown function, yhaG, is translationally regulated by tryptophan-activated TRAP and appears to be involved in tryptophan transport. *Journal of Bacteriology*, **182**(8):2329–31, 2000. doi:10.1128/Jb.182.8.2329-2331.2000.

Scapin, G. Molecular replacement then and now. *Acta Crystallogr D Biol Crystallogr*, **69**(Pt 11):2266–75, 2013. doi:10.1107/S0907444913011426.

Schiltz, M. and Bricogne, G. Modelling and refining site-specific radiation damage in SAD/MAD phasing. *J Synchrotron Radiat*, **14**(Pt 1):34–42, 2007. doi:10.1107/S0909049506038970.

Schlichting, I., Berendzen, J., Chu, K., Stock, A. M., Maves, S. A., Benson, D. E., Sweet, R. M., Ringe, D., Petsko, G. A., and Sligar, S. G. The Catalytic Pathway of Cytochrome P450cam at Atomic Resolution. *Science*, **287**(5458):1615–22, 2000. doi:10.1126/science.287.5458.1615.

Schmidt, B., Ho, L., and Hogg, P. J. Allosteric disulfide bonds. *Biochemistry*, **45**(24):7429–33, 2006. doi:10.1021/bi0603064.

Sevilla, M. D., D'Arcy, J. B., and Morehouse, K. M. An electron spin resonance study of .gamma.-irradiated frozen aqueous solutions containing dipeptides. Mechanisms of radical reaction. *The Journal of Physical Chemistry*, **83**(22):2887–92, 1979.

Shevtsov, M. B., Streeter, S. D., Thresh, S. J., Swiderska, A., McGeehan, J. E., and Kneale, G. G. Structural analysis of DNA binding by C.Csp231I, a member of a novel class of R-M controller proteins regulating gene expression. *Acta Crystallogr D Biol Crystallogr*, **71**(Pt 2):398–407, 2015. doi:10.1107/S139900471402690X.

Shimazu, F., Kumta, U. S., and Tappel, A. L. Radiation Damage to Methionine and Its Derivatives. *Radiat Res*, **22**(2):276–87, 1964.

Shimizu, N., Hirata, K., Hasegawa, K., Ueno, G., and Yamamoto, M. Dose dependence of radiation damage for protein crystals studied at various X-ray energies. *J Synchrotron Radiat*, **14**(Pt 1):4–10, 2007. doi:10.1107/S0909049506049296.

Shimotsu, H., Kuroda, M. I., Yanofsky, C., and Henner, D. J. Novel form of transcription attenuation regulates expression the Bacillus subtilis tryptophan operon. *J Bacteriol*, **166**(2):461–71, 1986.

Simons, J. How do low-energy (0.1-2 eV) electrons cause DNA-strand breaks? *Accounts of Chemical Research*, **39**(10):772–9, 2006. doi:10.1021/ar0680769.

Skarina, T., Xu, X., Evdokimova, E., and Savchenko, A. High-Throughput Crystallization Screening, pages 159–168. Springer New York, New York, NY, 2014. ISBN 978-1-4939-0354-2. doi:10.1007/978-1-4939-0354-2_12.

Sliz, P., Harrison, S. C., and Rosenbaum, G. How does Radiation Damage in Protein Crystals Depend on X-Ray Dose? *Structure*, **11**(1):13–9, 2003. doi:10.1016/s0969-2126(02)00910-3.

Solar, S., Solar, W., and Getoff, N. Reactivity of Oh with Tyrosine in Aqueous-Solution Studied by Pulse-Radiolysis. *Journal of Physical Chemistry*, **88**(10):2091–5, 1984. doi:DOI10.1021/j150654a030.

Solomun, T. and Skalický, T. The interaction of a protein-DNA surface complex with low-energy electrons. *Chemical Physics Letters*, **453**(1-3):101–4, 2008. doi:10.1016/j.cplett.2007.12.078.

Southworth-Davies, R. J. and Garman, E. F. Radioprotectant screening for cryocrystallography. *J Synchrotron Radiat*, **14**(Pt 1):73–83, 2007. doi:10.1107/S0909049506044177.

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):1531–41, 2007. doi:10.1016/j.str.2007.10.013.

Sponer, J., Gabb, H. A., Leszczynski, J., and Hobza, P. Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. *Biophys J*, **73**(1):76–87, 1997. doi:10.1016/S0006-3495(97)78049-4.

Spotheim-Maurizot, M. and Davídková, M. Radiation damage to DNA in DNA-protein complexes. *Mutat Res*, **711**(1-2):41–8, 2011. doi:10.1016/j.mrfmmm.2011.02.003.

Štísová, V., Goffinont, S., Spotheim-Maurizot, M., and Davídková, M. Radiation damage to DNA-protein specific complexes: estrogen response element-estrogen receptor complex. *Radiation Protection Dosimetry*, **122**(1-4):106–9, 2006. doi:10.1093/rpd/ncl443.

Stubbe, J., Nocera, D. G., Yee, C. S., and Chang, M. C. Radical initiation in the class I ribonucleotide reductase: long-range proton-coupled electron transfer? *Chem Rev*, **103**(6):2167–201, 2003. doi:10.1021/cr020421u.

Sutton, K. A., Black, P. J., Mercer, K. R., Garman, E. F., Owen, R. L., Snell, E. H., and Bernhard, W. A. Insights into the mechanism of X-ray-induced disulfide-bond cleavage in lysozyme crystals based on EPR, optical absorption and X-ray diffraction studies. *Acta Crystallogr D Biol Crystallogr*, **69**(Pt 12):2381–94, 2013. doi:10.1107/S0907444913022117.

Sygusch, J. and Allaire, M. Sequential radiation damage in protein crystallography. *Acta Crystallogr A*, **44 (Pt 4)**(4):443–8, 1988. doi:10.1107/S0108767388001394.

Symons, M. C. R. Electron movement through proteins and DNA. *Free Radical Biology and Medicine*, **22**(7):1271–6, 1997. doi:10.1016/S0891-5849(96)00548-5.

Taylor, G. L. Introduction to phasing. *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 4):325–38, 2010. doi:10.1107/S0907444910006694.

Ten Eyck, L. F. Crystallographic fast Fourier transforms. *Acta Crystallographica Section A*, **29**(2):183–91, 1973. doi:10.1107/s0567739473000458.

Teng, T. Y. Mounting of crystals for macromolecular crystallography in a free-standing thin film. *Journal of Applied Crystallography*, **23**(5):387–91, 1990. doi:10.1107/s0021889890005568.

Terwilliger, T. C., Bunkóczi, G., Hung, L. W., Zwart, P. H., Smith, J. L., Akey, D. L., and Adams, P. D. Can I solve my structure by SAD phasing? Planning an experiment, scaling data and evaluating the useful anomalous correlation and anomalous signal. *Acta Crystallogr D Struct Biol*, **72**(Pt 3):359–74, 2016. doi:10.1107/S2059798315019403.

Théodore, M., Sobczyk, M., and Simons, J. Cleavage of thymine N3âĂŞH bonds by low-energy electrons attached to base ĬĂâĹŮ orbitals. *Chemical Physics*, **329**(1-3):139–47, 2006. doi:10.1016/j.chemphys.2006.05.012.

Tickle, I. J. Statistical quality indicators for electron-density maps. *Acta Crystallogr D Biol Crystallogr*, **68**(Pt 4):454–67, 2012. doi:10.1107/S0907444911035918.

Tronrud, D. E. Introduction to macromolecular refinement. *Acta Crystallogr D Biol Crystallogr*, **60**(Pt 12 Pt 1):2156–68, 2004. doi:10.1107/S090744490402356X.

Tsai, A. L. and Kulmacz, R. J. Prostaglandin H synthase: resolved and unresolved mechanistic issues. *Arch Biochem Biophys*, **493**(1):103–24, 2010. doi:10.1016/j.abb.2009.08.019.

Urzhumtseva, L., Afonine, P. V., Adams, P. D., and Urzhumtsev, A. Crystallographic model quality at a glance. *Acta Crystallogr D Biol Crystallogr*, **65**(Pt 3):297–300, 2009. doi:10.1107/S0907444908044296.

Usón, I., Stevenson, C. E. M., Lawson, D. M., and Sheldrick, G. M. Structure determination of the O-methyltransferase NovP using the 'free lunch algorithm' as implemented in SHELXE. *Acta Crystallogr D Biol Crystallogr*, **63**:1069–74, 2007.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F., and Murshudov, G. N. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr D Biol Crystallogr*, **60**(Pt 12 Pt 1):2184–95, 2004. doi:10.1107/S0907444904023510.

Wagner, A., Duman, R., Henderson, K., and Mykhaylyk, V. In-vacuum long-wavelength macromolecular crystallography. *Acta Crystallogr D Struct Biol*, **72**(Pt 3):430–9, 2016. doi:10.1107/S2059798316001078.

Warkentin, M., Badeau, R., Hopkins, J. B., and Thorne, R. E. Spatial distribution of radiation damage to crystalline proteins at 25-300 K. *Acta Crystallogr D Biol Crystallogr*, **68**(Pt 9):1108–17, 2012. doi:10.1107/S0907444912021361.

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 Crystallogr D Struct Biol*, **72**(Pt 4):558–75, 2016. doi:10.1107/S2059798316002187.

Weik, M., Bergès, J., Raves, M. L., Gros, P., McSweeney, S., Silman, I., Sussman, J. L., Houée-Levin, C., and Ravelli, R. B. G. Evidence for the formation of disulfide radicals in protein crystals upon X-ray irradiation. *Journal of Synchrotron Radiation*, **9**(6):342–6, 2002. doi:10.1107/S0909049502014589.

Weik, M., Ravelli, R. B., 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. *Proc Natl Acad Sci U S A*, **97**(2):623–8, 2000.

Weik, M., Ravelli, R. B., 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 Sci*, **10**(10):1953–61, 2001. doi:10.1110/ps.09801.

Whittaker, J. W. Free radical catalysis by galactose oxidase. *Chem Rev*, **103**(6):2347–63, 2003. doi:10.1021/cr020425z.

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., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*, **67**(Pt 4):235–42, 2011. doi:10.1107/S0907444910045749.

Wisniowski, P., Carmichael, I., Fessenden, R. W., and Hug, G. L. Evidence for beta scission in the oxidation of amino acids. *Journal of Physical Chemistry A*, **106**(18):4573–80, 2002. doi:10.1021/jp0139383.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*, **34**(2):564–74, 2006. doi:10.1093/nar/gkj454.

Yang, M., de Saizieu, A., van Loon, A. P., and Gollnick, P. Translation of trpG in Bacillus subtilis is regulated by the trp RNA-binding attenuation protein (TRAP). *Journal of Bacteriology*, **177**(15):4272–8, 1995.

Yano, J., Kern, J., Irrgang, K. D., Latimer, M. J., Bergmann, U., Glatzel, P., Pushkar, Y., Biesiadka, J., Loll, B., Sauer, K., Messinger, J., Zouni, A., and Yachandra, V. K. X-ray damage to the Mn4Ca complex in single crystals of photosystem II: a case study for metalloprotein crystallography. *Proc Natl Acad Sci U S A*, **102**(34):12047–52, 2005. doi:10.1073/pnas.0505207102.

Yorke, B. A., Beddard, G. S., Owen, R. L., and Pearson, A. R. Time-resolved crystallography using the Hadamard transform. *Nat Methods*, **11**(11):1131–4, 2014. doi:10.1038/nmeth.3139.

Zeldin, O. B., Brockhauser, S., Bremridge, J., Holton, J. M., and Garman, E. F. Predicting the X-ray lifetime of protein crystals. *Proc Natl Acad Sci U S A*, **110**(51):20551–6, 2013a. doi:10.1073/pnas.1315879110.

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):1225–30, 2013b. doi:10.1107/S0021889813011461.

Zubavichus, Y., Zharnikov, M., Shaporenko, A., Fuchs, O., Weinhardt, L., Heske, C., Umbach, E., Denlinger, J. D., and Grunze, M. Soft X-ray induced decomposition of phenylalanine and tyrosine: A comparative study. *Journal of Physical Chemistry A*, **108**(20):4557–65, 2004. doi:10.1021/jp049376f.

Zwart, P. H., Banumathi, S., Dauter, M., and Dauter, Z. Radiation-damage-induced phasing with anomalous scattering: substructure solution and phasing. *Acta Crystallogr D Biol Crystallogr*, **60**(Pt 11):1958–63, 2004. doi:10.1107/S0907444904021730.

APPENDIX A

Bovine insulin data collection

Table A.1: MX data collection information for a bovine insulin dataset supplied by Dr. J. C. Brooks-Bartlett. The accumulated diffraction-weighted dose was calculated by *RADDOSE-3D*.

Beamline	P14, PETRA III, Hamburg
X-ray energy (keV)	12.7
Beam collimation (μm × μm)	140 × 140
Beam profile	Top Hat
Crystal dimensions (μm × μm × μm)	100 × 100 × 100
Number of images	900
Rotation per image (°)	0.1
Exposure time per image (s)	0.5
Diffraction-weighted dose (MGy)	0.89

For the work in Section 7.3, Dr. J. C. Brooks-Bartlett had supplied a single MX dataset (in the form of an unmerged `.mtz` file and refined coordinate model), corresponding to data collected from a single bovine insulin crystal. The dataset corresponded to 900 images collected over 90° (oscillation of 0.1°) per image (Table A.1). All data processing and model building was performed by J. C. Brooks-Bartlett, as reported in detail in Chapter 2 of the D.Phil thesis of Brooks-Bartlett. In short, *MOSFLM* was used to integrate images and the CCP4i data reduction pipeline (*AIMLESS* and *CTRUNCATE*) was used generate structure factor amplitudes. Molecular replacement was performed using *PHASER* to generate an initial phase set, with the Nanao *et al.* (2005) bovine insulin coordinate model (PDB: *2bn3*) used as a search model. Restrained refinement was performed using *REFMAC*, and the final coordinate model was refined to a maximum resolution of 1.4 Å. Table A.2 reports the data processing statistics for this supplied dataset.

Table A.2: Data processing and refinement statistics for bovine insulin dataset supplied by Dr. J. C. Brooks-Bartlett. The resolution range is $31.82 - 1.40$ Å, with values in parentheses corresponding to the highest-resolution shell ($1.42 - 1.40$ Å). The Wilson *B*-factor is estimated from the Wilson plot, as reported in each *AIMLESS* log file. Refinement was performed Dr. J. C. Brooks-Bartlett using *REFMAC*, full details of which can be found in Chapter 2 of the D.Phil thesis of Brooks-Bartlett. For observed F_{obs} and calculated F_{calc} structure factors, $R_{\text{work}} = \frac{\sum_{hkl} |F_{\text{obs}} - F_{\text{calc}}|}{\sum_{hkl} F_{\text{obs}}}$ and R_{free} is the R_{work} formula calculated from a small ($5\,\%$) test set of randomly selected reflections that were excluded throughout refinement. Both R_{work} and R_{free} were calculated using *REFMAC*.

Space group	$I2_13$
Unit-cell parameters	
$a = b = c$ (Å)	77.95
$\alpha = \beta = \gamma$ (°)	90
Total no. reflections	151103 (7528)
No. unique reflections	15673 (784)
Completeness (%)	99.9 (100)
Multiplicity	9.6 (9.6)
$\langle I/\sigma(I) \rangle$	33.2 (6.7)
R_{merge}	0.031 (0.288)
R_{meas}	0.033 (0.305)
$\text{CC}_{1/2}$	1.000 (0.970)
Wilson *B*-factor (Å^2)	14.1
Mosaicity (°)	0.42
Refinement	
R_{work}	0.156
R_{free}	0.183
Mean isotropic *B*-factor (Å^2)	18.9

APPENDIX B

Poster presentations

MX radiation damage to nucleoprotein complexes



Charles S. Bury[1], John E. McGeehan[2], Ian Carmichael[3], Markus Gerstel[1], Mikhail B. Shevtsov[4], Alfred A. Antson[5], Elspeth F. Garman[1]

[1]Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU [2]Biophysics Laboratories, Institute of Biomedical and Biomolecular Sciences, University of Portsmouth, King Henry I Street, Portsmouth, Hampshire PO1 2DY, UK [3]Notre Dame Radiation Laboratory, University of Notre Dame, Notre Dame, IN 46556, USA [4]Laboratory of Structural Biology of GPCRs, Moscow Institute of Physics and Technology, Dolgoprudniy 141700, Russia [5]York Structural Biology Laboratory, Department of Chemistry, University of York, York Y010 5DD, UK

Contact: charles.bury@dtc.ox.ac.uk

Radiation damage is still an inherent problem in MX diffraction data collection, even at cryo-temperatures (100 K)[1]

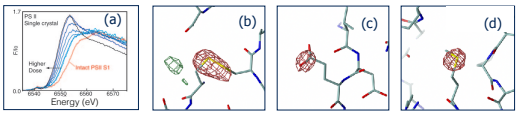
Global damage observed in reciprocal space in the diffraction pattern as:

- Overall loss of reflection intensities
- Loss of resolution
- Unit cell expansion
- Usually increasing mosaicity

Specific damage involves radiation-induced chemical and conformation changes to macromolecular structure. A reproducible order has been shown in proteins[2,3,4]:

- **(a)** Metallo-centre reduction
- **(b)** Disulphide bond cleavage
- **(c)** Acidic residue decarboxylation
- **(d)** Methionine C-S bond cleavage

(Below) (a) Mn XANES of single crystal of photosystem II as a function of X-ray dose[5] **(b-d)** $F_{o,1}$ - $F_{o,1}$ Fourier difference maps contoured at ±4σ for Torpedo californica acetyl-cholinesterase (TcAChE) pdb: 1QID[4]



Few MX investigations exist on **specific damage manifestations** for crystalline nucleic acids in their **complexes** with protein

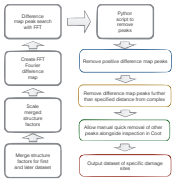
Protein-DNA radiation damage

The experiment

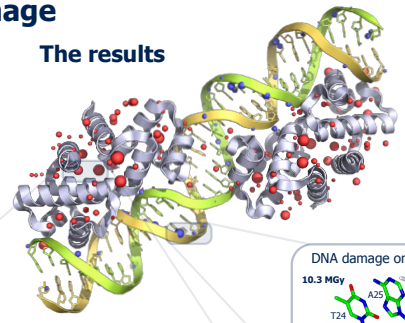
- Quantitative investigation[6] of specific damage mechanisms on model protein-DNA complex C.Esp1396I[7]
- 2.8 Å diffraction data collected at 100 K over large dose range (2.07–44.63 MGy)

The method

- Fourier difference maps ($F_{o,n}$ - $F_{o,1}$) between initial dose and higher dose datasets calculated to localise regions of electron density loss/disordering
- Difference map peak searching pipeline:
  1. Filter Fourier difference maps
  2. Remove noisy background peaks
  3. Locate sites of specific damage

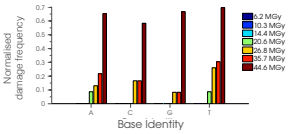


The results

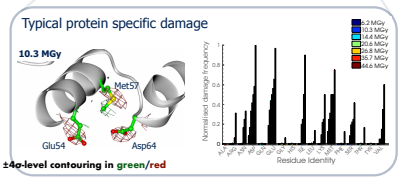


(Left) Detected damage sites within C-protein complex at 44.6 MGy Damage sites represented as spheres (radius proportional to electron density loss). Protein damage in red, damage within 2 Å of DNA in blue

(Below) Normalised frequency of detected specific damage against nucleotide type for each dose

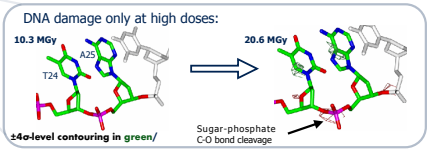


Typical protein specific damage



±4σ-level contouring in green/red

DNA damage only at high doses:

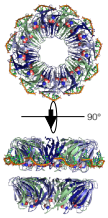


±4σ-level contouring in green/

Open Questions

- Is protein intrinsically more susceptible to damage or acting as electron scavenger to protect DNA?
- DNA phosphodiester bond damage localised to strained AT-rich DNA region. Does strained geometry enhance radiation damage effects in DNA?

Protein-RNA radiation damage



The aim

- To investigate effects of nucleic acid binding to a protein on specific damage susceptibility

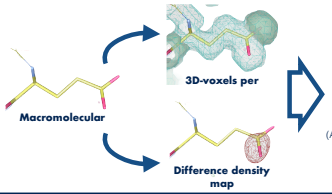
The experiment

- Model complex trp RNA-binding attenuation protein (TRAP) bound to 53 base RNA[8]
- Crystallises in 1-1 ratio of RNA-bound and non-bound TRAP within asymmetric unit, allowing controlled investigation under identical crystal conditions
- 10 sets of 1.98 Å diffraction data collected on same crystal at 100 K over large dose range (1.31 - 25.0 MGy)

The method

Per-atom damage quantification

Calculate damage metrics for each refined non-hydrogen atom in pdb file to locate specific damage sites



$\mathfrak{D}_{loss}$ (Max density loss per atom)

$\mathfrak{D}_{mean}$ (Average density change per atom)

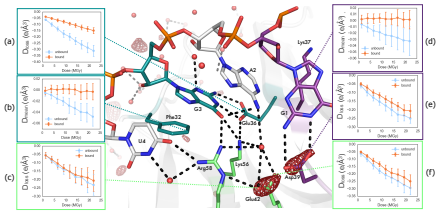
The results

- Glu/Asp decarboxylation rate highly dependent on proximity to RNA



(Left) D_{loss} calculated for Glu/Asp side-chain carboxyl oxygens within TRAP at 16.7 MGy. Each bar represents the mean calculated over 11 equivalent atoms around either RNA-bound or non-bound TRAP

- RNA binding reduces damage within binding interfaces



(Left) Reduction in decarboxylation rate most significant for Glu 36 **(a)** and Asp 39 **(e)** hydrogen bonding directly to RNA bases

Strong Glu 42 salt bridge interaction **(c)** preserved upon RNA-binding, so decarboxylation rate not reduced

Direct interactions with RNA bases stabilises Phe **(b)** and Lys **(d)** side-chains through physical stacking

Open Questions

- Is the RNA-protection physically or chemically mediated?
- Mechanism for enhanced damage susceptibility to residues forming crystal contacts?

UNIVERSITY OF OXFORD

Systems Biology Doctoral Training Centre

EPSRC Engineering and Physical Sciences Research Council

References

1. Garman, EF (2010). Acta Cryst. D, 66(PT 4), 339-51
2. Weik M, et al., (2000). PNAS, 97(2), 623-8
3. Ravelli RB & McSweeney SM (2000). Structure, 8(3), 315-28
4. Burmeister WP (2000). Acta Cryst. D, 56(3), 328-341
5. Yano J, et al., (2005). PNAS, 102(34),12047-52
6. Bury C, et al., (2015). J. Synchrotron Radiat. 22, 213-224
7. McGeehan JE, et al., (2008). Nucleic Acids Res. 36(14), 4778-87
8. Hopcroft NH, et al., (2002). Nature 58(4), 615-621

Figure B.1: Poster presentation at the 29th annual European Crystallographic Meeting (ECM29), Rovinj, Croatia, taking place on 23-28th August, 2015.

Tracking MX radiation damage in nucleoprotein complexes



Charles S. Bury[1], John E. McGeehan[2], Ian Carmichael[3], Markus Gerstel[1], Mikhail B. Shevtsov[4], Alfred A. Antson[5], Elspeth F. Garman[1]

[1]Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU [2]Biophysics Laboratories, Institute of Biomedical and Biomolecular Sciences, University of Portsmouth, King Henry I Street, Portsmouth, Hampshire PO1 2DY, UK [3]Notre Dame Radiation Laboratory, University of Notre Dame, Notre Dame, IN 46556, USA [4]Laboratory of Structural Biology of GPCRs, Moscow Institute of Physics and Technology, Dolgoprudniy 141700, Russia [5]York Structural Biology Laboratory, Department of Chemistry, University of York, York Y010 5DD, UK

Contact: charles.bury@dtc.ox.ac.uk

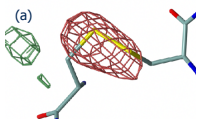
Radiation damage is still an inherent problem in MX diffraction data collection, even at cryo-temperatures (100 K)[1]

Specific damage involves radiation-induced chemical and conformation changes to macromolecular structure

Occurs before any observable loss in diffraction resolution

A reproducible order of damage is predicted within non-metalloproteins[2,3,4]

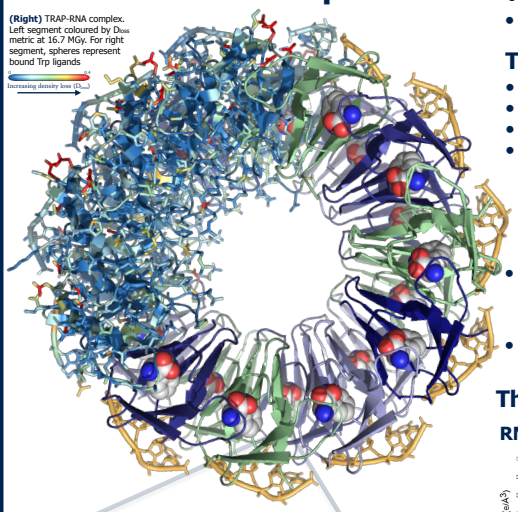
Disulphide bond cleavage | **Glu/Asp decarboxylation** | **Met S-C bond cleavage**



(Left) **(a-c)** F_{oS} - F_{o1} Fourier difference maps contoured at ±4σ for Torpedo californica acetyl-cholinesterase (TcAChE) pdb: 1QID[4]

Few MX investigations exist on **specific damage manifestations** for crystalline nucleic acids in **complex** with protein

The TRAP-RNA complex



(Right) TRAP-RNA complex. Left segment coloured by D_{loss} metric at 16.7 MGy. For right segment, spheres represent bound Trp ligands

Increasing density loss ($\mathfrak{D}_{loss}$)

The aim

- To quantify the effects of nucleic acid binding to a protein on specific damage susceptibility

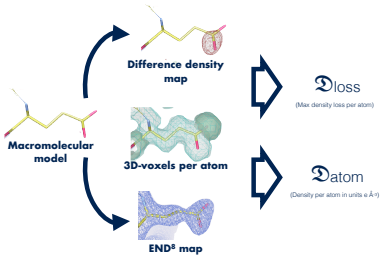
The experiment[5]

- Model complex: trp RNA-binding attenuation protein (TRAP) bound to 53 base RNA[6]
- 11 identical monomers per TRAP ring
- Glu & Asp rich (no Cysteines)
- 10 sets of 1.98 Å diffraction data collected on same crystal at 100 K over large dose range (DWD[7]: 1.31 - 25.0 MGy)

The method

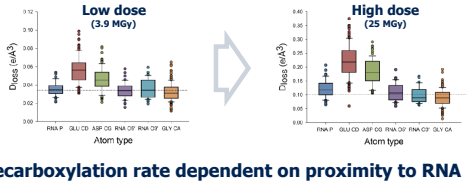
Per-atom damage quantification

- Calculate damage metrics based on electron density loss for each refined non-hydrogen atom in PDB file
- Pipeline developed (ETRACK) to locate specific damage sites



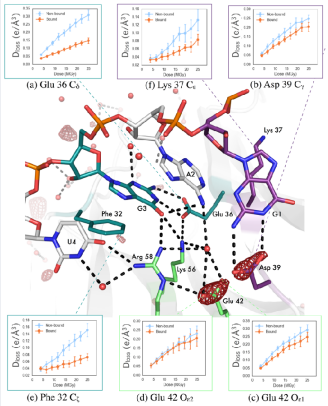
The Results

RNA is relatively radiation-insensitive



D_{loss} for radiation-sensitive protein atoms and RNA phosphate group atoms, for low dose **(far left)** and high dose **(left)** datasets

RNA binding reduces damage within binding interfaces



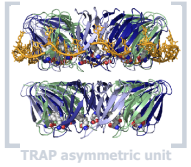
(Above) Reduction in decarboxylation rate most significant for Glu 36 **(a)** and Asp 39 **(b)** which hydrogen bond directly to RNA bases

Strong Glu 42 salt bridge interaction **(d)** preserved upon RNA-binding, so decarboxylation rate not reduced

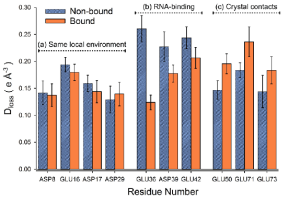
Direct interactions with RNA bases stabilises Phe **(e)** and Lys **(f)** side-chains through physical stacking

Glu/Asp decarboxylation rate dependent on proximity to RNA

TRAP crystallises in 1-1 ratio of RNA-bound and non-bound TRAP within asymmetric unit



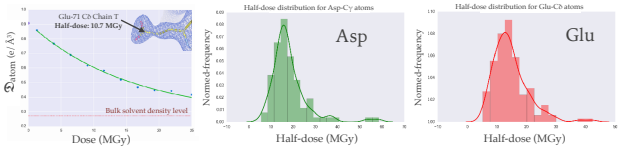
Allowed controlled investigation under identical crystal conditions



D_{loss} calculated for Glu/Asp side-chain carboxyl oxygens within TRAP at 16.7 MGy. Each bar represents the mean value calculated over 11 equivalent atoms around either RNA-bound or non-bound TRAP

Glu/Asp half-dose distributions

- Determination of dose at which the density attributed to an atom decays to half the initial value



D_{atom} half dose distributions for Glu and Asp. Glu & Asp have qualitatively similar half-dose distributions since both damage through decarboxylation mechanism

References

1. Garman, EF (2010). Acta Cryst. D, 66(Pt 4), 339–51
2. Weik M, et al., (2000). PNAS, 97(2), 623–8
3. Ravelli RB & McSweeney SM (2000). Structure, 8(3), 315–28
4. Burmeister WP (2000). Acta Cryst. D, 56(3), 328–341
5. Bury CS, et al., (2016). Acta Cryst. D, in press
6. Hozcroft NH, et al., (2002). Nature 58(4), 615–621
7. Zeldin PT, et al., (2013). PNAS 110(51), 20551-20556
8. Lang PT, et al., (2014). PNAS 111(1), 237-242

UNIVERSITY OF OXFORD

Systems Biology Doctoral Training Centre

EPSRC Engineering and Physical Sciences Research Council

Figure B.2: Poster presentation at the 9th International Workshop on X-ray Radiation Damage to Biological Crystalline Samples (RD9), Lund, Sweden, taking place on 9-11th March 2016.



Contact: charles.bury@dtc.ox.ac.uk
Github: https://github.com/charliebury

OH loss from Tyrosine: debunking a myth

Charles S. Bury[1], Ian Carmichael[2], Elspeth F. Garman[1]

[1]Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU
[2]Notre Dame Radiation Laboratory, University of Notre Dame, Notre Dame, IN 46556, USA

The Aim

To systematically characterise radiation-induced Tyr -OH electron density loss in general proteins

Specific Damage Trends

Radiation damage remains an inherent problem in MX diffraction data collection, even at cryo-temperatures (100 K)[1]

Specific damage involves localised radiation-induced chemical changes to macromolecular structure. It is known to follow a reproducible order in proteins at 100 K[2] (**left**)

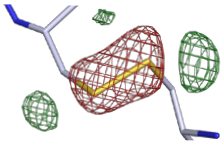
Tyr -OH electron density loss observed in $F_{obs}(n)$ - $F_{obs}(1)$ difference maps for a myrosinase (pdb: 1dwa) crystal at 100 K[3]

Tyr -OH cleavage is also commonly quoted as being amongst the most dominant specific damage mechanisms at 100 K.

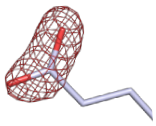
But:

1. Phenolic C–O bond cleavage has not been reported in the proposed mechanisms of Tyr radiation chemistry
2. Reported observations of Tyr -OH electron density loss in general protein systems is lacking

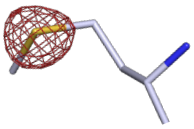
(Below)
$F_{obs}(4)$ - $F_{obs}(1)$ Fourier difference maps for myrosinase (pdb: 1dwa[3], 2.0 Å). All maps contoured at ±5σ (in green/red)



Disulphide Cleavage



Glu/Asp Decarboxylation



Methylthiol Cleavage



Tyr -OH loss

Analysis of PDB damage series

Complete collection of full MX damage series conducted at 100 K collated

.pdb and .mtz coordinate files retrieved from PDB_REDO server[4] prior to model rebuilding

RIDL program (**left**) used to calculate per-atom damage metric D_{loss} for each structure at each dose

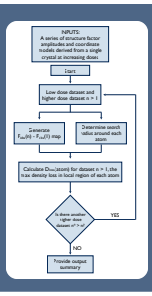
RIDL

Radiation-Induced Density Loss

Availability: https://github.com/GarmanGroup

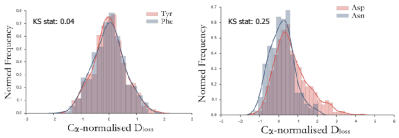
A tool to enable **objective** and **systematic** quantification of per-atom electron density loss for **generalised** protein crystals

Calculates a per-atom damage metric D_{loss}(atom), derived from $F_{obs}(n)$ - $F_{obs}(1)$ difference maps[5]



Results

Aromatic ring disordering



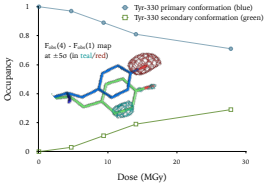
(**above**) Phe and Tyr undergo comparable radiation-induced aromatic ring disordering, quantified by Kolmogorov-Smirnov (KS) statistic (malate dehydrogenase[6] shown)

(**right**) Original observations of Tyr -OH electron density loss in myrosinase[3] (pdb: 1dwa) consistent with full aromatic group displacement. Occupancy of modelled secondary conformation (in green) increases as function of dose.

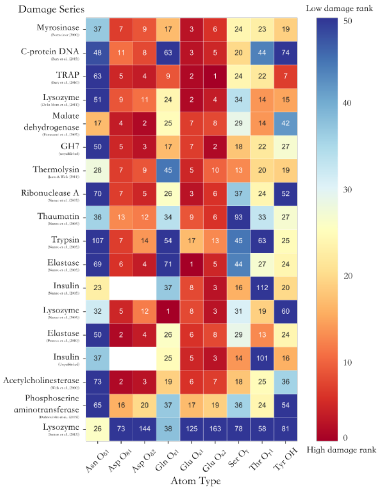
Ranking specific damage

(**right**) Glu and Asp decarboxylation detected across the range of protein damage series at 100 K

Tyr-OH electron density loss in general significantly lower than for Glu and Asp carboxylates, for doses > 1 MGy at which decarboxylation events were detectable



(**right**) For each damage series, each atom type A (e.g. the set of Tyr-OH atoms) is ranked with respect to the mean D_{loss} values attained by that atom type at each dose ($R_{A,i}$ for i = 2,..., n). The products of these per-dataset ranks ($R_{A,2}$ x $R_{A,3}$ x ... x $R_{A,n}$) are then ranked to provide an across-dataset rank for each atom type. The ranking scheme starts at 0 for the most damaged atom species within each protein.



Conclusion

Tyr is not generally radiation-sensitive in proteins at 100 K

RIDL provides an objective quantification of specific radiation damage in general protein systems

References

1. Garman, EF (2010). Acta Cryst. D, 66(4), 339–51
2. Weik M, et al., (2000). PNAS, 97(2), 623–8
3. Burmeister WP (2000). Acta Cryst. D, 56(3), 328–341
4. Joosten RP, et al., (2014). IUCrJ 1(4):213–20
5. Bury CS, et al., (2016). Acta Cryst. D, 72(5), 648–57
6. Fioravanti E, et al., (2007). J. Synch. Rad. 14(Pt 1), 84–91

Acknowledgements

We thank John McGeehan (Uni. of Portsmouth) for kindly supplying GH7 family protein crystals and Jonathan Brooks-Bartlett (Uni. of Oxford) for providing access to his complete insulin damage series data at 100 K, for use in the current investigation.

Figure B.3: Poster presentation at the Gordon Research Conference (GRC) on Diffraction Methods in Structural Biology, Lewiston, ME, USA, taking place on 17-22th July 2016.

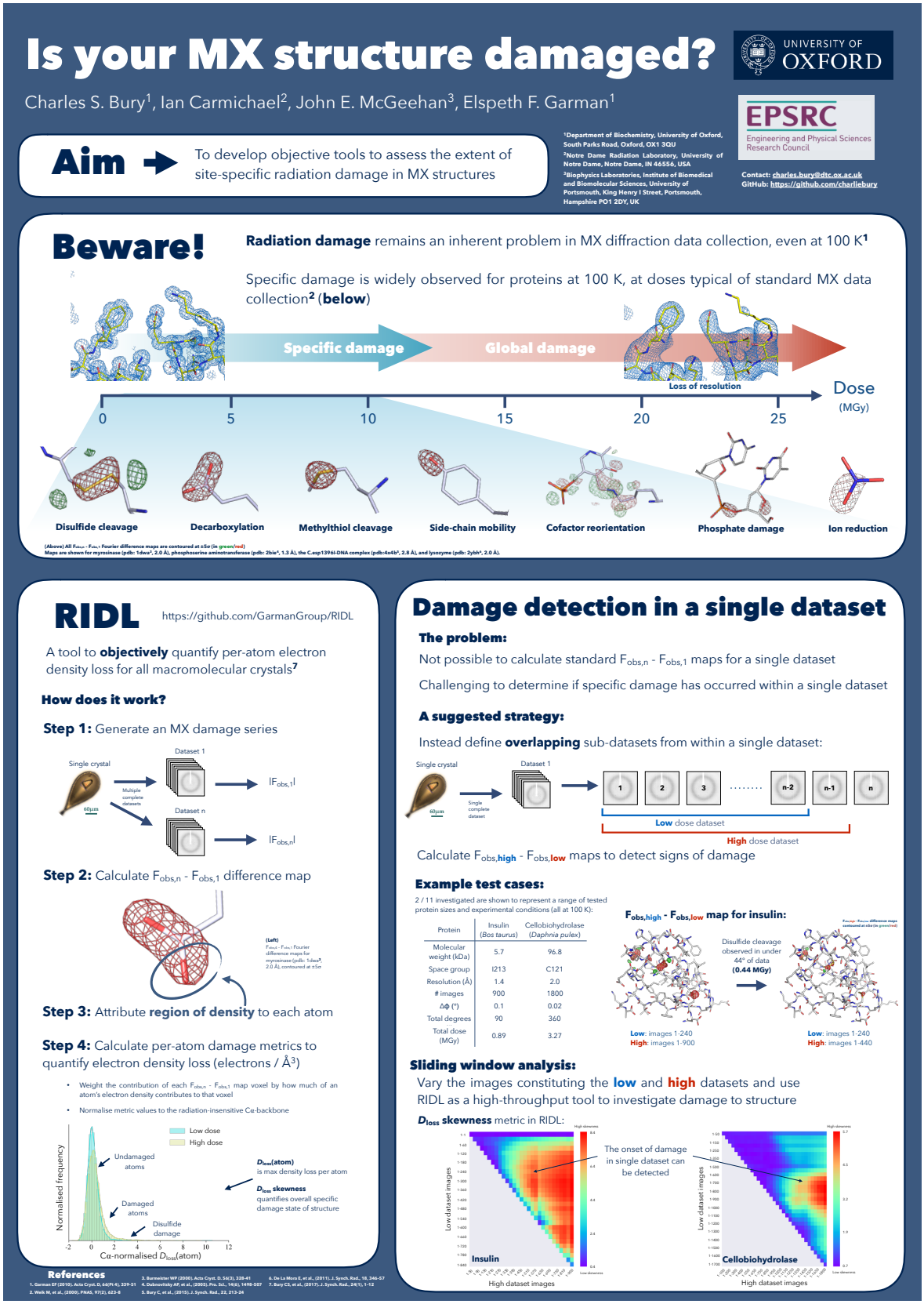


Figure B.4: Poster presentation at the 67th annual American Crystallographic Association meeting (ACA), New Orleans, ME, USA, taking place on 26-30th May 2017.

APPENDIX C

Journal covers

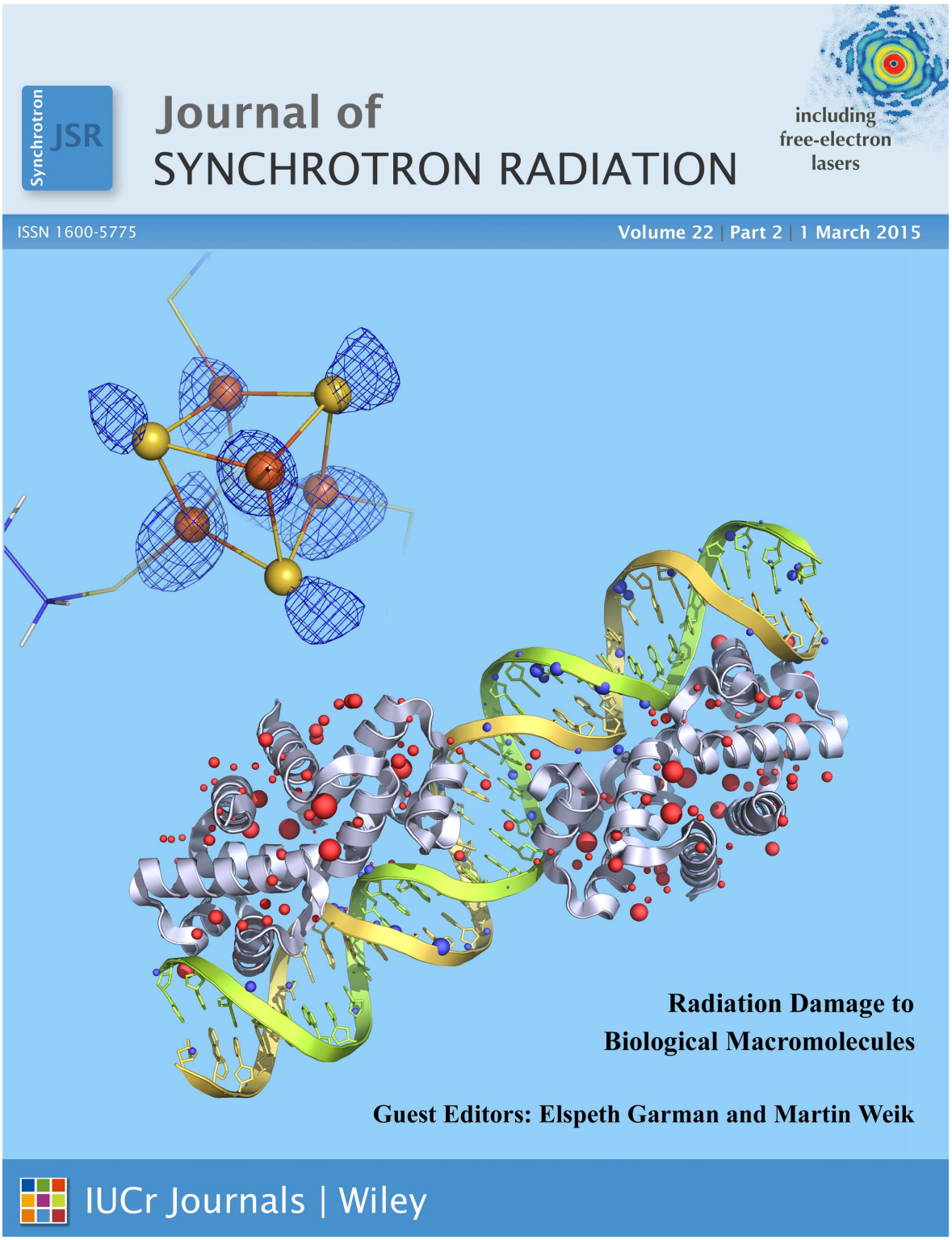


Figure C.1: Front cover image for Journal of Synchrotron Radiation 2015 March edition, co-created with Dr. K. Nass, Max Planck Institute for Medical Research, Germany: Specific radiation damage to an iron-sulphur cluster in ferredoxin caused by ultra-short and intense X-FEL pulses (upper panel; see (Nass *et al.*, 2015)) and specific radiation damage to the protein moiety of a protein-DNA complex cased by synchrotron radiation (lower panel; see Bury *et al.* (2015)).

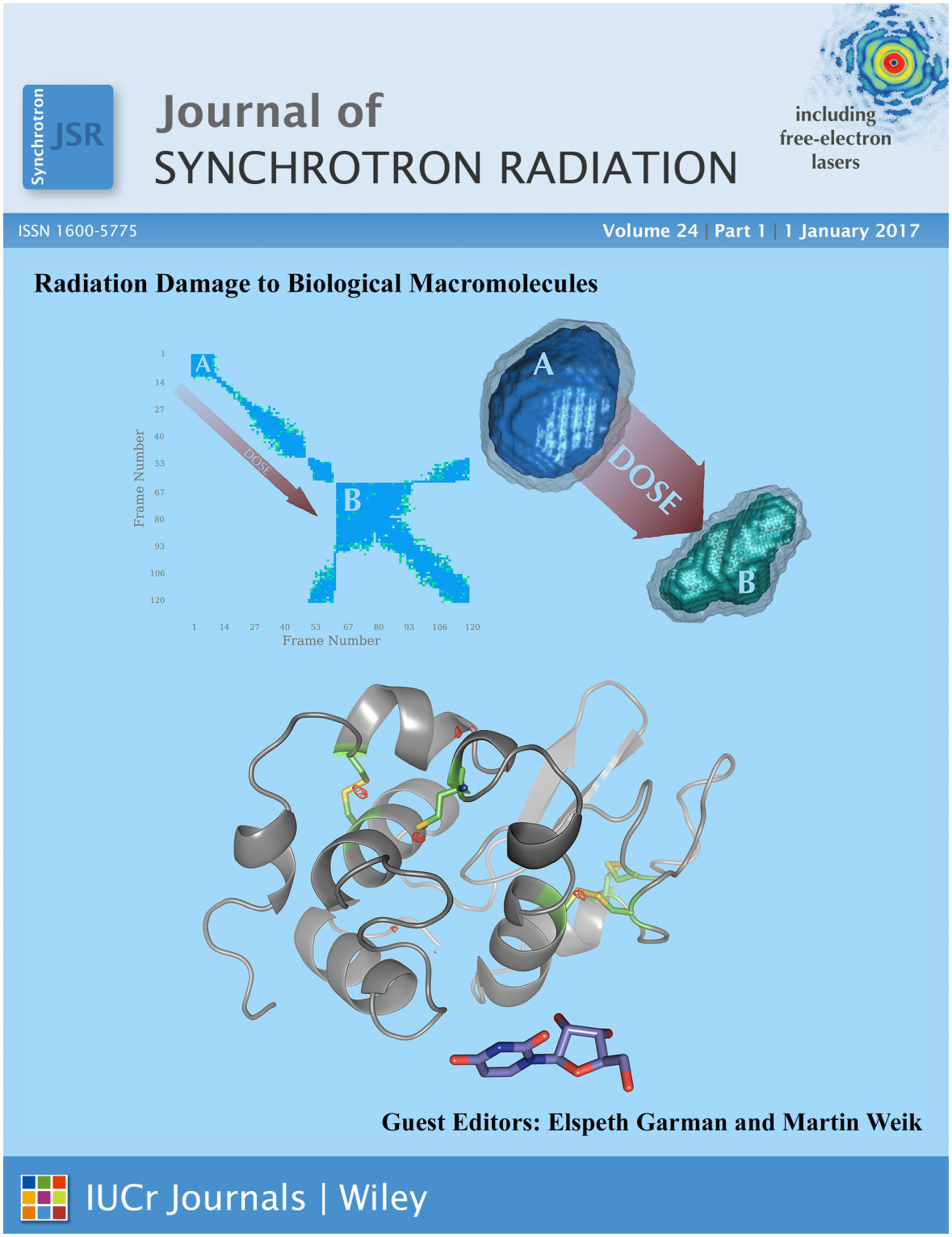


Figure C.2: Front cover image for Journal of Synchrotron Radiation 2017 March edition, co-created with J. C. Brooks-Bartlett, University of Oxford and E. Crosas, ALBA Synchrotron, Carrer de la Ilum.

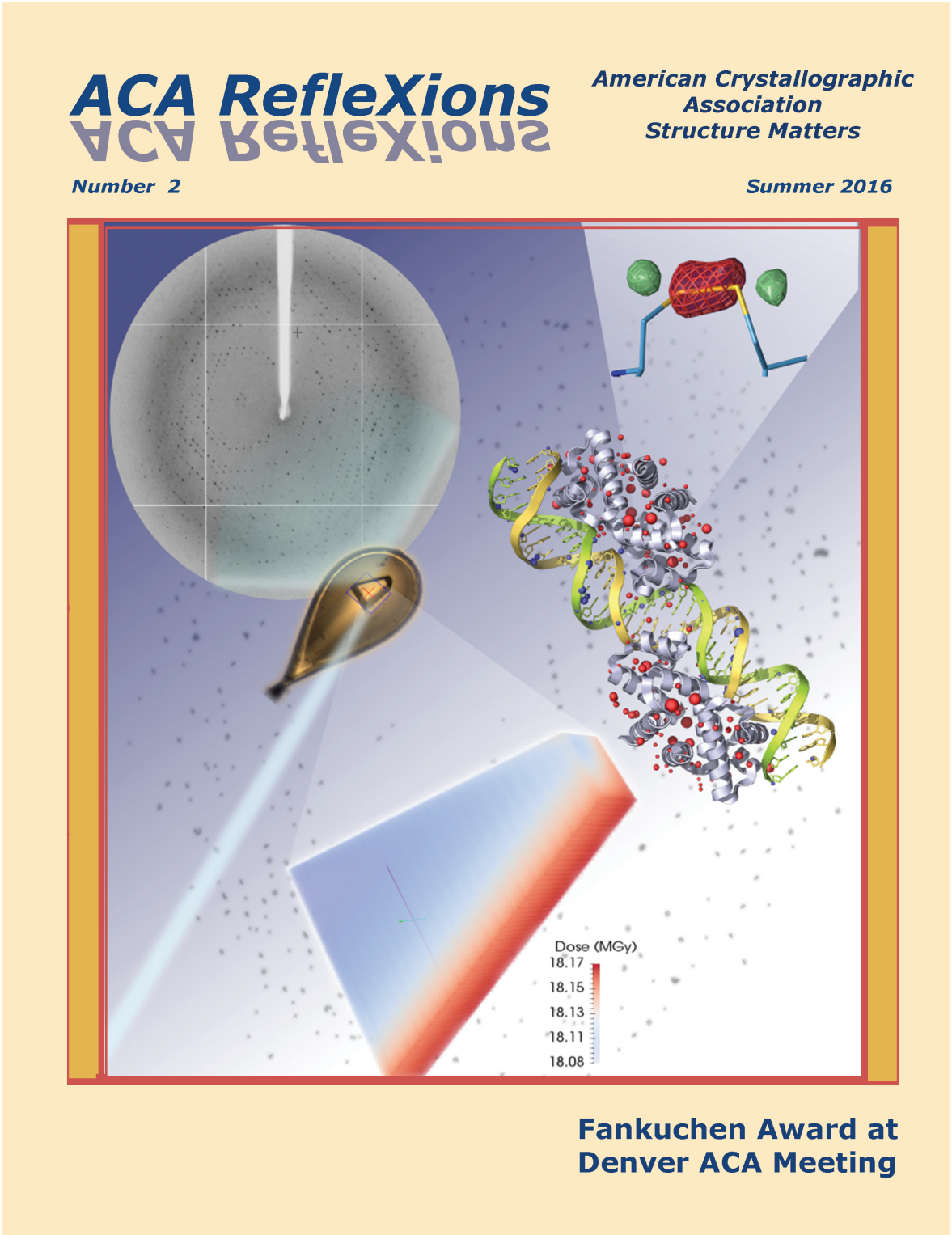


Figure C.3: Front cover image for ACA refleXions Summer 2016 edition co-created with Dr. J. C. Brooks-Bartlett.

APPENDIX D

Publications

Journal of
Synchrotron
Radiation
ISSN 1600-5775

Received 19 September 2014
Accepted 30 November 2014

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

Radiation damage to nucleoprotein complexes in macromolecular crystallography

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

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

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

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

2. Materials and methods

2.1. Crystallization and X-ray data collection

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

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

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

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

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

2.2. Dose calculation

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

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

2.3. Data processing

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

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

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

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

2.4. Automated specific damage location and large dataset collection

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

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

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

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

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

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

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

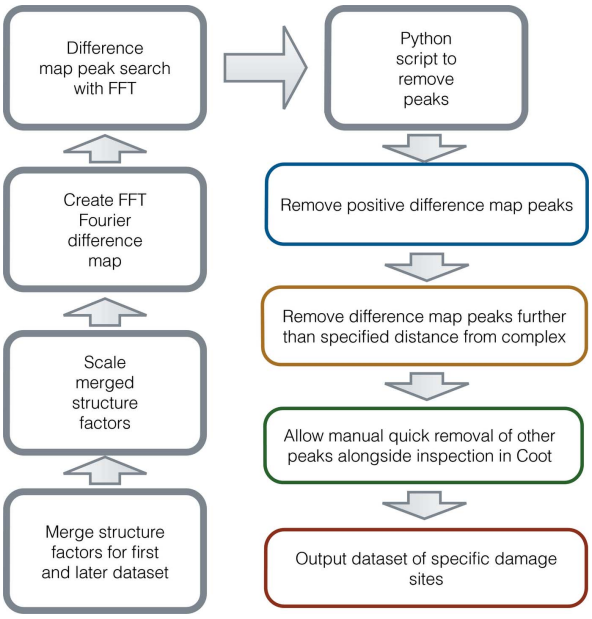


Figure 1
Flow chart for the difference map peak reduction process by systematic analysis, for a given later dataset $n \in \{2, \ldots, 8\}$. Steps performed in the *Python* scripts developed for this work are coloured blue/yellow/green/red.

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

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

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

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

3. Results

3.1. C.Esp1396I protein–DNA complex crystallography

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

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

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

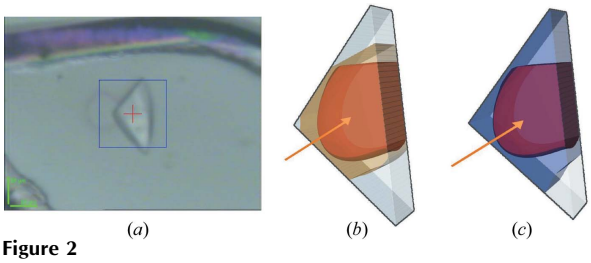


Figure 2
(*a*) The C.Esp1396I crystal within the rayon fibre loop during data collection. Crystal dimensions were x = 30 µm and y = 30 µm. The crystal is positioned in the loop such that dimension z = 10 µm is directly into the page. The blue box shown around the crystal is 50 µm × 50 µm. A faint purple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

Table 2
Data processing and refinement statistics.

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

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

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

3.2. Specific damage observations

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

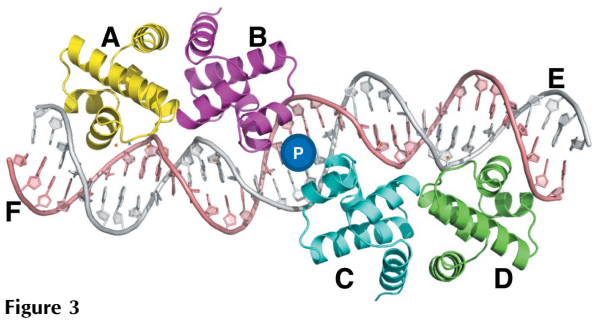


Figure 3
Visual representation of the C.Esp1396I complex rendered using *PyMOL*. Protein (*A*–*D*) and DNA (*E*, *F*) chains are labelled, and the 180° near-NCS symmetry axis is shown by marker *P*.

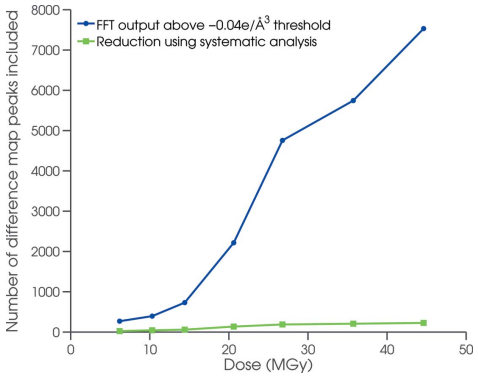


Figure 4
Reduction in the number of detected difference map peaks corresponding to potential specific damage obtained following a systematic analysis using the custom-made script for each dose (MGy).

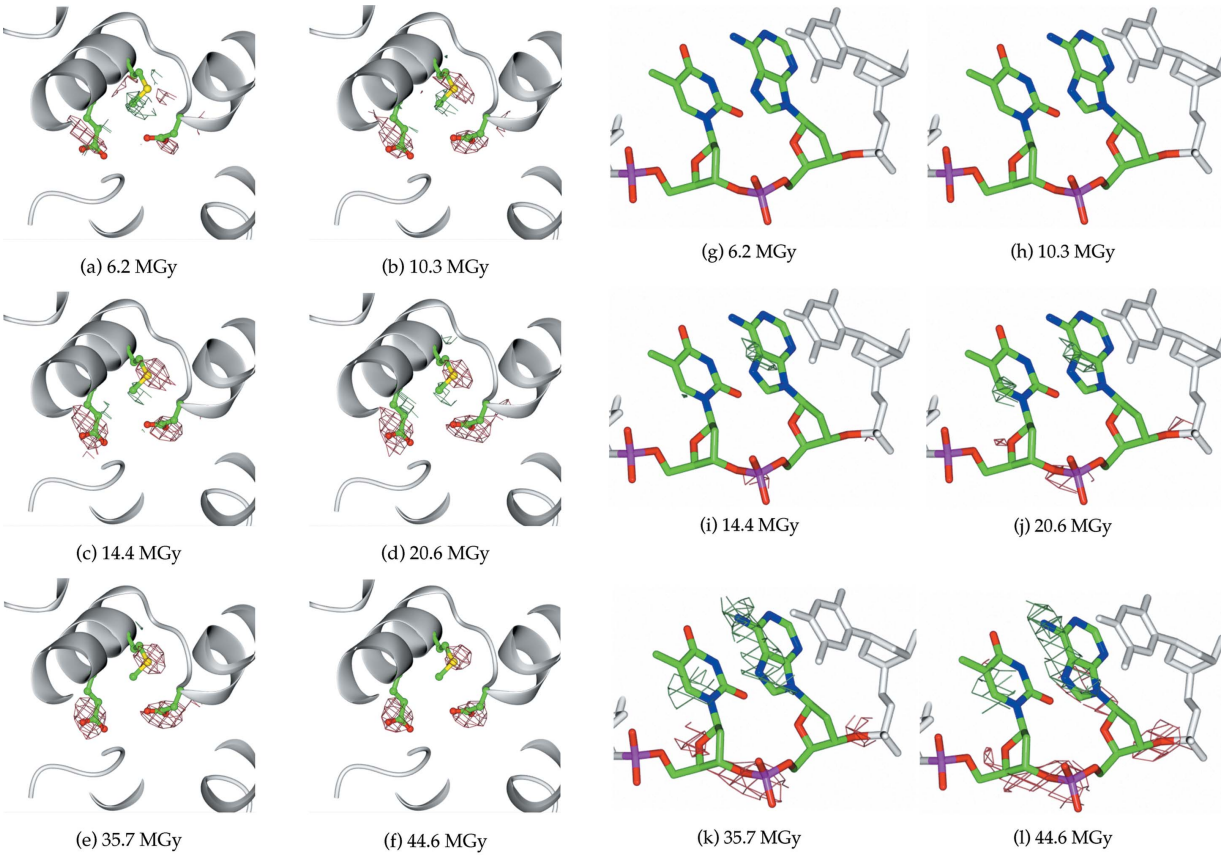


Figure 5
Protein and DNA damage sites in C.Esp1396I. (*a*)–(*f*) Visual representation of specific damage within protein chain *D* to Glu54, Met57 and Asp64 (green, left to right), displaying Fourier difference maps $F_{o,n} - F_{o,1}$, $n = 2, \ldots, 7$, over six increasing doses. Fourier maps are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

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

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

3.3. Chemical and topological distribution of specific damage

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

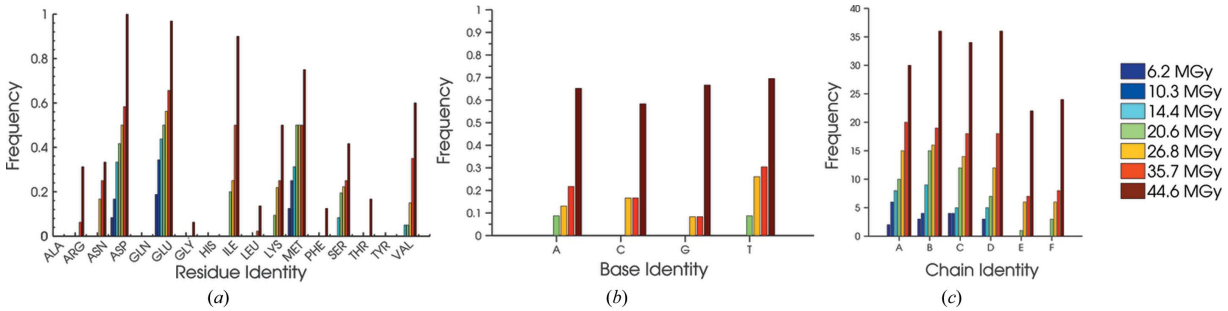
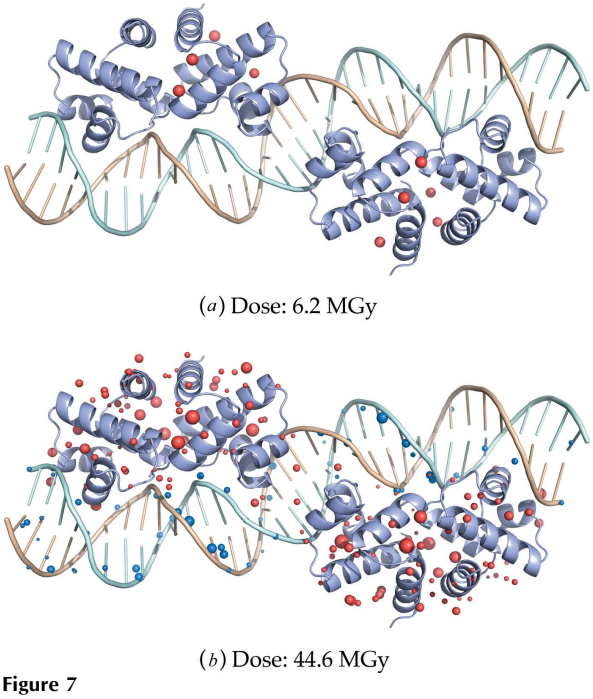


Figure 6
(*a*) Normalized frequency of detected specific damage against protein residue type (normalized to the frequency of occurrence of that residue in the structure) for each dose (MGy). Protein chains *A* to *D* are treated together. (*b*) Normalized frequency of detected specific damage for each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

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

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

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


(*a*) Dose: 6.2 MGy

(*b*) Dose: 44.6 MGy

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

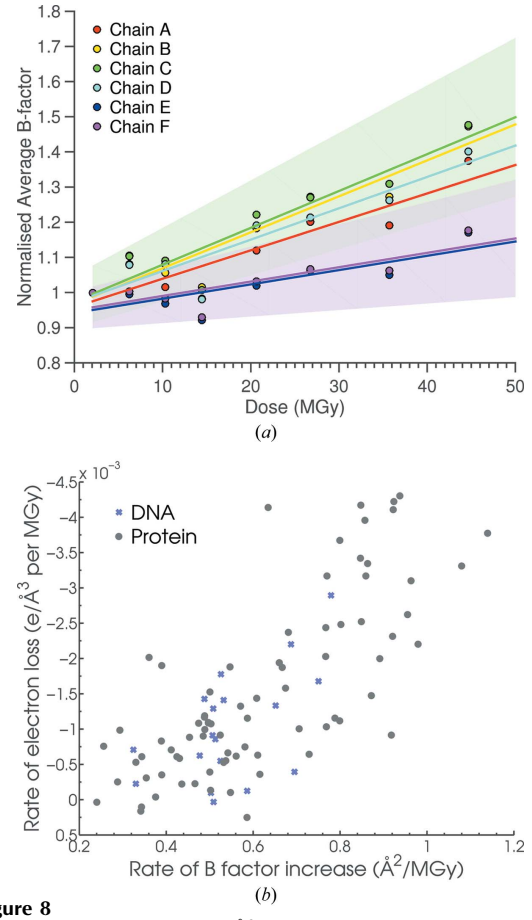


Figure 8
(*a*) Mean isotropic *B*-factor (Å^2) for each chain of the C.Esp1396I complex (*A* to *D* for protein monomers, *E* and *F* for DNA single strands) against dose (MGy), normalized to the *B*-factor of the first dataset for each chain, respectively. Data points are linearly fitted for each chain.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

and the existence of different damage mechanisms for the DNA.

3.4. Mean isotropic *B*-factor analysis

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

3.5. Specific damage dose-dynamics

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

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

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

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

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

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

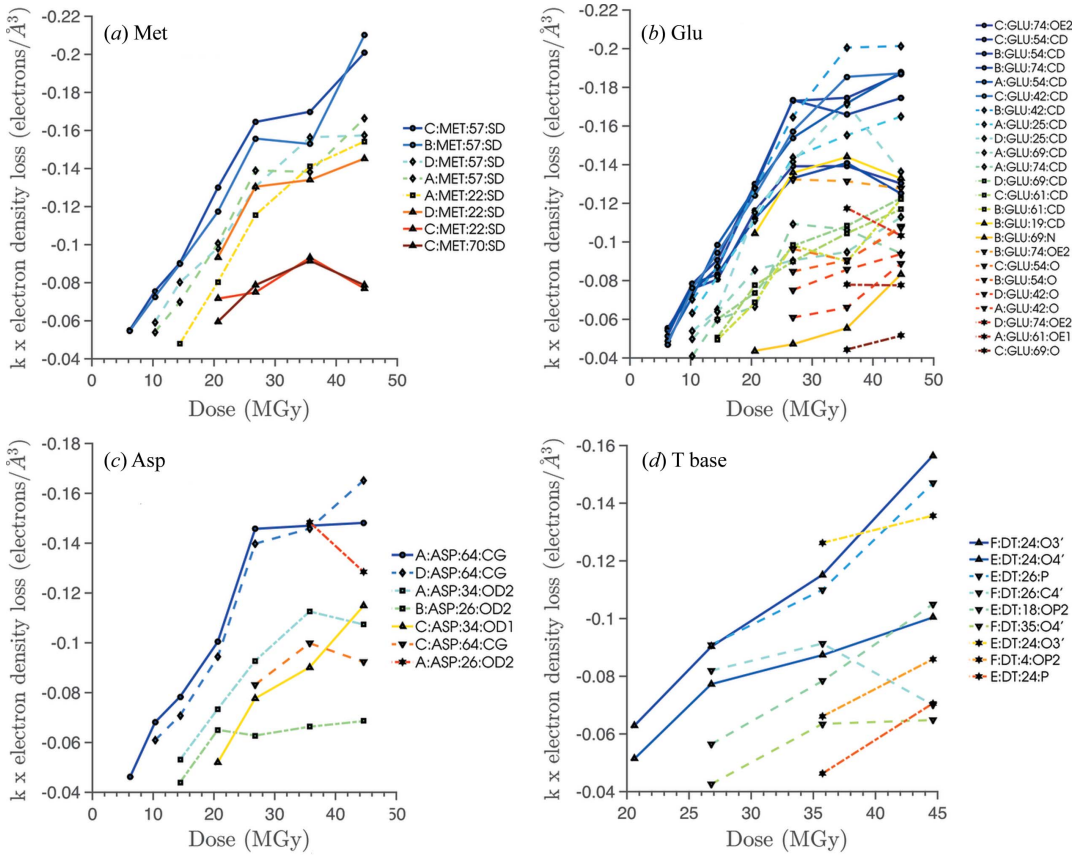


Figure 9
Electron density loss per Å^3 (scaled by constant factor k; see §2.4) against accumulated dose (MGy) for (*a*) methylthio group damage for methionine residues, decarboxylation of (*b*) glutamate and (*c*) aspartate groups, (*d*) DNA base T damage including base-sugar N_1—C bond and sugar-phosphate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

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

4. Discussion

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

STRUCTURAL BIOLOGY

ISSN 2059-7983

RNA protects a nucleoprotein complex against radiation damage

Charles S. Bury,[a] John E. McGeehan,[b] Alfred A. Antson,[c] Ian Carmichael,[d] Markus Gerstel,[a] Mikhail B. Shevtsov[e] and Elspeth F. Garman[a]*

[a]Laboratory of Molecular Biophysics, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, England, [b]Molecular Biophysics, Institute of Biomedical and Biomolecular Sciences, University of Portsmouth, King Henry I Street, Portsmouth PO1 2DY, England, [c]York Structural Biology Laboratory, Department of Chemistry, University of York, York Y010 5DD, England, [d]Notre Dame Radiation Laboratory, University of Notre Dame, Notre Dame, IN 46556, USA, and [e]Laboratory of Structural Biology of GPCRs, Moscow Institute of Physics and Technology, Dolgoprudniy 141700, Russian Federation. *Correspondence e-mail: elspeth.garman@bioch.ox.ac.uk

Received 1 February 2016
Accepted 26 February 2016

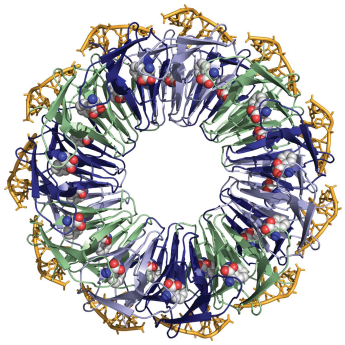
Edited by R. J. Read, University of Cambridge, England

Keywords: radiation damage; protein–RNA complex; electron difference density; specific damage; decarboxylation.

PDB references: TRAP–RNA, dose 1.3 MGy, 5eeu; dose 3.9 MGy, 5eev; dose 6.5 MGy, 5eew; dose 9.0 MGy, 5eex; dose 11.6 MGy, 5eey; dose 14.1 MGy, 5eez; dose 16.7 MGy, 5ef0; dose 19.3 MGy, 5ef1; dose 25.0 MGy, 5ef3; dose 21.9 MGy, 5ef2

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

Radiation damage during macromolecular X-ray crystallographic data collection is still the main impediment for many macromolecular structure determinations. Even when an eventual model results from the crystallographic pipeline, the manifestations of radiation-induced structural and conformation changes, the so-called specific damage, within crystalline macromolecules can lead to false interpretations of biological mechanisms. Although this has been well characterized within protein crystals, far less is known about specific damage effects within the larger class of nucleoprotein complexes. Here, a methodology has been developed whereby per-atom density changes could be quantified with increasing dose over a wide (1.3–25.0 MGy) range and at higher resolution (1.98 Å) than the previous systematic specific damage study on a protein–DNA complex. Specific damage manifestations were determined within the large *trp* RNA-binding attenuation protein (TRAP) bound to a single-stranded RNA that forms a belt around the protein. Over a large dose range, the RNA was found to be far less susceptible to radiation-induced chemical changes than the protein. The availability of two TRAP molecules in the asymmetric unit, of which only one contained bound RNA, allowed a controlled investigation into the exact role of RNA binding in protein specific damage susceptibility. The 11-fold symmetry within each TRAP ring permitted statistically significant analysis of the Glu and Asp damage patterns, with RNA binding unexpectedly being observed to protect these otherwise highly sensitive residues within the 11 RNA-binding pockets distributed around the outside of the protein molecule. Additionally, the method enabled a quantification of the reduction in radiation-induced Lys and Phe disordering upon RNA binding directly from the electron density.



OPEN ACCESS

1. Introduction

With the wide use of high-flux third-generation synchrotron sources, radiation damage (RD) has once again become a dominant reason for the failure of structure determination using macromolecular crystallography (MX) in experiments conducted both at room temperature and under cryocooled conditions (100 K). Significant progress has been made in recent years in understanding the inevitable manifestations of X-ray-induced RD within protein crystals, and there is now a body of literature on possible strategies to mitigate the effects of RD (*e.g.* Zeldin, Brockhauser *et al.*, 2013; Bourenkov & Popov, 2010). However, there is still no general consensus within the field on how to minimize RD during MX data collection, and debates on the dependence of RD progression on incident X-ray energy (Shimizu *et al.*, 2007; Liebschner *et*

al., 2015) and the efficacy of radical scavengers (Allan *et al.*, 2013) have yet to be resolved.

RD manifests in two forms. *Global radiation damage* is observed within reciprocal space as the overall decay of the summed intensity of reflections detected within the diffraction pattern as dose increases (Garman, 2010; Murray & Garman, 2002). Dose is defined as the absorbed energy per unit mass of crystal in grays (Gy; 1 Gy = 1 J kg^{-1}), and is the metric against which damage progression should be monitored during MX data collection, as opposed to time. At 100 K, an experimental dose limit of 30 MGy has been recommended as an upper limit beyond which the biological information derived from any macromolecular crystal may be compromised (Owen *et al.*, 2006).

Specific radiation damage (SRD) is observed in the real-space electron density, and has been detected at much lower doses than any observable decay in the intensity of reflections. Indeed, the C—Se bond in selenomethionine, the stability of which is key for the success of experimental phasing methods, can be cleaved at a dose as low as 2 MGy for a crystal maintained at 100 K (Holton, 2007). SRD has been well characterized in a large range of proteins, and is seen to follow a reproducible order: metallo-centre reduction, disulfide-bond cleavage, acidic residue decarboxylation and methionine methylthio cleavage (Ravelli & McSweeney, 2000; Burmeister, 2000; Weik *et al.*, 2000; Yano *et al.*, 2005). Furthermore, damage susceptibility within each residue type follows a preferential ordering influenced by a combination of local environment factors (solvent accessibility, conformational strain, proximity to active sites/high X-ray cross-section atoms; Holton, 2009). Deconvoluting the individual roles of these parameters has been surprisingly challenging, with factors such as solvent accessibility currently under active investigation (Weik *et al.*, 2000; Fioravanti *et al.*, 2007; Gerstel *et al.*, 2015).

There are a number of cases where SRD manifestations have compromised the biological information extracted from MX-determined structures at much lower doses than the recommended 30 MGy limit, leading to false structural interpretations of protein mechanisms. Active-site residues appear to be particularly susceptible, particularly for photosensitive proteins and in instances where chemical strain is an intrinsic feature of the reaction mechanism. For instance, structure determination of the purple membrane protein bacteriorhodopsin required careful corrections for radiation-induced structural changes before the correct photosensitive intermediate states could be isolated (Matsui *et al.*, 2002). The significant chemical strain required for catalysis within the active site of phosphoserine aminotransferase has been observed to diminish during X-ray exposure (Dubnovitsky *et al.*, 2005).

Since the majority of SRD studies to date have focused on proteins, much less is known about the effects of X-ray irradiation on the wider class of crystalline nucleoprotein complexes or how to correct for such radiation-induced structural changes. Understanding RD to such complexes is crucial, since DNA is rarely naked within a cell, instead dynamically interacting with proteins, facilitating replication, transcription, modification and DNA repair. As of early 2016, >5400 nucleoprotein complex structures have been deposited within the PDB, with 91% solved by MX. It is essential to understand how these increasingly complex macromolecular structures are affected by the radiation used to solve them. Nucleoproteins also represent one of the main targets of radiotherapy, and an insight into the damage mechanisms induced by X-ray irradiation could inform innovative treatments.

When a typical macromolecular crystal is irradiated with ionizing X-rays, each photoelectron produced *via* interactions with both the macromolecule (direct damage) and solvent (indirect damage) can induce cascades of up to 500 secondary low-energy electrons (LEEs) that are capable of inducing further ionizations. Investigations on sub-ionization-level LEEs (0–15 eV) interacting with both dried and aqueous oligonucleotides (Alizadeh & Sanche, 2014; Simons, 2006) concluded that resonant electron attachment to DNA bases and the sugar-phosphate backbone could lead to the preferential cleavage of strong (∼4 eV, 385 kJ mol^{-1}) sugar-phosphate C—O covalent bonds within the DNA backbone and then base-sugar N_1—C bonds, eventually leading to single-strand breakages (SSBs; Ptasińska & Sanche, 2007). Electrons have been shown to be mobile at 77 K by electron spin resonance spectroscopy studies (Symons, 1997; Jones *et al.*, 1987), with rapid electron quantum tunnelling and positive hole migration along the protein backbone and through stacked DNA bases indicated as a dominant mechanism by which oxidative and reductive damage localizes at distances from initial ionization sites at 100 K (O'Neill *et al.*, 2002).

The investigation of naturally forming nucleoprotein complexes circumvents the inherent challenges in making controlled comparisons of damage mechanisms between protein and nucleic acids crystallized separately. Recently, for a well characterized bacterial protein–DNA complex (C.Esp1396I; PDB entry 3clc; resolution 2.8 Å; McGeehan *et al.*, 2008) it was concluded that over a wide dose range (2.1–44.6 MGy) the protein was far more susceptible to SRD than the DNA within the crystal (Bury *et al.*, 2015). Only at doses above 20 MGy were precursors of phosphodiester-bond cleavage observed within AT-rich regions of the 35-mer DNA.

For crystalline complexes such as C.Esp1396I, whether the protein is intrinsically more susceptible to X-ray-induced damage or whether the protein scavenges electrons to protect the DNA remains unclear in the absence of a non-nucleic acid-bound protein control obtained under exactly the same crystallization and data-collection conditions. To monitor the effects of nucleic acid binding on protein damage susceptibility, a crystal containing two protein molecules per asymmetric unit, only one of which was bound to RNA, is reported here (Fig. 1). Using newly developed methodology, we present a controlled SRD investigation at 1.98 Å resolution using a large (∼91 kDa) crystalline protein–RNA complex: *trp* RNA-binding attenuation protein (TRAP) bound to a 53 bp RNA sequence $(GAGUU)_{10}GAG$ (PDB entry 1gtf; Hopcroft *et al.*, 2002). TRAP consists of 11 identical subunits assembled into a

ring with 11-fold rotational symmetry. It binds with high affinity ($K_d \simeq 1.0$ n*M*) to RNA segments containing 11 GAG/UAG triplets separated by two or three spacer nucleotides (Elliott *et al.*, 2001) to regulate the transcription of tryptophan biosynthetic genes in *Bacillus subtilis* (Antson *et al.*, 1999). In this structure, the bases of the G1-A2-G3 nucleotides form direct hydrogen bonds to the protein, unlike the U4-U5 nucleotides, which appear to be more flexible.

Ten successive 1.98 Å resolution MX data sets were collected from the same TRAP–RNA crystal to analyse X-ray-induced structural changes over a large dose range (d_1 = 1.3 MGy to d_{10} = 25.0 MGy). To avoid the previous necessity for visual inspection of electron-density maps to detect SRD sites, a computational approach was designed to quantify the electron-density change for each refined atom with increasing dose, thus providing a rapid systematic method for SRD study on such large multimeric complexes. By employing the high 11-fold structural symmetry within each TRAP macromolecule, this approach permitted a thorough statistical quantification of the RD effects of RNA binding to TRAP.

2. Materials and methods

2.1. RNA synthesis and protein preparation

As previously described (Hopcroft *et al.*, 2002), the 53-base RNA $(GAGUU)_{10}GAG$ was synthesized by *in vitro* transcription with T7 RNA polymerase and gel-purified. TRAP from *B. stearothermophilus* was overexpressed in *Escherichia coli* and purified.

2.2. Crystallization

TRAP–RNA crystals were prepared using a previously established hanging-drop crystallization protocol (Antson *et al.*, 1999). By using a 2:1 molar ratio of TRAP to RNA, crystals successfully formed from the protein–RNA complex (∼15 mg ml^{-1}) in a solution containing 70 m*M* potassium phosphate pH 7.8 and 10 m*M* L-tryptophan. The reservoir consisted of 0.2 *M* potassium glutamate, 50 m*M* triethanolamine pH 8.0, 10 m*M* $MgCl_2$, 8–11% monomethyl ether PEG 2000. In order to accelerate crystallization, a further gradient was induced by adding 0.4 *M* KCl to the reservoir after 1.5 µl protein solution had been mixed with an equal volume of the reservoir solution. Wedge-shaped crystals of approximate length 70 µm (longest dimension) grew within 3 d and were vitrified and stored in liquid nitrogen immediately after growth. The cryosolution consisted of 12% monomethyl ether PEG 2000, 30 m*M* triethanolamine pH 8.0, 6 m*M* L-tryptophan, 0.1 *M* potassium glutamate, 35 m*M* potassium phosphate pH 7.8, 5 m*M* $MgCl_2$ with 25% 2-methyl-2,4-pentanediol (MPD) included as a cryoprotectant.

2.3. X-ray data collection

Data were collected at 100 K from a wedge-shaped TRAP–RNA crystal of approximate dimensions 70 × 20 × 40 µm (see Supplementary Fig. S2) on beamline ID14-4 at the ESRF using an incident wavelength of 0.940 Å (13.2 keV) and an ADSC Q315R mosaic CCD detector at 304.5 mm from the crystal throughout the data collection. The beam size was

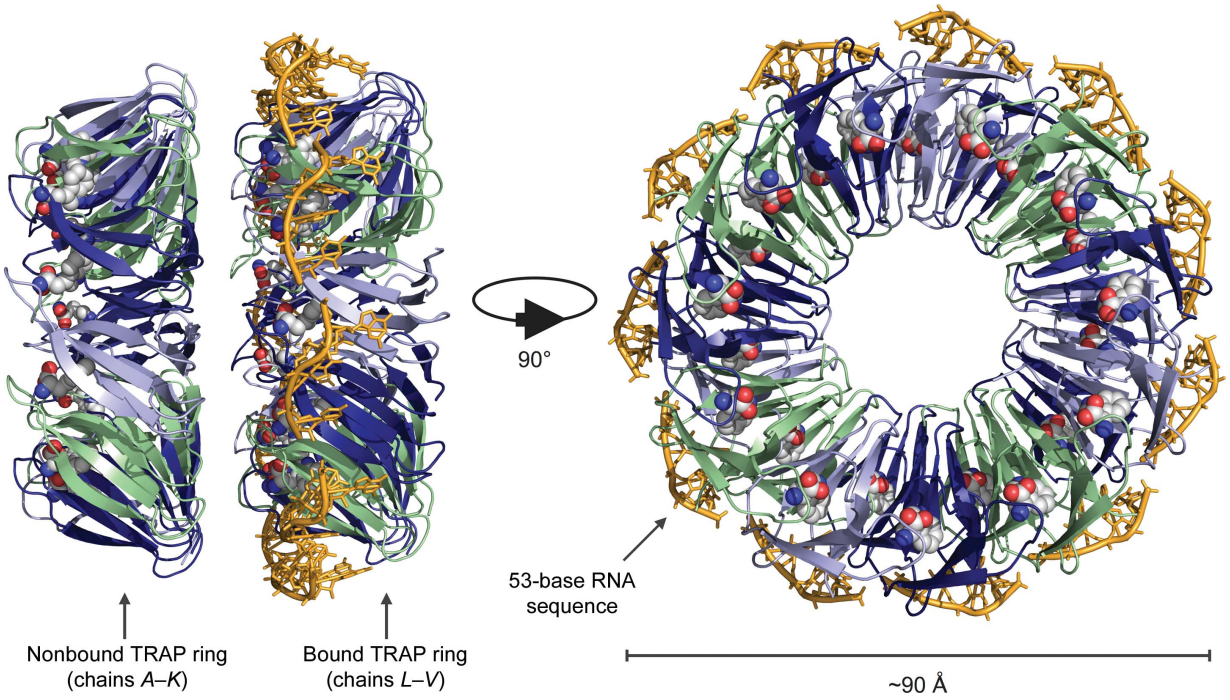


Figure 1
The TRAP–$(GAGUU)_{10}GAG$ complex asymmetric unit (PDB entry 1gtf; Hopcroft *et al.*, 2002). Bound tryptophan ligands are represented as coloured spheres. RNA is shown is yellow.

slitted to 0.100 mm (vertical) × 0.160 mm (horizontal), with a uniformly distributed profile, such that the crystal was completely bathed within the beam throughout data collection. Ten successive (1.98 Å resolution) 180° data sets (with $\Delta\varphi = 1^\circ$) were collected over the same angular range from a TRAP–RNA crystal at 28.9% beam transmission. The TRAP–RNA macromolecule crystallized in space group *C*2, with unit-cell parameters $a = 140.9$, $b = 110.9$, $c = 137.8$ Å, $\alpha = \gamma = 90$, $\beta = 137.8^\circ$ (the values quoted are for the first data set; see Supplementary Table S1 for subsequent values). For the first nine data sets the attenuated flux was recorded to be $\sim 5 \times 10^{11}$ photons s^{-1}. A beam refill took place immediately before data set 10, requiring a flux-scale factor increase of 1.42 to be applied, based on the ratio of observed relative intensity I_D/I_1 at data set 10 to that extrapolated from data set 9.

2.4. Dose calculation

RADDOSE-3D (Zeldin, Gerstel *et al.*, 2013) was used to calculate the absorbed dose distribution during each data set (see input file; Supplementary Figs. S1 and S2). The crystal composition was calculated from the deposited TRAP–RNA structure (PDB entry 1gtf; Hopcroft *et al.*, 2002). Crystal absorption coefficients were calculated in *RADDOSE-3D* using the concentration (mmol l^{-1}) of solvent heavy elements from the crystallization conditions. The beam-intensity profile was modelled as a uniform ('top-hat') distribution. The diffraction-weighted dose (DWD) values (Zeldin, Brockhauser *et al.*, 2013) are given in Supplementary Table S1.

2.5. Data processing and model refinement

Each data set was integrated using *iMosflm* (Leslie & Powell, 2007) and was scaled using *AIMLESS* (Evans & Murshudov, 2013; Winn *et al.*, 2011) using the same 5% $R_{\rm free}$ set of test reflections for each data set. To phase the structure obtained from the first data set, molecular replacement was carried out with *Phaser* (McCoy *et al.*, 2007), using an identical TRAP–RNA structure (PDB entry 1gtf; resolution 1.75 Å; Hopcroft *et al.*, 2002) as a search model. The resulting TRAP–RNA structure (TR1) was refined using *REFMAC*5 (Murshudov *et al.*, 2011), initially using rigid-body refinement, followed by repeated cycles of restrained, TLS and isotropic *B*-factor refinement, coupled with visual inspection in *Coot* (Emsley *et al.*, 2010). TR1 was refined to 1.98 Å resolution, with a dimeric assembly of non-RNA-bound and RNA-bound TRAP rings within the asymmetric unit. Consistent with previous structures of the TRAP–RNA complex, the RNA sequence termini were not observed within the $2F_o - F_c$ map; the first spacer (U4) was then modelled at all 11 repeats around the TRAP ring and the second spacer (U5) was omitted from the final refined structure. For the later data sets, the observed structure-factor amplitudes from each separately scaled data set (output from *AIMLESS*) were combined with the phases of TR1 and the resulting higher-dose model was refined with *phenix.refine* (Adams *et al.*, 2010) using only rigid-body and isotropic *B*-factor refinement. During this refinement, the TRAP–RNA complex and nonbound TRAP ring were treated as two separate rigid bodies within the asymmetric unit. Supplementary Table S1 shows the relevant summary statistics.

2.6. $D_{\rm loss}$ metric calculation

The *CCP*4 program *CAD* was used to create a series of nine merged .mtz files combining observed structure-factor amplitudes for the first data set $F_{\rm obs}(d_1)$ with each later data set $F_{\rm obs}(d_n)$ (for $n = 2, \ldots, 10$). All later data sets were scaled against the initial low-dose data set in *SCALEIT*. For each data set an atom-tagged .map file was generated using the ATMMAP mode in *SFALL* (Winn *et al.*, 2011). A full set of nine Fourier difference maps $F_{\rm obs}(d_n) - F_{\rm obs}(d_1)$ were calculated using *FFT* (Ten Eyck, 1973) over the full TRAP–RNA unit-cell dimensions, with the same grid-sampling dimensions as the atom-tagged .map file. All maps were cropped to the TRAP asymmetric unit in *MAPMASK*. Comparing the atom-tagged .map file and $F_{\rm obs}(d_n) - F_{\rm obs}(d_1)$ difference map at each dose, each refined atom was assigned a set of density-change values *X*. The maximum density-loss metric, $D_{\rm loss}$ (units of e $Å^{-3}$), was calculated to quantify the per-atom electron-density decay at each dose, assigned as the absolute magnitude of the most negative Fourier difference map voxel value in a local volume around each atom as defined by the set X.

2.7. Model system calculation

Model calculations were run for the simple amino acids glutamate and aspartate. In order to avoid decarboxylation at the C-terminus instead of the side chain on the C^α atom, the C-terminus of each amino acid was methylated. While the structures of the closed shell acids are well known, the same is not true of those in the oxidized state. The quantum-chemical calculations employed were chosen to provide a satisfactory description of the structure of such radical species and also provide a reliable estimation of the relative $C-C(O_2)$ bond strengths, which are otherwise not available.

Structures of methyl-terminated (at the N- and C-termini) carboxylates were determined using analytic energy gradients with density functional theory (B3LYP functional; Becke, 1993) and a flexible basis set of polarized valence triple-zeta size with diffuse functions on the non-H atoms [6-311+G(d,p)] in the *Gaussian* 09 computational chemistry package (Frisch *et al.*, 2009). The stationary points obtained were characterized as at least local minima by examination of the associated analytic Hessian. Effects of the medium were modelled using a dielectric cavity approach (Tomasi *et al.*, 1999) parameterized for water.

3. Results

3.1. Per-atom quantification of electron density

To quantify the exact effects of nucleic acid binding to a protein on SRD susceptibility, a high-throughput and automated pipeline was created to systematically calculate the electron-density change for every refined atom within the TRAP–RNA structure as a function of dose. This provides an atom-specific quantification of density–dose dynamics, which

was previously lacking within the field. Previous studies have characterized SRD sites by reporting magnitudes of $F_{obs}(d_n) - F_{obs}(d_1)$ Fourier difference map peaks in terms of the sigma (σ) contour level (the number of standard deviations from the mean map electron-density value) at which peaks become visible. However, these σ levels depend on the standard deviation values of the map, which can deviate between data

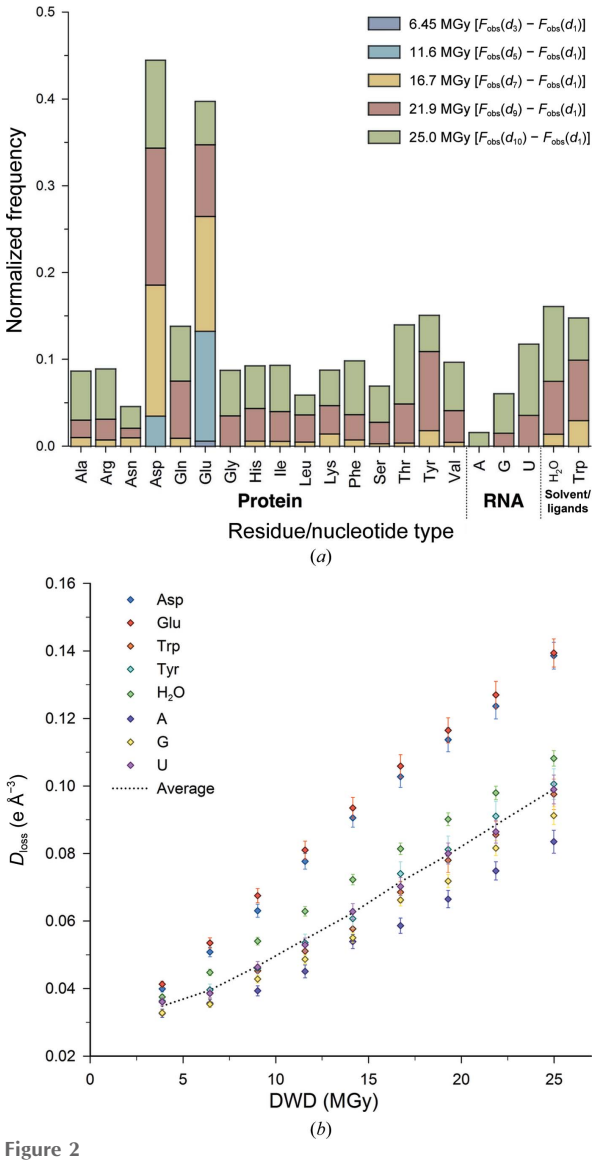


Figure 2
(*a*) Electron-density loss sites as indicated by D_{loss} in the TRAP–RNA complex crystal by residue/nucleotide type for five doses [sites determined above the 4× average D_{loss} threshold, calculated over the TRAP–RNA structure for the first difference map: $F_{obs}(d_2) - F_{obs}(d_1)$]. Cumulative frequencies are normalized to both the total number of non-H atoms per residue/nucleotide and the total number of that residue/nucleotide type present. (*b*) Average D_{loss} for each residue/nucleotide type with respect to the DWD (diffraction-weighted dose; Zeldin, Brockhauser *et al.*, 2013). 95% confidence intervals (CI) are shown. Only a subset of key TRAP residue types are included. The average D_{loss} (calculated over the whole TRAP asymmetric unit) is shown at each dose (dashed line).

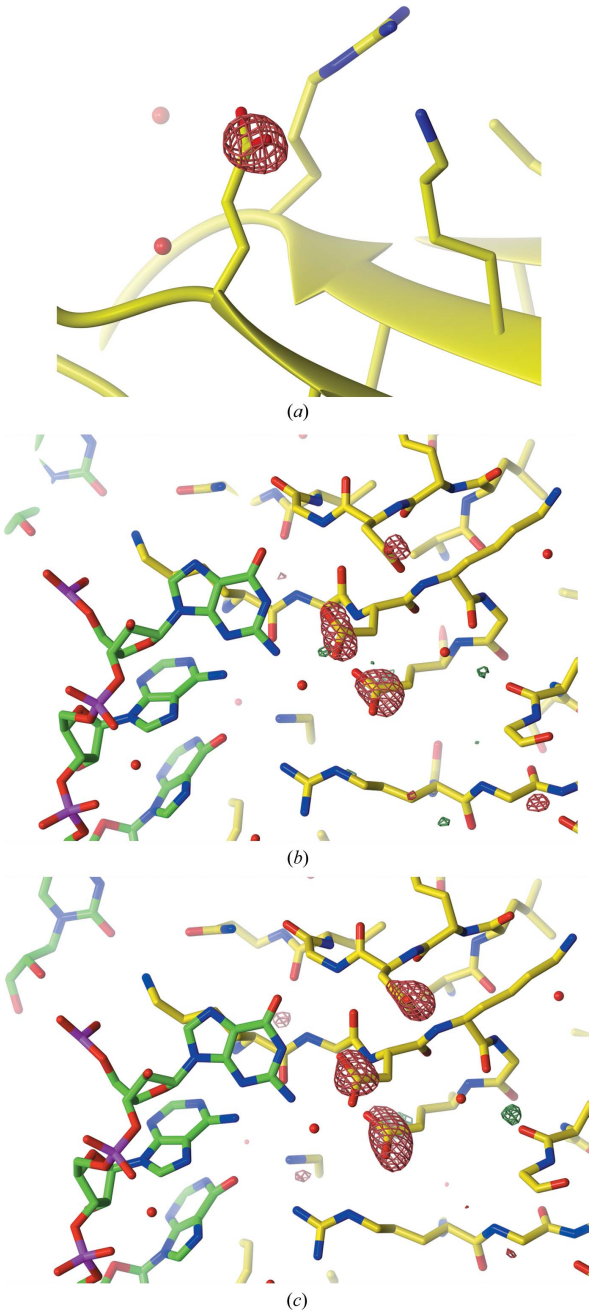


Figure 3
$F_{obs}(d_n) - F_{obs}(d_1)$ Fourier difference maps for (*a*) n = 2 (3.9 MGy), (*b*) n = 3 (6.5 MGy) and (*c*) n = 7 (16.7 MGy) contoured at $\pm 4\sigma$ (*a*) and $\pm 3.5\sigma$ (*b*, *c*). In (*a*) clear difference density is observed around the Glu42 carboxyl side chain in chain *H*, within the lowest dose difference map at d_2 = 3.9 MGy. Radiation-induced protein disordering is evident across the large dose range (*b*, *c*); in comparison, no clear deterioration of the RNA density was observed.

sets, and are thus unsuitable for quantitative comparison of density between different dose data sets. Instead, we use here a maximum density-loss metric ($D_{\rm loss}$), which is the per-atom equivalent of the magnitude of these negative Fourier difference map peaks in units of e Å^{-3}. Large positive $D_{\rm loss}$ values indicate radiation-induced atomic disordering reproducibly throughout the unit cells with respect to the initial low-dose data set.

For each TRAP–RNA data set, the $D_{\rm loss}$ metric successfully identified the recognized forms of protein SRD (Fig. 2*a*), with clear Glu and Asp side-chain decarboxylation even in the first difference map calculated (3.9 MGy; Fig. 3*a*). The main sequence of TRAP does not contain any Trp and Cys residues (and thus contains no disulfide bonds). The substrate Trp amino-acid ligands also exhibited disordering of the free terminal carboxyl groups at higher doses (Fig. 2*a*); however, no clear Fourier difference peaks could be observed visually. Even for radiation-insensitive residues (*e.g.* Gly) the average $D_{\rm loss}$ increases with dose: this is the effect of global radiation damage, since as dose increases the electron density associated with each refined atom becomes weaker as the atomic occupancy decreases (Fig. 2*b*). Only Glu and Asp residues exhibit a rate of $D_{\rm loss}$ increase that consistently exceeds the average decay (Fig. 2*b*, dashed line) at each dose. Additionally, the density surrounding ordered solvent molecules was determined to significantly diminish with increasing dose (Fig. 2*b*). The rate of $D_{\rm loss}$ (attributed to side-chain decarboxylation) was consistently larger for Glu compared with Asp residues over the large dose range (Fig. 2*b* and Supplementary Fig. S3); this observation is consistent with our calculations on model systems (see above) that suggest that, without considering differential hydrogen-bonding environments, CO_2 loss is more exothermic by around 8 kJ mol^{-1} from oxidized Glu residues than from their Asp counterparts.

3.2. RNA is less susceptible to electron-density loss than protein within the TRAP–RNA complex

Visual inspection of Fourier difference maps illustrated the clear lack of RNA electron-density degradation with increasing dose compared with the obvious protein damage manifestations (Figs. 3*b* and 3*c*). Only at the highest doses investigated (>20 MGy) was density loss observed at the RNA phosphate and C—O bonds of the phosphodiester backbone. However, the median $D_{\rm loss}$ was lower by a factor of >2 for RNA P atoms than for Glu and Asp side-chain groups at 25.0 MGy (Supplementary Fig. S4), and furthermore could not be numerically distinguished from Gly C$^{\alpha}$ atoms within TRAP, which are not radiation-sensitive at the doses tested here (Supplementary Fig. S3).

3.3. RNA binding protects radiation-sensitive residues

For the large number of acidic residues per TRAP ring (four Asp and six Glu residues per protein monomer), a strong dependence of decarboxylation susceptibility on local environment was observed (Fig. 4). For each Glu C$^{\delta}$ or Asp C$^{\gamma}$ atom, $D_{\rm loss}$ provided a direct measure of the rate of side-chain carboxyl-group disordering and subsequent decarboxylation. For acidic residues with no differing interactions between nonbound and bound TRAP (Fig. 4*a*), similar damage was apparent between the two rings within the asymmetric unit, as expected. However, TRAP residues directly on the RNA-binding interfaces exhibited greater damage accumulation in nonbound TRAP (Fig. 4*b*), and for residues at the ring–ring interfaces (where crystal contacts were detected) bound TRAP exhibited enhanced SRD accumulation (Fig. 4*c*).

Three acidic residues (Glu36, Asp39 and Glu42) are involved in RNA interactions within each of the 11 TRAP ring subunits, and Fig. 5 shows their density changes with increasing dose. Hotelling's T-squared test (the multivariate counterpart of Student's t-test) was used to reject the null hypothesis that the means of the $D_{\rm loss}$ metric were equal for the bound and nonbound groups in Fig. 5.

A significant reduction in $D_{\rm loss}$ is seen for Glu36 in RNA-bound compared with nonbound TRAP, indicative of a lower rate of side-chain decarboxylation (Fig. 5*a*; $p = 6.06 \times 10^{-5}$). For each TRAP ring subunit, the Glu36 side-chain carboxyl group accepts a pair of hydrogen bonds from the two N atoms of the G3 RNA base. In our analysis, Asp39 in the TRAP–(GAGUU)$_{10}$GAG structure appears to exhibit two distinct hydrogen bonds to the G1 base within each of the 11 TRAP–RNA interfaces, as does Glu36 to G3; however, the reduction in density disordering upon RNA binding is far less significant for Asp39 than for Glu36 (Fig. 5*b*, $p = 0.0925$).

3.4. RNA binding reduces radiation-induced disorder on the atomic scale

One oxygen (O$^{\varepsilon 1}$) of Glu42 appears to form a hydrogen bond to a nearby water within each TRAP RNA-binding pocket, with the other (O$^{\varepsilon 2}$) being involved in a salt-bridge interaction with Arg58 (Hopcroft *et al.*, 2002; Antson *et al.*,

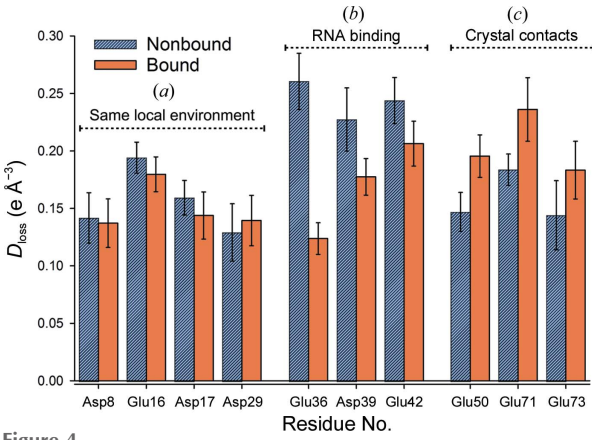


Figure 4
$D_{\rm loss}$ calculated for all side-chain carboxyl group Glu C$^{\delta}$ and Asp C$^{\gamma}$ atoms within the TRAP–RNA complex for a dose of 19.3 MGy (d_8). Residues have been grouped by amino-acid number, and split into bound and nonbound groupings, with each bar representing the mean calculated over 11 equivalent atoms around a TRAP ring. Whiskers indicate 95% CI. The $D_{\rm loss}$ behaviour shown here was consistently exhibited across the entire investigated dose range.

1999). Salt-bridge interactions have previously been suggested to reduce the glutamate decarboxylation rate within the large (∼62.4 kDa) myrosinase protein structure (Burmeister, 2000). A significant difference was observed between the $D_{\rm loss}$ dynamics for the nonbound/bound Glu42 $O^{\varepsilon1}$ atoms (Fig. 5*c*; $p = 0.007$) but not for the Glu42 $O^{\varepsilon2}$ atoms (Fig. 5*d*; $p = 0.239$), indicating that the stabilizing strength of this salt-bridge interaction was conserved upon RNA binding and that the water-mediated hydrogen bond had a greater relative susceptibility to atomic disordering in the absence of RNA. The density-change dynamics were statistically indistinguishable between bound and nonbound TRAP for each Glu42 carboxyl group C^{δ} atom ($p = 0.435$), indicating that upon RNA binding the conserved salt-bridge interaction ultimately dictated the overall Glu42 decarboxylation rate.

The RNA-stabilizing effect was not restricted to radiation-sensitive acidic residues. The side chain of Phe32 stacks against the G3 base within the 11 TRAP RNA-binding interfaces (Antson *et al.*, 1999). With increasing dose, the $D_{\rm loss}$ associated with the Phe32 side chain was significantly reduced upon RNA binding (Fig. 5*e*; Phe32 C^{ζ}; $p = 0.0014$), an indication that radiation-induced conformation disordering of Phe32 had been reduced. The extended aliphatic Lys37 side chain stacks against the nearby G1 base, making a series of nonpolar contacts within each RNA-binding interface. The $D_{\rm loss}$ for Lys37 side-chain atoms was also reduced when stacked against the G1 base (Fig. 5*f*; $p = 0.0243$ for Lys37 C^{ε} atoms). Representative Phe32 and Lys37 atoms were selected to illustrate these trends.

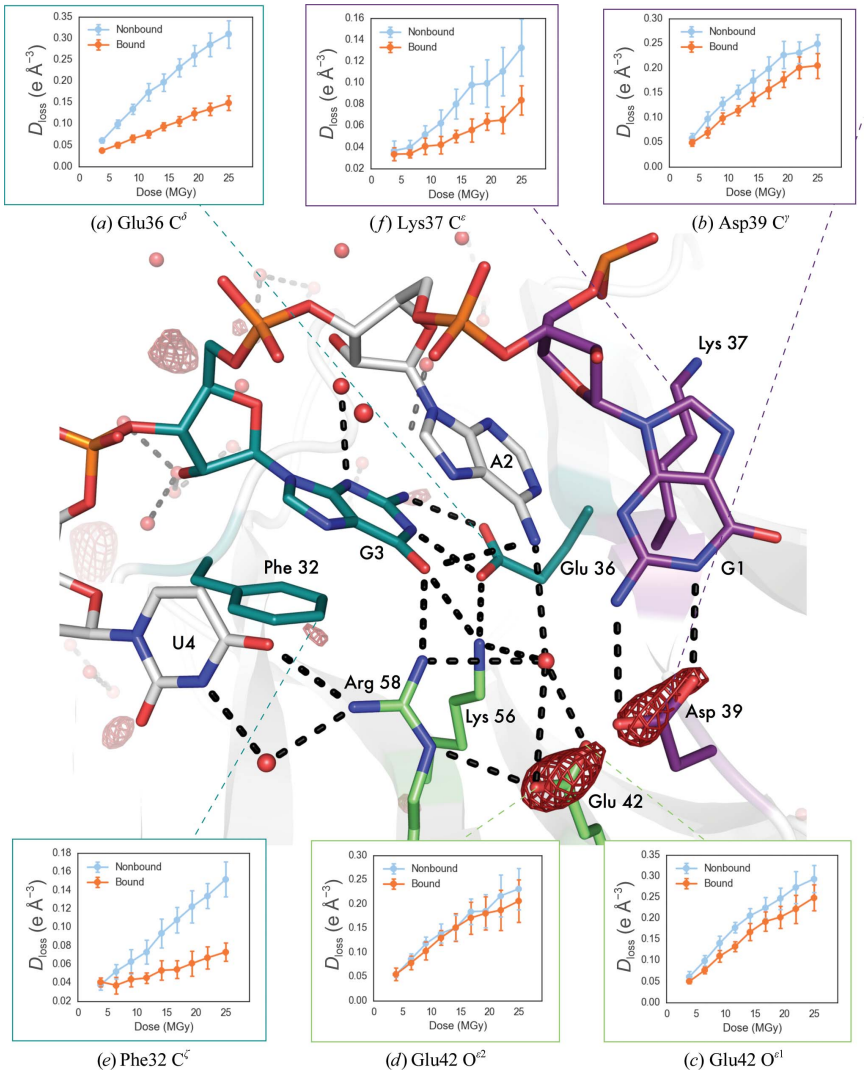


Figure 5
$D_{\rm loss}$ against dose for (*a*) Glu36 C^{δ}, (*b*) Asp39 C^{γ}, (*c*) Glu42 $O^{\varepsilon1}$, (*d*) Glu42 $O^{\varepsilon2}$, (*e*) Phe32 C^{ζ} and (*f*) Lys37 C^{ε} atoms. 95% CI are included for each set of 11 equivalent atoms grouped as bound/nonbound. RNA-binding interface interactions are shown for TRAP chain *N*, with the $F_{\rm obs}(d_7) - F_{\rm obs}(d_1)$ Fourier difference map (dose 16.7 MGy) overlaid and contoured at a $\pm4\sigma$ level.

4. Discussion

Here, MX radiation-induced specific structural changes within the large TRAP–RNA assembly over a large dose range (1.3–25.0 MGy) have been analysed using a high-throughput quantitative approach, providing a measure of the electron-density distribution for each refined atom with increasing dose, $D_{\rm loss}$. Compared with previous studies, the results provide a further step in the detailed characterization of SRD effects in MX. Our methodology, which eliminated tedious and error-prone visual inspection, permitted the determination on a per-atom basis of the most damaged sites, as characterized by $F_{\rm obs}(d_n) - F_{\rm obs}(d_1)$ Fourier difference map peaks between successive data sets collected from the same crystal. Here, it provided the precision required to quantify the role of RNA in the damage susceptibilities of equivalent atoms between RNA-bound and nonbound TRAP, but it is applicable to any MX SRD study.

The RNA was found to be substantially more radiation-resistant than the protein, even at the highest doses investigated (∼25.0 MGy), which is in strong concurrence with our previous

SRD investigation of the C.Esp1396I protein–DNA complex (Bury *et al.*, 2015). Consistent with that study, at high doses of above ∼20 MGy, $F_{obs}(d_n) - F_{obs}(d_1)$ map density was detected around P, O3′ and O5′ atoms of the RNA backbone, with no significant difference density localized to RNA ribose and basic subunits. RNA backbone disordering thus appears to be the main radiation-induced effect in RNA, with the protein–base interactions maintained even at high doses (>20 MGy). The U4 phosphate exhibited marginally larger D_{loss} values above 20 MGy than G1, A2 and G3 (Supplementary Fig. S4). Since U4 is the only refined nucleotide not to exhibit significant base–protein interactions around TRAP (with a water-mediated hydrogen bond detected in only three of the 11 subunits and a single Arg58 hydrogen bond suggested in a further four subunits), this increased U4 D_{loss} can be explained owing to its greater flexibility. At 25.0 MGy, the magnitude of the RNA backbone D_{loss} was of the same order as for the radiation-insensitive Gly C^{α} atoms and on average less than half that of the acidic residues of the protein (Supplementary Fig. S3). Consequently, no clear single-strand breaks could be located, and since RNA-binding within the current TRAP–$(GAGUU)_{10}GAG$ complex is mediated predominantly through base–protein interactions, the biological integrity of the RNA complex was dictated by the rate at which protein decarboxylation occurred.

RNA interacting with TRAP was shown to offer significant protection against radiation-induced structural changes. Both Glu36 and Asp39 bind directly to RNA, each through two hydrogen bonds to guanine bases (G3 and G1, respectively). However, compared with Asp39, Glu36 is strikingly less decarboxylated when bound to RNA (Fig. 4). This is in good agreement with previous mutagenesis and nucleoside analogue studies (Elliott *et al.*, 2001), which indicated that the G1 nucleotide does not bind to TRAP as strongly as do A2 and G3, and plays little role in the high RNA-binding affinity of TRAP ($K_d \simeq 1.1 \pm 0.4$ n*M*). For Glu36 and Asp39, no direct quantitative correlation could be established between hydrogen-bond length and D_{loss} (linear R^2 of <0.23 for all doses; Supplementary Fig. S5). Thus, another factor must be responsible for this clear reduction in Glu36 CO_2 decarboxylation in RNA-bound TRAP. The Glu36 carboxyl side chain also potentially forms hydrogen bonds to His34 and Lys56, but since these interactions are conserved irrespective of G3 nucleotide binding, this cannot directly account for the stabilization effect on Glu36 in RNA-bound TRAP. Radiation-induced decarboxylation has been proposed to be mediated by preferential positive-hole migration to the side-chain carboxyl group, with rapid proton transfer trapping the hole at the carboxyl group (Burmeister, 2000; Symons, 1997):

$$(p)\text{-}CH_2CH_2COO^- \underset{K_{-1}}{\overset{K_1}{\rightleftharpoons}} (p)\text{-}CH_2CH_2COO^{\bullet} + e^-, \qquad (1)$$

where the forward rate is K_1 and the backward rate is K_{-1},

$$(p)\text{-}CH_2CH_2COO^{\bullet} \overset{K_2}{\rightarrow} (p)\text{-}CH_2CH_2^{\bullet} + CO_2, \qquad (2)$$

where the forward rate is K_2.

When bound to RNA, the average solvent-accessible area of the Glu36 side-chain O atoms is reduced from ∼15 to 0 Å^2. We propose that with no solvent accessibility Glu36 decarboxylation is inhibited, since the CO_2-formation rate K_2 is greatly reduced, and suggest that steric hindrance prevents each radicalized Glu36 CO_2 group from achieving the planar conformation required for complete dissociation from TRAP. The electron-recombination rate K_{-1} remains high, however, owing to rapid electron migration through the protein–RNA complex to refill the Glu36 positive hole (the precursor for Glu decarboxylation). Upon RNA binding, the Asp39 side-chain carboxyl group solvent-accessible area changes from ∼75 to 35 Å^2, still allowing a high CO_2-formation rate K_2.

Previous studies have reported inconsistent results concerning the dependence of the acidic residue decarboxylation rate on solvent accessibility (Weik *et al.*, 2000; Fioravanti *et al.*, 2007; Gerstel *et al.*, 2015). The prevalence of radical attack from solvent channels surrounding the protein in the crystal is a questionable cause, considering previous observations indicating that the strongly oxidizing hydroxyl radical is immobile at 100 K (Allan *et al.*, 2013; Owen *et al.*, 2012). Furthermore, the suggested electron hole-trapping mechanism which induces decarboxylation within proteins at 100 K has no clear mechanistic dependence on the solvent-accessible area of each carboxyl group. By comparing equivalent acidic residues with and without RNA, we have now deconvoluted the role of solvent accessibility from other local protein environment factors, and thus propose a suitable mechanism by which exceptionally low solvent accessibility can reduce the rate of decarboxylation. Overall, no direct correlation between solvent accessibility and decarboxylation susceptibility was observed, but it is very clear that inaccessible residues are protected.

Apart from these RNA-binding interfaces, RNA binding was seen to enhance decarboxylation for residues Glu50, Glu71 and Glu73, all of which are involved in crystal contacts between TRAP rings (Fig. 4*c*). However, for each of these residues the exact crystal contacts are not preserved between bound and nonbound TRAP or even between monomers within one TRAP ring. For example, in bound TRAP, Glu73 hydrogen-bonds to a nearby lysine on each of the 11 subunits, whereas in nonbound TRAP no such interaction exists and Glu73 interacts with a variable number of refined waters in each subunit. Thus, the dependence of decarboxylation rates on these interactions could not be established.

Radiation-induced side-chain conformational changes have been poorly characterized in previous SRD investigations owing to their strong dependence on packing density and geometric strain. Such structural changes are known to have significant roles within enzymatic pathways, and experimenters must be aware of these possible confounding factors when assigning true functional mechanisms using MX. Our results show that RNA binding to TRAP physically stabilizes non-acidic residues within the TRAP macromolecule, most notably Lys37 and Phe32, which stack against the G1 and G3 bases, respectively. It has been suggested (Burmeister, 2000) that Tyr residues can lose their aromatic –OH group owing to

radiation-induced effects; however, no energetically favourable pathway for –OH cleavage exists and this has not been detected in aqueous radiation-chemistry studies. In TRAP, $D_{\rm loss}$ increased at a similar rate for both the Tyr O atoms and aromatic ring atoms, suggesting that full ring conformational disordering is more likely. Indeed, no convincing reproducible Fourier difference peaks above the background map noise were observed around any Tyr terminal –OH groups.

The RNA-stabilization effects on protein are observed at short ranges and are restricted to within the RNA-binding interfaces around the TRAP ring. For example, Asp17 is located ∼6.8 Å from the G1 base, outside the RNA-binding interfaces, and has indistinguishable C^{γ} atom $D_{\rm loss}$ dose-dynamics between RNA-bound and nonbound TRAP ($p > 0.9$). An increase in the dose at which functionally important residues remain intact has biological ramifications for understanding the mechanisms at which ionizing radiation damage is mitigated within naturally forming DNA–protein and RNA–protein complexes. Observations of lower protein radiation-sensitivity in DNA-bound forms have been recorded in solution at RT at much lower doses (∼1 kGy) than those used for typical MX experiments [*e.g.* an oestrogen response element–receptor complex (Stísová *et al.*, 2006) and a DNA glycosylase and its abasic DNA target site (Gillard *et al.*, 2004)]. In these studies, the main damaging species is predicted to be the oxidizing hydroxyl radical produced through solvent irradiation, which is known to add to double covalent bonds within both DNA and RNA bases to induce strand breaks and base modification (Spotheim-Maurizot & Davídková, 2011; Chance *et al.*, 1997). It was suggested that physical screening of DNA by protein shielded the DNA–protein interaction sites from radical damage, yielding an extended life-dose for the nucleoprotein complex compared with separate protein and DNA constituents at RT.

However, in the current MX study at 100 K, the main damaging species are believed to be migrating LEEs and holes produced directly within the protein–RNA components or in closely associated solvent. The results presented here suggest that biologically relevant nucleoprotein complexes also exhibit prolonged life-doses under the effect of LEE-induced structural changes, involving direct physical protection of key RNA-binding residues. Such reduced radiation-sensitivity in this case ensures that the interacting protein remains bound long enough to the RNA to complete its function, even whilst exposed to ionizing radiation. Within the nonbound TRAP macromolecule, the acidic residues within the unoccupied RNA-binding interfaces (Asp39, Glu36, Glu42) are notably amongst the most susceptible residues within the asymmetric unit (Fig. 4). When exposed to X-rays, these residues will be preferentially damaged by X-rays and subsequently reduce the affinity with which TRAP binds to RNA. Within the cellular environment, this mechanism could reduce the risk that radiation-damaged proteins might bind to RNA, thus avoiding the detrimental introduction of incorrect DNA-repair, transcriptional and base-modification pathways.

The Python scripts written to calculate the per atom $D_{\rm loss}$ metric are available from the authors on request.

5. Related literature

The following references are cited in the Supporting Information for this article: Chen *et al.* (2010).

Acknowledgements

We thank Paul Gollnick from Buffalo University for providing the protein and RNA, and the ESRF beamline staff for help during Radiation Damage BAG beamtime. We gratefully acknowledge the UK Engineering and Physical Sciences Research Council for studentship funding in the Systems Biology Programme of the University of Oxford Doctoral Training Centre (CSB and MG). IC is supported by the US Department of Energy Office of Science, Office of Basic Energy Sciences under Award No. DE- FC02-04ER1553.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Alizadeh, E. & Sanche, L. (2014). *Eur. Phys. J. D*, **68**, 97.
Allan, E. G., Kander, M. C., Carmichael, I. & Garman, E. F. (2013). *J. Synchrotron Rad.* **20**, 23–36.
Antson, A. A., Dodson, E. J., Dodson, G., Greaves, R. B., Chen, X. & Gollnick, P. (1999). *Nature (London)*, **401**, 235–242.
Becke, A. D. (1993). *J. Chem. Phys.* **98**, 5648–5652.
Bourenkov, G. P. & Popov, A. N. (2010). *Acta Cryst.* D**66**, 409–419.
Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.
Bury, C., Garman, E. F., Ginn, H. M., Ravelli, R. B. G., Carmichael, I., Kneale, G. & McGeehan, J. E. (2015). *J. Synchrotron Rad.* **22**, 213–224.
Chance, M. R., Sclavi, B., Woodson, S. A. & Brenowitz, M. (1997). *Structure*, **5**, 865–869.
Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.
Dubnovitsky, A. P., Ravelli, R. B. G., Popov, A. N. & Papageorgiou, A. C. (2005). *Protein Sci.* **14**, 1498–1507.
Elliott, M. B., Gottlieb, P. A. & Gollnick, P. (2001). *RNA*, **7**, 85–93.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.
Fioravanti, E., Vellieux, F. M. D., Amara, P., Madern, D. & Weik, M. (2007). *J. Synchrotron Rad.* **14**, 84–91.
Frisch, M. J. *et al.* (2009). *Gaussian* 09. Gaussian Inc., Wallingford, Connecticut, USA.
Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.
Gerstel, M., Deane, C. M. & Garman, E. F. (2015). *J. Synchrotron Rad.* **22**, 201–212.
Gillard, N., Begusova, M., Castaing, B. & Spotheim-Maurizot, M. (2004). *Radiat. Res.* **162**, 566–571.
Holton, J. M. (2007). *J. Synchrotron Rad.* **14**, 51–72.
Holton, J. M. (2009). *J. Synchrotron Rad.* **16**, 133–142.
Hopcroft, N. H., Wendt, A. L., Gollnick, P. & Antson, A. A. (2002). *Acta Cryst.* D**58**, 615–621.
Jones, G. D., Lea, J. S., Symons, M. C. & Taiwo, F. A. (1987). *Nature (London)*, **330**, 772–773.
Leslie, A. G. W. & Powell, H. R. (2007). *Evolving Methods for Macromolecular Crystallography*, edited by R. J. Read & J. L. Sussman, pp. 41–51. Dordrecht: Springer.
Liebschner, D., Rosenbaum, G., Dauter, M. & Dauter, Z. (2015). *Acta Cryst.* D**71**, 772–778.
Matsui, Y., Sakai, K., Murakami, M., Shiro, Y., Adachi, S., Okumura, H. & Kouyama, T. (2002). *J. Mol. Biol.* **324**, 469–481.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

McGeehan, J. E., Streeter, S. D., Thresh, S. J., Ball, N., Ravelli, R. B. G. & Kneale, G. G. (2008). *Nucleic Acids Res.* **36**, 4778–4787.
Murray, J. & Garman, E. (2002). *J. Synchrotron Rad.* **9**, 347–354.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
O'Neill, P., Stevens, D. L. & Garman, E. (2002). *J. Synchrotron Rad.* **9**, 329–332.
Owen, R. L., Axford, D., Nettleship, J. E., Owens, R. J., Robinson, J. I., Morgan, A. W., Doré, A. S., Lebon, G., Tate, C. G., Fry, E. E., Ren, J., Stuart, D. I. & Evans, G. (2012). *Acta Cryst.* D**68**, 810–818.
Owen, R. L., Rudiño-Piñera, E. & Garman, E. F. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 4912–4917.
Ptasińska, S. & Sanche, L. (2007). *Phys. Rev. E*, **75**, 031915.
Ravelli, R. B. G. & McSweeney, S. M. (2000). *Structure*, **8**, 315–328.
Shimizu, N., Hirata, K., Hasegawa, K., Ueno, G. & Yamamoto, M. (2007). *J. Synchrotron Rad.* **14**, 4–10.
Simons, J. (2006). *Acc. Chem. Res.* **39**, 772–779.
Spotheim-Maurizot, M. & Davídková, M. (2011). *Mutat. Res.* **711**, 41–48.
Stísová, V., Goffinont, S., Spotheim-Maurizot, M. & Davídková, M. (2006). *Radiat. Prot. Dosimetry*, **122**, 106–109.
Symons, M. C. R. (1997). *Free Radical Biol. Med.* **22**, 1271–1276.
Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.
Tomasi, J., Mennucci, B. & Cancès, E. (1999). *J. Mol. Struct.* **464**, 211–226.
Weik, M., Ravelli, R. B. G., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 623–628.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Yano, J., Kern, J., Irrgang, K. D., Latimer, M. J., Bergmann, U., Glatzel, P., Pushkar, Y., Biesiadka, J., Loll, B., Sauer, K., Messinger, J., Zouni, A. & Yachandra, V. K. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 12047–12052.
Zeldin, O. B., Brockhauser, S., Bremridge, J., Holton, J. M. & Garman, E. F. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 20551–20556.
Zeldin, O. B., Gerstel, M. & Garman, E. F. (2013). *J. Appl. Cryst.* **46**, 1225–1230.

ARTICLE IN PRESS

Radiation Physics and Chemistry ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Radiation Physics and Chemistry

journal homepage: www.elsevier.com/locate/radphyschem

Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography

Charles S. Bury [a], Ian Carmichael [b], John E. McGeehan [c], Elspeth F. Garman [a,*]

[a] Laboratory of Molecular Biophysics, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK
[b] Notre Dame Radiation Laboratory, University of Notre Dame, Notre Dame, IN 46556, USA
[c] Molecular Biophysics, Institute of Biomedical and Biomolecular Sciences, University of Portsmouth, King Henry 1st Street, Portsmouth PO1 2DY, UK

HIGHLIGHTS

- We review radiation damages nucleoprotein complexes during X-ray crystallography.
- We detect radiation-induced chemical changes from electron density difference maps.
- We use a systematic pipeline to track electron density loss with increasing dose.
- Nucleic acids are radiation-insensitive compared to protein within crystals at 100 K.
- RNA protects key RNA-binding residues from radiation-induced decarboxylation.

ARTICLE INFO

Article history:
Received 1 April 2016
Accepted 25 May 2016

Keywords:
Protein-nucleic acid complexes
Macromolecular X-ray crystallography
Radiation damage
Electron density
Dose

ABSTRACT

In X-ray crystallography, for the determination of the 3-D structure of macromolecules, radiation damage is still an inherent problem at modern third generation synchrotron sources, even when utilising cryo-crystallographic techniques (sample held at 100 K). At doses of just several MGy, at which a typical diffraction dataset is collected, site-specific radiation-induced chemical changes are known to manifest within protein crystals, and a wide body of literature is now devoted to understanding the mechanisms behind such damage. Far less is known regarding radiation-induced damage to crystalline nucleic acids and the wider class of nucleoprotein complexes during macromolecular X-ray crystallography (MX) data collection. As the MX structural biology community now strives to solve structures for increasingly larger and complex macromolecular assemblies, it essential to understand how such structures are affected by the X-ray radiation used to solve them. The purpose of this review is to summarise advances in the field of specific damage to nucleoprotein complexes and to present case studies of MX damage investigations on both protein-DNA (C.Esp1396I) and protein-RNA (TRAP) complexes. To motivate further investigations into MX damage mechanisms within nucleoprotein complexes, current and emerging protocols for investigating specific damage within $F_{obs}(n) - F_{obs}(1)$ electron density difference maps are discussed.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Over the past century, macromolecular X-ray crystallography (MX) has proved an invaluable tool within structural biology, permitting the three-dimensional visualisation of proteins and nucleic acids at near atomic resolution. As of early 2016, a total of 104,244 structures of proteins and nucleoprotein complexes have been deposited within the Protein Data Bank (PDB); this world-wide effort has been instrumental in the correct characterisation of the molecular interactions underpinning the function of highly biologically-important macromolecules (Ferreira et al., 2004; Voorhees et al., 2009).

During a standard MX experiment, crystalline macromolecular samples are exposed to a monochromatic beam of ionising X-rays (with typical energies 8–15 keV) in order to generate a series of diffraction images, from which a temporal and spatial average of the macromolecular electron density throughout a crystal can be derived. However, for a typical 100 μm thick protein crystal exposed to a 12.4 keV X-ray beam, only a small proportion (2%) of incident photons are predicted to actually interact directly with the crystal, with only 8% of these elastically scattering to contribute to the desired diffraction pattern. The photoelectric and Compton effects are the dominant processes by which the incident

* Corresponding author.
E-mail addresses: charles.bury@dtc.ox.ac.uk (C.S. Bury), elspeth.garman@bioch.ox.ac.uk (E.F. Garman).

http://dx.doi.org/10.1016/j.radphyschem.2016.05.023
0969-806X/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

photon energy is absorbed by the crystal (84% and 8% respectively) (Garman, 2010). Consequently, for the current third generation high-brilliance synchrotron sources with photon fluxes in the range of 10^{10}–10^{13} ph/s into a 5–100 μm diameter spot, this results in a large deposition of energy within typical macromolecular crystals upon an X-ray exposure of a few frames of data collection.

The dose is defined as the energy absorbed in the crystal per unit mass (with units of gray, Gy=J/kg); protein crystals typically accumulate doses of the order of several MGy within the course of one diffraction dataset collection at 100 K. Increasing dose is coupled with increasing crystal non-isomorphism, eventually leading to *global radiation* damage effects: an overall reduction in the mean diffraction intensity, loss in resolution, and increases in both crystal mosaicity and unit cell volume (Murray and Garman, 2002; Ravelli and Garman, 2006; Weik et al., 2000). An experimental dose limit ($D_{0.7}$) of 30 MGy has been determined, cited as the upper dose at which the average diffraction intensity for *any* protein crystal held at 100 K will have decayed to 70%, and at which the biological information derived from the inferred structure of the crystalline protein may be compromised (Owen et al., 2006). Due to these various effects, radiation damage is still an inevitable hindrance to successful MX data collection and structure determination, even with the crystal held at cryocooled temperatures (100 K) and with modern advances in multi-crystal data collection methods (Stellato et al., 2014), high X-ray flux density nano- and micro-beamlines (I24) and faster pixel based detectors, also at room temperature (RT) (Owen et al., 2014). Such issues are exacerbated as experimenters push towards structure determination of larger and more complex macromolecules, which in turn typically involve collecting data from smaller crystals (< 1–5 μm). These have fewer unit cells compounded with intrinsic crystal disorder, resulting in weaker diffraction. There must always be a compromise between collecting more diffraction data and the inevitable progression of radiation damage with increasing dose.

Even before any observed reduction in diffraction resolution, at doses of several MGy, synchrotron X-ray irradiation has been widely reported to induce site-specific chemical and conformation changes to crystalline proteins held at 100 K. Such damage has been directly observable within electron density maps reconstructed from sequential diffraction patterns obtained with increasing dose (Weik et al., 2000) (Fig. 1), and a reproducible order of susceptibility has now been established within a variety of proteins at 100 K: metallocentre reduction, disulphide bond elongation and cleavage, acidic residue decarboxylation, and reported Tyr –OH group disordering and methionine S_{δ}–C_{ε} bond cleavage (Burmeister, 2000; Weik et al., 2000; Yano et al., 2005). The prevalence of these *specific radiation damage* (SRD) effects are suspected to be highly dependent on local crystalline protein environment, with factors such as solvent accessibility, proximity to high X-ray cross-section atoms, acidic residue protonation state and packing density (Fioravanti et al., 2007; Gerstel et al., 2015) predicted to affect the damage rates. However, investigations into a number of protein systems have revealed minimal correlation between such individual factors and SRD events, and it is thus suggested that a multitude of parameters contribute to a particular residue's susceptibility to SRD (Holton, 2009). For instance, the addition of a single ordered nitrate anion (NO_3^-) proximal to a disulphide bond will result in substantial stability of the otherwise highly reducible bond (De La Mora et al., 2011). At doses below 1 MGy, disulphide bond radicalisation (a precursor for disulphide breakage) has been detected by electron paramagnetic resonance (EPR) and *in situ* UV–visible absorption microspectrophotometry (Sutton et al., 2013). Radiation-induced structural changes within active sites of photosensitive proteins have also been detected at doses as low as 0.06 MGy (Borshchevskiy et al., 2014). As such, it is unclear whether a universal safe dose is achievable in MX before the onset of protein specific damage events.

Somewhat paradoxically, given the wealth of radiation damage studies on nucleic acids conducted by radiation chemists (e.g. (Alizadeh et al., 2015; Cadet et al., 1999; Michaud et al., 2012)), crystallographic investigations regarding MX radiation-induced changes to nucleic acids and the larger class of nucleoprotein complexes have been substantially less comprehensive to date, and a governing MX specific damage *rulebook* for them has not yet been established. Nucleic acid and nucleoprotein complexes now comprise approximately 6.4% of MX-derived structures deposited within the PDB. The structural biology community are currently pursuing increasingly large (> 200 kDa) and complex macromolecular systems encompassing those bound to nucleic acids, and a thorough characterisation of MX radiation damage within such complexes is essential to ensure correct structural interpretations at the atomistic scale provided by crystallography.

Nucleic acids have diverse roles in information exchange and control but also make up some of the fundamental structural and catalytic components of large macromolecular machines such as ribosomes and spliceosomes. A multitude of studies have investigated both nucleic acids and nucleoprotein complexes in solution at RT, where secondary diffusive hydroxyl radicals produced through solvent radiolysis can add to double covalent bonds within both DNA and RNA bases to induce SSBs and base modification (Chance et al., 1997; Spotheim-Maurizot and Davidkova, 2011), and oxidise protein residues with differential degrees of susceptibility (O'Neill et al., 2002). However many radical species (such as hydroxyl radicals) are immobilised below 110 K (Allan et al., 2013; Owen et al., 2012); since most modern MX experiments take place at 100 K, such hydroxyl-mediated damage is not

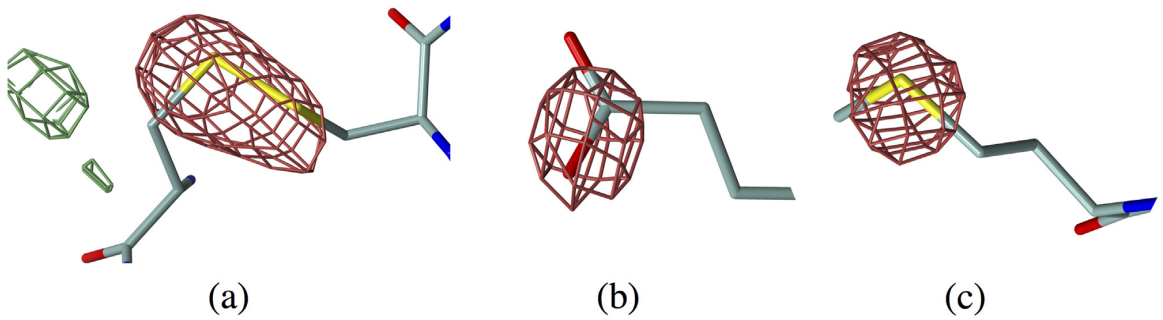


Fig. 1. (a) Disulphide bond cleavage, (b) Glu decarboxylation, and (c) Met sulphur disordering within *Torpedo californica* acetyl-cholinesterase (TcAChE) pdb: 1QID (Weik et al., 2000). $F_{obs}(5)-F_{obs}(1)$ Fourier difference maps between dataset 1 and 5 collected on the same crystal are shown, contoured at ±4σ, with negative difference density (red) indicating disordering of the atomic positions with accumulated dose. F_{obs} is the set of observed structure factors, proportional to the square root of the observed intensities recorded on the diffraction pattern. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

anticipated. Instead, the main damaging species are predicted to be low energy electrons (LEEs) with energies of several eV. After primary photoabsorption of an incoming X-ray photon by both the macromolecule and crystal solvent channels, each ejected high-energy primary photoelectron (as well as associated Auger electrons) is predicted to have enough energy to cause the production of up to ~500 secondary LEEs along their track length (3–4 μm for the incident energies employed in MX (Nave and Hill, 2005; Sanishvili et al., 2011)). These LEEs can then cause further ionisation events as they gradually thermalise throughout the crystal (O'Neill et al., 2002).

Electrons have been shown to be mobile at 77 K (Jones et al., 1987), with electron spin resonance (ESR) spectroscopy studies postulating rapid quantum tunnelling along both the protein backbone and nucleic acid stacked base π-systems as the dominant mechanism by which LEEs can relocate from initial ionisation sites at 100 K (Symons, 1997). Ultimately, these LEEs become trapped by electron-affinic sites in both protein and nucleic acid to form electron gain centres in proteins and $T^{\bullet-}$ and $C^{\bullet-}$ centres in nucleic acids. Positive holes are also believed to migrate at 100 K, but to a lesser extent, being more readily trapped, predominantly by deprotonation at protein backbone amide units and at acidic residue side-chain carboxyl groups, and to form $G^{\bullet+}$ centres in nucleic acids.

Extensive studies on sub-ionisation level LEEs (0–15 eV) interacting with short DNA oligonucleotides deposited on films at varying degrees of hydration at RT, 4 K, and 70 K have been performed to date (Alizadeh and Sanche, 2014; Ptasinska and Sanche, 2007). Transient resonant attachment by LEEs to π^* orbitals of DNA bases and the sugar-phosphate backbone, followed by rapid bond dissociation, is now believed sufficient for cleavage of both base-sugar N_1C bonds and strong (~4 eV) covalent bonds within the DNA phosphodiester backbone (Barrios et al., 2002; Berdys et al., 2004; Simons, 2006), with computational chemistry calculations indicating preferential cleavage of sugar-phosphate C–O bonds over other bonds present within the DNA structure (Théodore et al., 2006). Such bond ruptures are ultimately detected as yields of single strand breakages (SSBs), typically through gel electrophoresis analysis (Boudaïffa, 2000). Similar preferential rupture of C–O bonds (in particular the 5′ C–O bond) has been reported in a recent X-ray absorption spectroscopy study for DNA oligonucleotides in solution under either UVA photon or proton irradiation (Czapla-Masztafiak et al., 2016).

Such LEE-mediated damage observations are thus also anticipated within crystalline DNA at 100 K upon ionising X-ray radiation exposure. However, an MX investigation coupled with Raman spectroscopy (McGeehan et al., 2007) into X-ray radiation-induced damage to cryocooled brominated oligonucleotide crystals observed no clear signs of native DNA specific damage, with DNA debromination as the only chemical change detected, even at the highest accumulated dose tested (> 15 MGy). Although anecdotal reports suggest that native DNA is substantially resilient to X-ray irradiation at the MGy doses typically reported for MX data collection (with no clear or reproducible deterioration of the electron density with accumulated dose), this remains to be systematically tested and characterised.

The following sections provide a discussion of MX radiation damage case studies on two bacterial nucleoprotein complexes: the C.Esp1396I-DNA complex and the TRAP-RNA complex. In previous studies, MX protein specific damage has been typically characterised by visually interpreting $F_{obs}(n) - F_{obs}(1)$ Fourier difference maps. However such time consuming visual inspection is limited by the subjective bias of the investigator. The development of scripted pipelines is discussed, designed to mitigate such bias and systematically categorise radiation-induced chemistry within the current crystalline nucleoprotein case studies. Such automated approaches are applicable to any MX protein or nucleic acid crystal study, with the case studies presented here designed to motivate further MX specific damage systematic investigations.

2. The C.Esp1396I-DNA complex: a model protein-DNA MX damage investigation

To characterise and investigate the apparent DNA radiation-insensitivity that was observed for native DNA crystals at 100 K, the relative rates of specific damage have recently been determined for a well-studied bacterial protein-DNA complex (Bury et al., 2015) (C.Esp1396I, pdb: 3clc, resolution: 2.8 Å (McGeehan et al., 2008)).

'Controller' C-protein dimers, such as C.Esp1396I, are part of a large helix-turn-helix protein family, acting as transcriptional regulators of gene expression in both eukaryotic and prokaryotic systems. In the C.Esp1396I complex (Fig. 2), C-protein dimers (chains A and B, and C and D) bind to the 35bp DNA operator sequence (chains E and F) of a bacterial restriction-modification (R-M) system Esp1396I, to regulate C-protein and endonuclease gene expression in a concentration-dependent manner (Ball et al., 2009). The C.Esp1396I complex (~60 kDa) was chosen as a model biologically-relevant protein-DNA complex due to the large 35bp dsDNA component, resulting in the number of protein and DNA atoms being of the same order (protein: 2496 atoms, DNA: 1429 atoms). A statistical comparison was conducted between SRD rates to interacting DNA and protein constituents, determining protein to be significantly more damage susceptible within the complex over the wide dose range investigated (2.1–44.6 MGy over 8 MX datasets, n = 1,..,8, collected on the same crystal held at 100 K) (Bury et al., 2015).

As is conventionally performed in protein MX specific damage studies (Southworth-Davies et al., 2007), sites of significant radiation-induced electron density disordering within the C.Esp1396I complex were located by calculating Fourier difference maps with increasing dose ($F_{obs}(n) - F_{obs}(1)$ for n = 2, …, 8) over the crystal asymmetric unit. Difference peaks within $F_{obs}(n) - F_{obs}(1)$ maps provide a visual representation of regions of high electron density loss or gain with accumulated crystal dose. For the C.Esp1396I complex, within the first difference map ($F_{obs}(2) - F_{obs}(1)$, 6.2 MGy) clear density loss was localised around acidic residue side-chain carboxyl groups, indicative of radiation-induced decarboxylation, and around methylthio sidechain groups, consistent with Met CH_3–S covalent bond cleavage (Fig. 2(a)-(b)). Such observations were highly consistent with previous reports of protein SRD (Weik et al., 2000) (no disulphide bonds are present within this protein).

At higher doses (> 14.4 MGy), several locations of SRD were detected in close proximity (within 2 Å) of the DNA, including possible sugar-phosphate C–O bond cleavage between the T24 and A25 nucleotides of DNA chain F (Fig. 2(c)–(d)). Additionally, positive electron density accumulation was observed with increasing dose near the T24 and A25 bases; this is consistent with mechanisms of LEE attachment to nucleobases as suggested within oligonucleotide film studies (Alizadeh et al., 2013), or base modification induced by close proximity solvent free radicals (Cadet et al., 1999) (for example, hydroxyl radical binding to carbon 6 in T24). The location of such a SSB correlates with a DNA region that is both AT-rich and under significant strain as a consequence of large-scale deformation due to protein binding. Our suggestion is that such strained geometries could enhance radiation damage effects in DNA and would have major biological consequences, since eukaryotic DNA wrapping around histones in part relies on the distortion of DNA around such AT-rich sites.

Whereas $F_{obs}(n) - F_{obs}(1)$ Fourier electron density difference

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

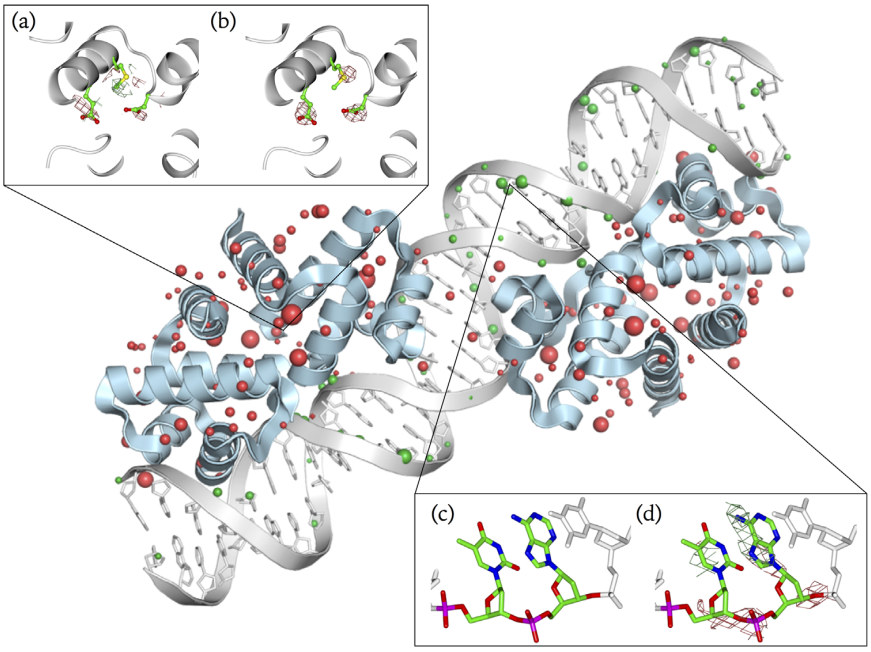


Fig. 2. (centre) The C.Esp1396I complex crystal asymmetric unit. Two C-protein dimers (chains A and B, left, and chains C and D, right) bind to the 35bp DNA operator sequence (chains E and F). Spheres represent sites of detected SRD within the highest dose dataset (44.6 MGy); red/green spheres indicate protein/DNA damage, with the sphere radii representing the magnitude of electron density loss at each SRD site. (inserts) Protein and DNA damage sites in C.Esp1396I: (a)–(b) protein chain D, Glu-54, Met-57 and Asp-64 (green, left to right), and (c)–(d) DNA chain F nucleotides T24 and A25 (with 5′ to 3′ end from left to right in each image) at (a), (c) low-dose (6.2 MGy) and (b), (d) high dose (44.6 MGy). $F_{obs}(n)-F_{obs}(1)$ Fourier difference maps are contoured at $\pm 3.0\sigma$ in green/red throughout. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

map peaks coincident with the C.Esp1396I complex provide an insight into SRD events, with increasing dose, difference maps become intrinsically noisy due to overall degradation of diffraction data quality (global radiation damage) and increasingly unmodelled chemistry within crystal bulk solvent regions. Accordingly, for the C.Esp1396I complex at 44.6 MGy, 10^3 difference peaks were detected above a 0.04 $eÅ^{-3}$ threshold (at which damage sites had been visually distinguishable from noise within the lowest dose $F_{obs}(2)-F_{obs}(1)$ difference map). To facilitate case-by-case inspection of difference peaks, a custom *python*-scripted pipeline (Bury et al., 2015) was implemented to systematically and objectively isolate damage sites throughout the C.Esp1396I complex with increasing dose (Fig. 3). Using this protocol, significant damage was detected for Asp, Glu, Met and Ser residues (electron density loss indicating disordering of the Ser side-chain –OH) even at the lowest doses (Fig. 3(a)). At higher doses, Arg and Asn exhibit electron density loss/disorder to their main-chain carboxyl group oxygen, and Ile and Lys experience partial density loss around the side-chain aliphatic and lysyl side-chains respectively. The remaining amino acids exhibited minimal SRD even at very high dose. Overall the detected damage onset was at significantly higher doses for DNA than for protein component (>20.6 MGy versus >6.2 MGy), with more homogeneously distributed damage between the four nucleotide types in contrast to the damage preferentially distributed amongst specific protein residues (Fig. 3(b)).

Our findings suggest DNA damage dynamics are indiscriminate of nucleotide identity and are evident only at high doses, where global radiation damage effects are anticipated. Such preferential damage to protein constituents within nucleoprotein complexes in solution has been indicated for the lac repressor-operator (Beguszová et al., 2001; Charlier et al., 2002; Eon et al., 2001) and estrogen response element-receptor complexes (Stisova et al., 2006). The accumulated doses typically reported for these aqueous studies (~1 kGy) were much lower than in the C.Esp1396I MX experiment, and conducted at RT where damage is predominantly mediated by $OH^{\bullet}$ radical attack. The C.Esp1396I study additionally provides evidence for preferential protein damage within crystalline nucleoprotein complexes under a high dose regime (MGy) at 100 K, where LEEs act as the dominant damaging species.

Multiple mechanisms have been proposed for DNA radiation-protection by interacting amino acids upon exposure to LEEs, both in solution and the condensed phase through: (a) short-range protein-DNA proton transfer to stabilise LEEs captured by nucleobases, (b) direct electron scavenging by proximal amino acids, and (c) inaccessibility of the nucleotides' lowest unoccupied molecular orbitals (LUMOs) for LEE attachment due to direct engagement with interacting protein (Gu et al., 2014; Solomun and Skalický, 2008). However, for crystalline complexes such as C.Esp1396I, the relevance of such protective mechanisms is unclear, and whether protein is intrinsically more susceptible to X-ray induced damage, or whether the protein scavenges electrons to protect DNA remains to be determined.

3. The TRAP-RNA complex: a model protein-RNA MX damage investigation

An MX SRD experiment has been conducted on the TRAP-RNA complex (Bury et al., 2016), with 10 diffraction datasets (n=1, …,10) collected at 100 K on the same crystal over a large dose range (1.31–25.0 MGy). TRAP fortuitously crystallises in a 1:1 ratio of RNA-bound and non-bound protein within the same crystal asymmetric unit (Hopcroft et al., 2002). It thus provides an ideal controlled experiment under identical crystallisation and MX data collection conditions to investigate the exact role of nucleic acid

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

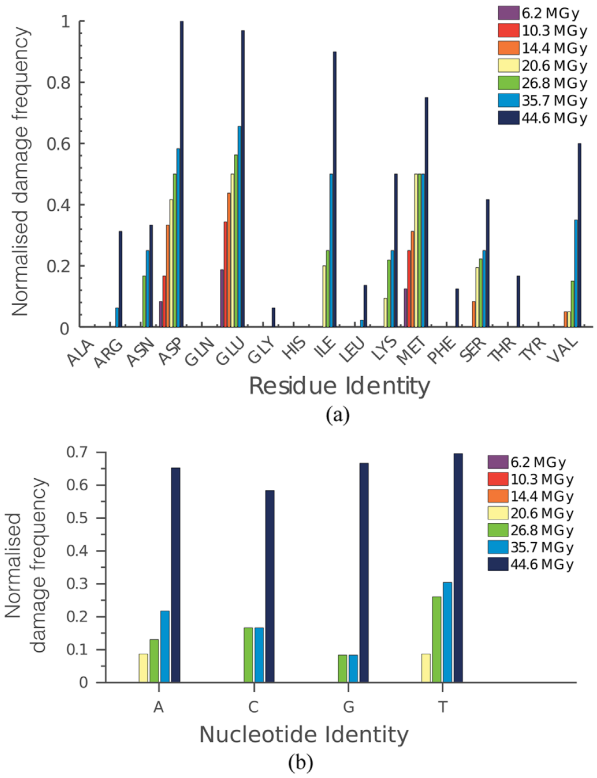


Fig. 3. Normalised frequency of detected SRD events against (a) protein residue type over chains A to D and (b) DNA nucleotides over chains E and F (normalised to the frequency of occurrence of each residue/nucleotide throughout the structure) for each dose (all doses quoted in this review from our work are DWD (Zeldin et al., 2013), diffraction-weighted dose, MGy).

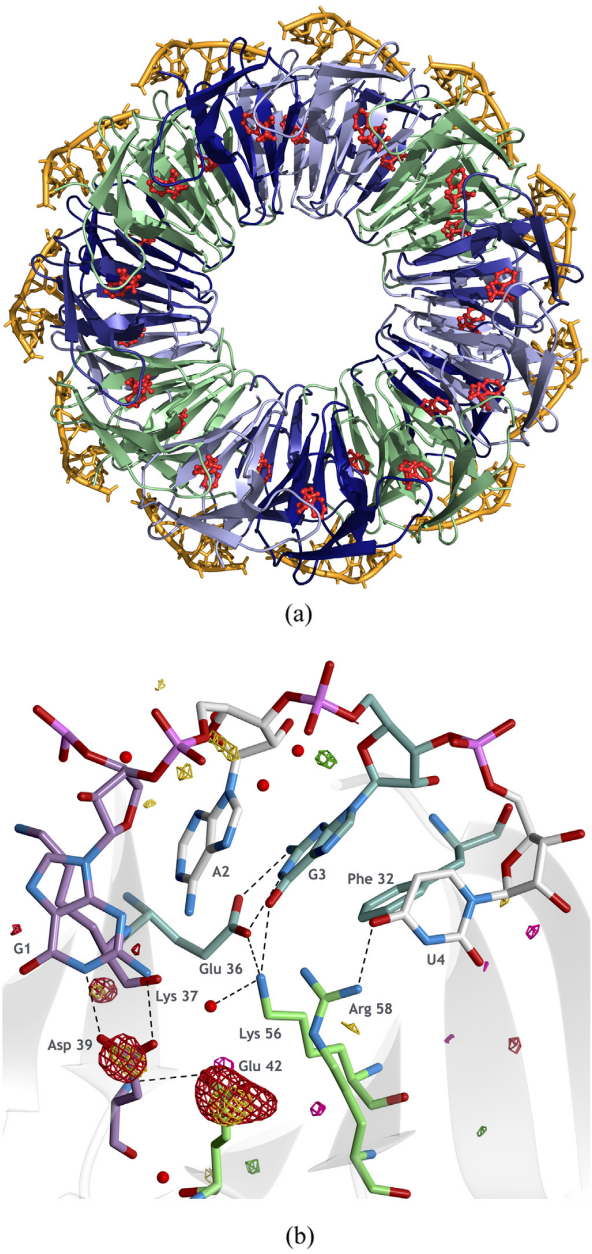


Fig. 4. (a) The TRAP-$(GAGUU)_{10}GAG$ complex asymmetric unit (resolution: 1.98 Å). Bound tryptophan ligands and RNA are represented in red and orange respectively. (b) One of 11 identical RNA-binding interfaces around the RNA-bound TRAP ring (chain N). $F_{obs}(n)-F_{obs}(1)$ Fourier difference maps are overlaid at ± 3.5 σ-level, at low-dose (n=2; DWD: 3.88 MGy) in pink/yellow, and high-dose (n=10; DWD: 25.0 MGy) in green/red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

binding on protein damage susceptibility within a crystalline biologically-relevant nucleoprotein complex.

Trp RNA-binding attenuation protein (TRAP) regulates transcription of tryptophan biosynthetic genes in *Bacillus subtilis* and several other bacteria through an attenuation mechanism (Antson et al., 1999). TRAP consists of eleven identical protein chains, arranged as a ring structure with 11-fold rotational symmetry (~91 kDa, Fig. 4(a)). Once activated through binding to L-tryptophan within eleven inter-subunit hydrophobic pockets symmetrically spaced around the TRAP ring, TRAP binds with high-affinity (Elliott et al., 2001) (K_d ~1.1 nM) to a specific RNA sequence within the leader segment of the nascent *trpEDCFBA* operon mRNA. In the current investigation, TRAP is bound to the 53 base RNA sequence $(GAGUU)_{10}GAG$. Bases of G1-A2-G3 nucleotides form direct hydrogen bonds to TRAP, whereas the U4-U5 nucleotides do not bind directly to the protein and are more flexible.

In a procedure similar to that employed in the C.Esp1396I complex investigation, a series of Fourier difference maps ($F_{obs}(n)-F_{obs}(1)$ for datasets n=2, …, 10) were generated to locate SRD sites throughout the large TRAP-RNA complex with increasing dose. A new damage metric, D_{loss}, was introduced as the magnitude of the most negative Fourier difference map value assigned within a local region around each TRAP atom. It was computed for each refined atom at each dose state, in order to quantify, on a per-atom basis, radiation-induced electron density changes within the series of Fourier difference maps (see (Bury et al., 2016) for details). Large positive D_{loss} values indicate radiation-induced atomic disordering reproducibly throughout crystal unit cells, relative to an initial low-dose dataset (here at 1.31 MGy). This approach avoided the previous necessity for visually inspection of Fourier difference maps to detect SRD sites, which has proved a cumbersome limiting factor to previous specific damage investigations in MX (Bury et al., 2015; Fioravanti et al., 2007; Weik et al., 2000).

TRAP is highly Glu and Asp rich (220 in total within the asymmetric unit). However, it does not contain any cysteine or refinable methionine residues. The D_{loss} damage metric

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

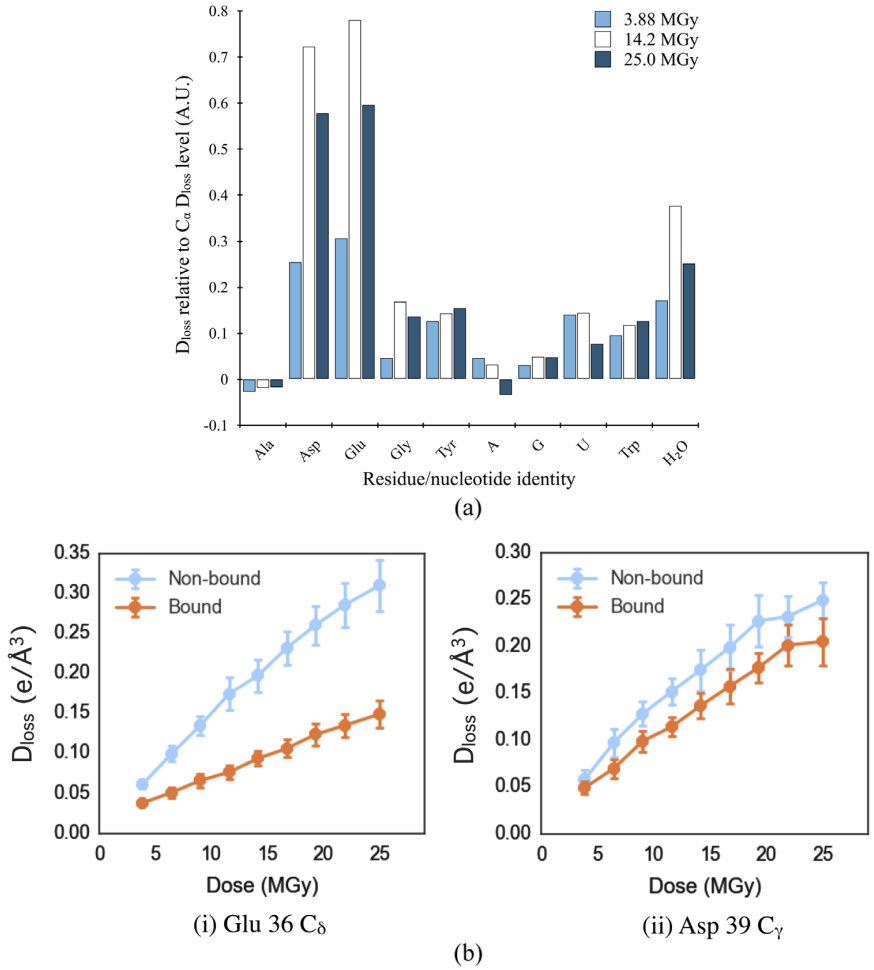


Fig. 5. (a) D_{loss} metric (relative to mean D_{loss} metric over the set of all main-chain protein C_α atoms within TRAP) for a subset of residue and nucleotide types within the asymmetric unit for 3 increasing doses. Positive values indicate greater electron density loss/disordering than that detected for the set of TRAP C_α atoms at the same dose. The TRAP structure only contains non-covalently bound tryptophan ligands, shown at the right of the plot. Damage to ordered solvent is also shown. Note that at the highest doses, global damage increases the non-specific electron density disorder, as evidenced by larger C_α atom D_{loss} values, and this effect decreases the relative D_{loss} of susceptible residue types. (b) D_{loss} progression with increasing dose for (i) Glu-36C_δ and (ii) Asp-39C_γ atoms within TRAP, with atoms grouped by location within the asymmetric unit (RNA-bound or non-bound TRAP ring). 95% confidence intervals are shown, calculated over the 11 equivalent atoms within each ring.

successfully detected typical forms of protein SRD within the TRAP-RNA complex, with Glu and Asp side-chain decarboxylation events interpretable within the first difference map $F_{obs}(2)-F_{obs}(1)$ (Figs. 4(b), 3.9 MGy) and flagged by the D_{loss} metric (Fig. 5(a)). Below 20 MGy, Fourier difference maps revealed negligible radiation-induced electron density deterioration in the region of the refined RNA nucleotides, in direct contrast to the clear protein SRD events noted above. Only above 20 MGy was evidence present for density loss observable around C–O bonds and phosphate groups of the RNA phosphodiester backbone (at dose levels consistent with the C.Esp1396I-DNA study). However, the median D_{loss} calculated over all RNA P atoms at 25 MGy was more than a factor of 2 lower than that of Glu-C_δ and Asp-C_γ atoms, and of the same order as the radiation-insensitive Gly-C_α atoms within TRAP. Therefore little evidence was provided for RNA SSB events consistently throughout the unit cells of the TRAP-RNA crystal, even at high dose values. It was thus concluded that RNA, similarly to DNA, was highly radiation-insensitive relative to protein within crystalline nucleoproteins at 100 K.

For the large number of Glu and Asp residues in TRAP (6 Glu and 4 Asp residues per TRAP monomer) a strong dependence of acidic residue decarboxylation susceptibility on local environment was statistically established using the high 11-fold symmetry around TRAP. The class of acidic residues situated within the 11 identical RNA-binding interfaces distributed around the TRAP ring exhibited greater radiation-induced disordering and subsequent decarboxylation for non-bound TRAP (Fig. 5(b)). The reduction in D_{loss} upon RNA-binding was most significant for Glu-36 (Hotelling T-squared test: p-value$=6.1\times10^{-5}$), the carboxyl side-group of which interacts directly with the G3 nucleobase by accepting two similar length hydrogen bonds ($\sim$2.7 Å on average). In the current TRAP-RNA asymmetric unit, Asp-39 was observed to accept two hydrogen bonds from the two nitrogens of the G1 RNA base, similar to the bonding of Glu-36 to G3. However, a less significant reduction in D_{loss} was observed for Asp-39 upon RNA-binding (Hotelling T-squared test: p-value$=0.093$). We suggest that the relative protection provided by RNA-binding to Glu-36 and Asp-39 directly correlates with the functional importance of each residue in high affinity RNA-binding to TRAP. Whereas nucleoside analogue studies have demonstrated that guanine replacement with

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

inosine at position G3 within each repeating GAG triplet resulted in >1000-fold decrease on TRAP binding affinity, inosine substitution at position G1 resulted in less than a 2-fold decrease in binding affinity (Elliott et al., 1999). Indeed, in other crystallographic studies on TRAP (Hopcroft et al., 2004, 2002), Asp-39 is documented to exhibit only a single distorted hydrogen bond to G1, or none at all in the case where G1 is substituted for uracil, without any loss in binding affinity (for TRAP bound to a repeating UAGAU RNA sequence, pdb accession code: 1utf).

The significant reduction in radiation-induced disordering upon RNA-binding was also exhibited for both the aliphatic side-chain of Lys-37 (Hotelling T-squared test, $p=0.024$ for Lys-37C_ε atoms) and the aromatic ring of Phe-32 (Hotelling T-squared test, Phe-32C_ζ, $p=0.0014$), which stack against bases G1 and G3 within each of the 11 RNA-binding interfaces around TRAP. This finding corresponds well with previous alanine mutagenesis studies, which identified the Lys-37 stacking interaction as essential for high affinity RNA-binding (Yang et al., 1997).

An aqueous RT study on the DNA glycosylase Fpg and its abasic DNA target site (Gillard et al., 2004) has highlighted a similar lower radiation-sensitivity for the protein in its DNA-bound form, through a combination of spectroscopy and mass spectrometry, although at much lower doses (~1 kGy). From this study, it was suggested that the DNA physically shields key protein-DNA interactions sites from hydroxyl mediated-damage, ultimately extending the total *life-dose* of the complex under irradiation. The above findings indicate a short-range protective ability of the RNA to otherwise highly radiation-sensitive protein residues in a crystalline nucleoprotein complex held at 100 K. The stability of the complex is governed by these repeated interactions between the protein and RNA around the TRAP ring. Such complexes may thus also exhibit enhanced life-dose values in a crystalline environment where LEEs are perceived to be the main mediator of radiation-induced structural changes.

4. Conclusion

The striking radiation-insensitivity of crystalline nucleic acids when exposed to X-ray ionising radiation at 100 K during MX data collection is perhaps somewhat surprising given the abundance of observations of LEE-induced covalent bond breakage in RT solution studies on both DNA and RNA, in which much lower doses are accumulated (~1 kGy). It could be postulated that in the crystal environment at 100 K, tight crystal packing could obstruct the development of large nucleic acid deformation events (single strand breakage and base-sugar cleavage) reproducibly throughout crystal unit cells. We also note that much of the radiation chemical investigation of nucleic acids in aqueous solution revolves around OH radical interactions, which we believe to be minimised at the 100 K cryo-temperatures typically used in MX. The dissociative electron attachment (DEA) studies have essentially been conducted on isolated bases in the gas phase (or on solid surfaces under ultrahigh vacuum conditions) where the base is the sole target (Ptasinska and Sanche., 2007). As such, the applicability of these DEA studies is unclear for crystalline nucleic acid polymers interacting with protein at 100 K.

In previous MX radiation damage studies, radiation chemistry literature has been consulted to aid explanation of protein SRD mechanisms within crystals at 100 K. It has been suggested that disulphide bond breakage results from LEE-mediated RSSR group reduction, and that Glu and Asp decarboxylation is induced by oxidative electron-loss centre trapping at their carboxylate group (Burmeister, 2000; Garrison, 1987; Sevilla et al., 1979). However, again the validity in extending such mechanisms to describe the crystalline state at 100 K is questionable, and indeed a more complex, multi-ionisation model for disulphide bridge cleavage (triggered by intersecting LEE tracks) has now been proposed (Sutton et al., 2013). Similarly, for MX experiments conducted on nucleic acids/nucleoprotein crystals held at 100 K, in which accumulated doses approach the order of MGy, it is thus now vital to continue characterising radiation-induced chemistry within the crystalline state, such that damage mechanisms are not falsely extended from studies at quite different regimes (RT, relative low-dose (~kGy), in solution, small oligonucleotides).

The current nucleoprotein MX radiation damage case studies have been presented to motivate further systematic investigations on other complex macromolecules, whilst also describing the introduction of systematic approaches to investigate $F_{obs}(n)-F_{obs}(1)$ Fourier difference maps with minimal subjective bias. In crystals at 100 K, we observe X-ray radiation protection of key RNA-binding residues and localisation of possible DNA phosphate backbone damage at AT-rich sites of high conformational strain. These findings provide an indication of the complexity at which LEE-mediated damage can manifest within intact nucleoprotein complexes. Only with further investigations will the prevalence of such damage events within the diverse community of nucleoprotein complexes be determined and the underlying mechanisms governing such damage dynamics elucidated.

Acknowledgements

We gratefully acknowledge the UK Engineering and Physical Sciences Research Council (Grant: EP/G03706X/1) for studentship funding in the Systems Biology Programme of the University of Oxford Doctoral Training Centre (CSB). IC is supported by the US Department of Energy Office of Science, Office of Basic Energy Sciences under Award Number DE-FC02–04ER1553.

References

Alizadeh, E., Orlando, T.M., Sanche, L., 2015. Biomolecular damage induced by ionizing radiation: the direct and indirect effects of low-energy electrons on DNA. Annu. Rev. Phys. Chem. 66, 379–398.
Alizadeh, E., Sanche, L., 2014. Low-energy-electron interactions with DNA: approaching cellular conditions with atmospheric experiments. Eur. Phys. J. D 68, 1–13.
Alizadeh, E., Sanz, A.G., García, G., Sanche, L., 2013. Radiation damage to DNA: the indirect effect of low energy electrons. J. Phys. Chem. Lett. 19, 820–825.
Allan, E.G., Kander, M.C., Carmichael, I., Garman, E.F., 2013. To scavenge or not to scavenge, that is STILL the question. J. Synchrotron Radiat. 20, 23–36.
Antson, A.A., Dodson, E.J., Dodson, G., Greaves, R.B., Chen, X., Gollnick, P., 1999. Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. Nature 401, 235–242.
Ball, N., Streeter, S.D., Kneale, G.G., McGeehan, J.E., 2009. Structure of the restriction-modification controller protein C.Esp1396I. Acta Cryst. D Biol. Cryst. 65, 900–905.
Barrios, R., Skurski, P., Simons, J., 2002. Mechanism for damage to DNA by low-energy electrons. J. Phys. Chem. B 106, 7991–7994.
Begusová, M., Eon, S., Sy, D., Culard, F., Charlier, M., Spotheim-Maurizot, M., 2001. Radiosensitivity of DNA in a specific protein-DNA complex: the lac repressor-lac operator complex. Int. J. Radiat. Biol. 77, 645–654.
Berdys, J., Skurski, P., Simons, J., 2004. Damage to model DNA fragments by 0.25–1.0 eV electrons attached to a thymine pi* orbital. J. Phys. Chem. B 108, 5800–5805.
Borshchevskiy, V., Round, E., Erofeev, I., Weik, M., Ishchenko, A., Gushchin, I., Mishin, A., Willbold, D., Buldt, G., Gordeliy, V., 2014. Low-dose X-ray radiation induces structural alterations in proteins. Acta Cryst. D Biol. Cryst. 70, 2675–2685.
Boudaïffa, B., 2000. Resonant formation of DNA strand breaks by low-energy (3–20 eV) electrons. Science 287, 1658–1660.
Burmeister, W.P., 2000. Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Cryst. D Biol. Cryst. 56, 328–341.
Bury, C., Garman, E.F., Ginn, H.M., Ravelli, R.B., Carmichael, I., Kneale, G., McGeehan, J.E., 2015. Radiation damage to nucleoprotein complexes in macromolecular crystallography. J. Synchrotron Radiat. 22, 213–224.
Bury, C.S., McGeehan, J.E., Antson, A.A., Carmichael, I., Gerstel, M., Shevtsov, M.B., Garman, E.F., 2016. RNA protects a nucleoprotein complex against radiation

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

damage. Acta Cryst. D. Biol. Cryst. 75, 648–657.
Cadet, J., Delatour, T., Douki, T., Gasparutto, D., Pouget, J.P., Ravanat, J.L., Sauvaigo, S., 1999. Hydroxyl radicals and DNA base damage. Mutat. Res. 424, 9–21.
Chance, M.R., Sclavi, B., Woodson, S.A., Brenowitz, M., 1997. Examining the conformational dynamics of macromolecules with time-resolved synchrotron X-ray 'footprinting'. Structure 5, 865–869.
Charlier, M., Eon, S., Sèche, E., Bouffard, S., Culard, F., Spotheim-Maurizot, M., 2002. Radiolysis of lac repressor by gamma-rays and heavy ions: a two-hit model for protein inactivation. Biophys. J. 82, 2373–2382.
Czapla-Masztafiak, J., Szlachetko, J., Milne Christopher, J., Lipiec, E., Sá, J., Penfold Thomas, J., Huthwelker, T., Borca, C., Abela, R., Kwiatek Wojciech, M., 2016. Investigating DNA radiation damage using X-ray absorption spectroscopy. Biophys. J. 110, 1304–1311.
De La Mora, E., Carmichael, I., Garman, E.F., 2011. Effective scavenging at cryotemperatures: further increasing the dose tolerance of protein crystals. J. Synchrotron Radiat. 18, 346–357.
Elliott, M.B., Gottlieb, Pa, Gollnick, P., 1999. Probing the TRAP-RNA interaction with nucleoside analogs. RNA 5, 1277–1289.
Elliott, M.B., Gottlieb, P.A., Gollnick, P., 2001. The mechanism of RNA binding to TRAP: initiation and cooperative interactions. RNA 7, 85–93.
Eon, S., Culard, F., Sy, D., Charlier, M., Spotheim-Maurizot, M., 2001. Radiation disrupts protein–DNA complexes through damage to the protein. The lac repressor–operator system. Radiat. Res. 156, 110–117.
Ferreira, K.N., Iverson, T.M., Maghlaoui, K., Barber, J., Iwata, S., 2004. Architecture of the photosynthetic oxygen-evolving center. Science 303, 1831–1838.
Fioravanti, E., Vellieux, F.M., Amara, P., Madern, D., Weik, M., 2007. Specific radiation damage to acidic residues and its relation to their chemical and structural environment. J. Synchrotron Radiat. 14, 84–91.
Garman, E.F., 2010. Radiation damage in macromolecular crystallography: what is it and why should we care? Acta Cryst. D Biol. Cryst. 66, 339–351.
Garrison, W.M., 1987. Reaction mechanisms in the radiolysis of peptides, polypeptides, and proteins. Chem. Rev. 87, 381–398.
Gerstel, M., Deane, C.M., Garman, E.F., 2015. Identifying and quantifying radiation damage at the atomic level. J. Synchrotron Radiat. 22, 201–212.
Gillard, N., Begusova, M., Castaing, B., Spotheim-Maurizot, M., 2004. Radiation affects binding of Fpg repair protein to an abasic site containing DNA. Radiat. Res. 162, 566–571.
Gu, B., Smyth, M., Kohanoff, J., 2014. Protection of DNA against low-energy electrons by amino acids: a first-principles molecular dynamics study. PCCP 16, 24350–24538.
Holton, J.M., 2009. A beginner's guide to radiation damage. J. Synchrotron Radiat. 16, 133–142.
Hopcroft, N.H., Manfredo, A., Wendt, A.L., Brzozowski, A.M., Gollnick, P., Antson Aa, 2004. The interaction of RNA with TRAP: the role of triplet repeats and separating spacer nucleotides. J. Mol. Biol. 338, 43–53.
Hopcroft, N.H., Wendt, A.L., Gollnick, P., Antson, A.A., 2002. Specificity of TRAP-RNA interactions: crystal structures of two complexes with different RNA sequences. Acta Cryst. D Biol. Cryst. 58, 615–621.
Jones, G.D., Lea, J.S., Symons, M.C., Taiwo, F.A., 1987. Structure and mobility of electron gain and loss centres in proteins. Nature 330, 772–773.
McGeehan, J.E., Carpentier, P., Royant, A., Bourgeois, D., Ravelli, R.B.G., 2007. X-ray radiation-induced damage in DNA monitored by online Raman. J. Synchrotron Radiat. 14, 99–108.
McGeehan, J.E., Streeter, S.D., Thresh, S.J., Ball, N., Ravelli, R.B., Kneale, G.G., 2008. Structural analysis of the genetic switch that regulates the expression of restriction-modification genes. Nucleic Acids Res. 36, 4778–4787.
Michaud, M., Bazin, M., Sanche, L., 2012. Measurement of inelastic cross sections for low-energy electron scattering from DNA bases. Int. J. Radiat. Biol. 88, 15–21.
Murray, J., Garman, E., 2002. Investigation of possible free-radical scavengers and metrics for radiation damage in protein cryocrystallography. J. Synchrotron Radiat. 9, 347–354.
Nave, C., Hill, M.A., 2005. Will reduced radiation damage occur with very small crystals? J. Synchrotron Radiat. 12, 299–303.
O'Neill, P., Stevens, D.L., Garman, E.F., 2002. Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective. J. Synchrotron Radiat. 9, 329–332.
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., Ren, J., Stuart, D.I., Evans, G., 2012. Outrunning free radicals in room-temperature macromolecular crystallography. Acta Cryst. D Biol. Cryst. 68, 810–818.
Owen, R.L., Paterson, N., Axford, D., Aishima, J., Schulze-Briese, C., Ren, J., Fry, E.E., Stuart, D.I., Evans, G., 2014. Exploiting fast detectors to enter a new dimension in room-temperature crystallography. Acta Cryst. D Biol. Cryst. 70, 1248–1256.
Owen, R.L., Rudino-Pinera, E., Garman, E.F., 2006. Experimental determination of the radiation dose limit for cryocooled protein crystals. Proc. Natl. Acad. Sci. USA 103, 4912–4917.
Ptasinska, S., Sanche, L., 2007. Dissociative electron attachment to hydrated single DNA strands. Phys. Rev. E 75, 031915.
Ravelli, R.B.G., Garman, E.F., 2006. Radiation damage in macromolecular cryocrystallography. Curr. Opin. Struct. Biol. 16, 624–629.
Sanishvili, R., Yoder, D.W., Pothineni, S.B., Rosenbaum, G., Xu, S., Vogt, S., Stepanov, S., Makarov, Oa, Corcoran, S., Benn, R., NagarajanV, R., Smith, J.L., Fischetti, R.F., 2011. Radiation damage in protein crystals is reduced with a micron-sized X-ray beam. Proc. Natl. Acad. Sci. USA 108, 6127–6132.
Sevilla, M.D., D'Arcy, J.B., Morehouse, K.M., 1979. An electron spin resonance study of γ-irradiated frozen aqueous solutions containing dipeptides. Mechanisms of radical reaction. J. Phys. Chem. 83, 2887–2892.
Simons, J., 2006. How do low-energy (0.1–2 eV) electrons cause DNA-strand breaks? Acc. Chem. Res. 39, 772–779.
Solomun, T., Skalický, T., 2008. The interaction of a protein–DNA surface complex with low-energy electrons. Chem. Phys. Lett. 453, 101–104.
Southworth-Davies, R.J., Medina, M.A., Carmichael, I., Garman, E.F., 2007. Observation of decreased radiation damage at higher dose rates in room temperature protein crystallography. Structure 15, 1531–1541.
Spotheim-Maurizot, M., Davidkova, M., 2011. Radiation damage to DNA in DNA-protein complexes. Mutat. Res. 711, 41–48.
Stellato, F., Oberthur, D., Liang, M., Bean, R., Gati, C., Yefanov, O., Barty, A., Burkhardt, A., Fischer, P., Galli, L., Kirian, R.A., Meyer, J., Panneerselvam, S., Yoon, C.H., Chervinskii, F., Speller, E., White, T.A., Betzel, C., Meents, A., Chapman, H.N., 2014. Room-temperature macromolecular serial crystallography using synchrotron radiation. IUCrJ 1, 204–212.
Stisova, V., Goffinont, S., Spotheim-Maurizot, M., Davidkova, M., 2006. Radiation damage to DNA-protein specific complexes: estrogen response element-estrogen receptor complex. Radiat. Prot. Dosim. 122, 106–109.
Sutton, K.A., Black, P.J., Mercer, K.R., Garman, E.F., Owen, R.L., Snell, E.H., Bernhard, W.A., 2013. Insights into the mechanism of X-ray-induced disulfide-bond cleavage in lysozyme crystals based on EPR, optical absorption and X-ray diffraction studies. Acta Cryst. D Biol. Cryst. 69, 2381–2394.
Symons, M.C.R., 1997. Electron movement through proteins and DNA. Free Radic. Biol. Med. 22, 1271–1276.
Théodore, M., Sobczyk, M., Simons, J., 2006. Cleavage of thymine N3–H bonds by low-energy electrons attached to base π^* orbitals. Chem. Phys. 329, 139–147.
Voorhees, R.M., Weixlbaumer, A., Loakes, D., Kelley, A.C., Ramakrishnan, V., 2009. Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome. Nat. Struct. Mol. Biol. 16, 528–533.
Weik, M., Ravelli, R.B., Kryger, G., McSweeney, S., Raves, M.L., Harel, M., Gros, P., Silman, I., Kroon, J., Sussman, J.L., 2000. Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc. Natl. Acad. Sci. USA 97, 623–628.
Yang, M., Chen, Xp, Militello, K., Hoffman, R., Fernandez, B., Baumann, C., Gollnick, P., 1997. Alanine-scanning mutagenesis of *Bacillus Subtilis* trp RNA-binding attenuation protein (TRAP) reveals residues involved in tryptophan binding and RNA binding. J. Mol. Biol. 270, 696–710.
Yano, J., Kern, J., Irrgang, K.D., Latimer, M.J., Bergmann, U., Glatzel, P., Pushkar, Y., Biesiadka, J., Loll, B., Sauer, K., Messinger, J., Zouni, A., Yachandra, V.K., 2005. X-ray damage to the Mn4Ca complex in single crystals of photosystem II: a case study for metalloprotein crystallography. Proc. Natl. Acad. Sci. USA 102, 12047–12052.
Zeldin, O.B., Brockhauser, S., Bremridge, J., Holton, J.M., Garman, E.F., 2013. Predicting the X-ray lifetime of protein crystals. Proc. Natl. Acad. Sci. USA 110, 20551–20556.

Please cite this article as: Bury, C.S., et al., Radiation damage within nucleoprotein complexes studied by macromolecular X-ray crystallography. Radiat. Phys. Chem. (2016), http://dx.doi.org/10.1016/j.radphyschem.2016.05.023

JOURNAL OF SYNCHROTRON RADIATION
ISSN 1600-5775

OH cleavage from tyrosine: debunking a myth

Charles S. Bury,[a] Ian Carmichael[b] and Elspeth F Garman[a]*

[a]Laboratory of Molecular Biophysics, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK, and [b]Notre Dame Radiation Laboratory, University of Notre Dame, Notre Dame, IN 46556, USA. *Correspondence e-mail: elspeth.garman@bioch.ox.ac.uk

Received 15 July 2016
Accepted 19 October 2016

Edited by M. Weik, Université Grenoble Alpes, France

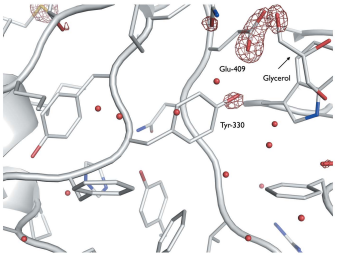
Keywords: tyrosine; specific damage; radiation chemistry; electron density loss; Fourier difference maps.

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

During macromolecular X-ray crystallography experiments, protein crystals held at 100 K have been widely reported to exhibit reproducible bond scission events at doses on the order of several MGy. With the objective to mitigate the impact of radiation damage events on valid structure determination, it is essential to correctly understand the radiation chemistry mechanisms at play. OH-cleavage from tyrosine residues is regularly cited as amongst the most available damage pathways in protein crystals at 100 K, despite a lack of widespread reports of this phenomenon in protein crystal radiation damage studies. Furthermore, no clear mechanism for phenolic C—O bond cleavage in tyrosine has been reported, with the tyrosyl radical known to be relatively robust and long-lived in both aqueous solutions and the solid state. Here, the initial findings of Tyr –OH group damage in a myrosinase protein crystal have been reviewed. Consistent with that study, at increasing doses, clear electron density loss was detectable local to Tyr –OH groups. A systematic investigation performed on a range of protein crystal damage series deposited in the Protein Data Bank has established that Tyr –OH electron density loss is not generally a dominant damage pathway in protein crystals at 100 K. Full Tyr aromatic ring displacement is here proposed to account for instances of observable Tyr –OH electron density loss, with the original myrosinase data shown to be consistent with such a damage model. Systematic analysis of the effects of other environmental factors, including solvent accessibility and proximity to disulfide bonds or hydrogen bond interactions, is also presented. Residues in known active sites showed enhanced sensitivity to radiation-induced disordering, as has previously been reported.

1. Introduction

The effects of radiation damage incurred during data collection continue to impede successful structure determination of proteins by macromolecular X-ray crystallography (MX), even at cryo temperatures (Garman & Weik, 2015). Global damage is manifest in the fading of the diffraction pattern with those reflections pertinent to high-resolution diminishing first, principally as a result of increasing crystal non-isomorphism (Murray & Garman, 2002). Multiple global damage *dose limits* have been proposed as upper bounds for the accumulated dose required to reduce the diffractive power of any macromolecular crystal below a specified fraction (typically 50%). These MX dose limits appear consistent across studies; indeed the originally reported half-dose ($D_{0.5}$ = 20 MGy), theoretically derived by comparison with two-dimensional electron diffraction experiments at 77 K (Henderson, 1990), is on the same order as subsequent, experimentally derived values: $D_{0.5}$ = 43 MGy (Owen *et al.*, 2006) and, for a resolution of s Å, a resolution-dependent dose limit of $D_{0.5} = 10 \times s$ MGy (Howells *et al.*, 2009). An experimental dose limit has been determined as 30 MGy at cryo-temperatures (typically ∼100 K), corresponding to the total recorded intensity



OPEN ACCESS

dropping to 0.7 of its original value, beyond which biological information may well be compromised (Owen *et al.*, 2006). These upper limits appear to be constrained by the associated physics, although chemical effects, so-called specific damage, have been observed at much lower doses, with damage reportedly occurring in a well defined order across a broad range of proteins. Specific damage is typically observed by reconstructing electron density maps, more precisely $F_{obs}(n) - F_{obs}(1)$ difference maps, to monitor changes to electron density as dose increases. Already, at doses of the order of 0.35 MGy, redox-active metal centres are reduced (Corbett *et al.*, 2007) and, as the dose increases, apparent damage, *i.e.* areas of negative difference electron density, have been reported around disulfide bonds, side-chain carboxylic groups on aspartates and glutamates, the OH group on tyrosines and around the C—S bonds of methionines (Burmeister, 2000; Weik *et al.*, 2000; Ravelli & McSweeney, 2000).

Extensive radiation chemical studies in aqueous solution have shown that disulfide bonds are susceptible to both oxidation, for example by hydroxyl radicals, and reduction by the associated solvated electrons. A convincing demonstration that the latter mechanism prevails in lysozyme crystals X-irradiated at cryo-temperatures has been published (De La Mora *et al.*, 2011). In that work, the complementary technique of microspectrophotometry revealed changes in optical absorption as a function of dose, with a peak around 500 nm, associated with radiation-induced trapped electrons, diminishing in concert with the growth of a peak around 400 nm. The latter band could be reliably attributed to the formation of a disulfide radical anion, in which the S—S internuclear distance is substantially increased from that in the neutral disulfide residue, a feature visible in reconstructed electron densities as a function of increasing dose (Weik *et al.*, 2002).

Oxidative decarboxylation of amino acids is also well known from aqueous radiation chemistry, where electron loss in the zwitterionic residue leads to $C_\alpha—CO_2$ bond lengthening at the C-terminus and eventual loss of CO_2 (Wisniowski *et al.*, 2002). In MX, side-chain decarboxylation is also seen in aspartates and glutamates, with quantum chemical calculations indicating that the process is more thermodynamically favourable in the latter (Bury *et al.*, 2016).

Additionally, methionine residues are also readily oxidized under radiation chemical conditions (Kopoldova *et al.*, 1967). The methionine sulfur-centred radical cations which are formed predominantly enter irreversible reaction channels. These ultimately yield carbon-centred and/or peroxyl radicals, which are the starting points for chain reactions of protein oxidation (Schöneich, 2005). Cleavage of the $C_\beta—S$ bond is thus well documented.

Conversely, though the tyrosine residue is well known to be redox active, cleavage of the phenolic C—O bond has not been reported in the proposed mechanisms of tyrosine radiation chemistry. It was thus surprising that the radiation damage study of Burmeister (2000) provided such a clear display of electron density loss, and thus specific radiation damage, around the phenolic oxygen on a tyrosine residue during diffraction data collection from X-irradiated myrosinase crystals (Burmeister, 2000). Indeed the redox couple involving tyrosine and the tyrosyl radical plays a key role in high potential redox catalysis for a wide range of natural enzymes including, for example, photosystem II (Rappaport & Diner, 2008), ribonucleotide reductase (Stubbe *et al.*, 2003), cyctochrome *c* oxidase (Kaila *et al.*, 2010), galactose oxidase (Whittaker, 2003) and prostaglandin synthase (Tsai & Kulmacz, 2010), suggesting a robustness of the phenolic linkage.

Extensive experience from the aqueous radiation chemistry of tyrosine-containing solutions at room temperature also suggests the relative persistence of the tyrosyl radical. For example, from pulse radiolysis studies it is clear that attack by the strongly oxidizing hydroxyl radical, which is formed by the radiolysis of solvent water, leads predominantly to substitution on the aromatic ring. In the presence of oxygen, the eventual products are 2,4- and 3,4-dihydroxyphenylalanines (Chrysochoos, 1968; Solar *et al.*, 1984). Minor pathways involve ring–ring dimerization and H-abstraction from the β-carbon (Solar *et al.*, 1984). Other, less potent, one-electron oxidants initially lead to the formation of the tyrosyl radical, opening pathways leading to products which have recently been thoroughly reviewed (Houée-Levin *et al.*, 2015). In the solid state, soft X-ray irradiation of the respective powders showed phenylalanine degrading more rapidly than tyrosine (Zubavichus *et al.*, 2004). Also, γ-radiolysis at low temperature cleaves the former to give a benzyl radical, but again produces the tyrosyl radical in the latter (Fasanella & Gordy, 1969).

The original Burmeister (2000) study reported progressive accumulation of negative $F_{obs}(n) - F_{obs}(1)$ difference density centred on the –OH group at active site residue Tyr-330 in myrosinase, coupled with observable positive difference density in the adjacent solvent region. In addition, a decaying –OH group occupancy with increasing dose was reported. Myrosinase functions to hydrolyze glucosinolates predominantly in the *Brassicaceae* family, with Tyr-330 situated in a glucose binding site and exhibiting a hydrogen bond to an acidic residue (Glu-409) in the absence of bound glucose (Burmeister *et al.*, 1997). Tyr-330 is predicted to reorientate to accommodate glucose insertion, with the Tyr hydroxyl group then positioned adjacent to the sugar ring oxygen (with no possibility to form a hydrogen bond to Glu-409). In the Burmeister (2000) study, an ordered glycerol molecule instead occupies the glucose-binding site, and Tyr-330 initially exhibits the hydrogen bond to Glu-409. Glu-409 was observed to undergo radiation-induced decarboxylation, leading Burmeister to postulate the existence of a correlation between Tyr –OH damage and such hydrogen bonding interactions. The loss of the Tyr-330-stabilizing hydrogen bond would permit increased lability of the Tyr-330 aromatic ring, and would manifest itself in $F_{obs}(n) - F_{obs}(1)$ maps mostly clearly at positions furthest from the Tyr C_β hinge-point (*i.e.* at the Tyr –OH group). Burmeister also suggested the possibility that an –OH cleavage event had occurred [legend Fig. 3(*c*), Burmeister (2000)].

However, despite the ambiguity in the reason behind Burmeister's finding of density loss around the tyrosyl oxygen,

and the weight of evidence from radiation chemistry studies for tyrosyl persistence, such Tyr –OH electron density loss has subsequently been interpreted as OH loss for protein crystals at 100 K. Tyr –OH cleavage is now regularly cited [including by some of the current authors, see for example Garman (2010); Dubnovitsky *et al.* (2005); Owen & Sherrell (2016)] as one of the four dominant non-metalloprotein specific damage events present in MX experiments conducted at 100 K, despite there being no reported mechanism for phenolic C—O bond cleavage. Conversely, other than for myrosinase, to our knowledge there have been no other reports of clear Tyr –OH damage for protein crystal systems irradiated at 100 K.

To better characterize the reproducibility of Tyr –OH group electron density loss in irradiated protein crystals, we have performed a systematic investigation on a wide range of structures at 100 K. A subset of previously conducted MX radiation damage investigations have included deposition of readily accessible coordinate and structure factor information in the Protein Data Bank (PDB), providing an ideal database of protein damage series. Each damage series has been refined independently by different research groups, mitigating the risk of systematic and subjective errors in the refinements between different structures. Moreover, the deposited damage series have all *a priori* satisfied the PDB validation criteria, and can thus be deemed of suitable quality for the current study.

For the analysis of multiple MX damage series for distinct protein structures, a systematic visual interpretation of $F_{\rm obs}(n) - F_{\rm obs}(1)$ maps would be impractical, with a high risk of subjective bias by the investigators. The $D_{\rm loss}$(atom) metric provides an objective dose-dependent atom-by-atom quantification of electron density loss. $D_{\rm loss}$ is derived from $F_{\rm obs}(n) - F_{\rm obs}(1)$ difference maps at each subsequent accumulated dose ($n = 2, \ldots$) as the maximum density loss value in the local vicinity of each atom [for full construction see Bury *et al.* (2016)]. Here, $F_{\rm obs}(n) - F_{\rm obs}(1)$ difference maps are calculated from the differences of reflection amplitudes between the first and each successive data set and with a fixed phase set, $\varphi_{\rm calc}(1)$, derived from model refinement against the first data set. Our $D_{\rm loss}$ metric was originally introduced to enable analysis of the distribution of specific damage sites within a large (∼204 kDa) bacterial protein-RNA complex (Bury *et al.*, 2016).

This paper reports the results of an investigation into tyrosine electron density loss in the available damage series using this $D_{\rm loss}$ metric.

2. Materials and methods

2.1. Reanalysis of the myrosinase damage series (Burmeister, 2000)

To assess the quality of the originally deposited myrosinase damage series (Burmeister, 2000), the coordinate and `.mmcif` structure factor files were retrieved from the PDB (accession codes: 1dwa, 1dwf, 1dwg, 1dwh, 1dwi, 1dwj). For the original publication, two coordinate models (PDB: 1dwi and 1dwj) had been deposited for the highest dose state, corresponding to occupancy and positional refinement, respectively. The 1dwi coordinate model was used for the current analysis. For comparison, coordinate and structure factor files were extracted from *PDB_REDO* (Joosten *et al.*, 2014), both after initial file curation [ten cycles of rigid body refinement in *REFMAC* (Murshudov *et al.*, 2011)] and after full model re-building. Overall atomic fits to electron density [real-space $Z_{\rm obs}$ scores and real-space correlation coefficient (RSCC)] were computed using *EDSTATS* (Tickle, 2012).

For the initial coordinate model (PDB: 1dwa) of the Burmeister (2000) damage series retrieved from *PDB_REDO* following full model re-building, a secondary Tyr-330 side-chain conformation was built manually in *COOT* (Emsley *et al.*, 2010). Positive difference density in the $F_{\rm obs}(4) - F_{\rm obs}(1)$ map nearby the Tyr-330 aromatic ring (contoured at $\pm 5\sigma$) was used as a guide for suitable placement. To assess the conformational variability of Tyr-330 with dose, for each dataset within this damage series, occupancy refinement was performed in *phenix.refine* (Adams *et al.*, 2010) with atomic *B*-factors held constant, using this modified coordinate model coupled with the structure factor amplitudes for that dataset. The presence of unmodelled alternate Tyr side-chain conformations at each dose was detected using *RINGER* (Lang *et al.*, 2010). *Bdamage* (Gerstel *et al.*, 2015) was calculated for myrosinase as a quantifier for radiation-induced flexibility of Tyr aromatic side groups. Prior to this calculation, to account for the dependence of *Bdamage* on non-unity atomic occupancies and reported atomic *B*-factor values, all occupancies in each myrosinase coordinate model were set to 1 and a further round of isotropic *B*-factor refinement was performed in *phenix.refine*.

2.2. *PDB_REDO* damage series analysis

A complete (to our knowledge) list of MX damage series with both coordinate models and structure factor files deposited in the PDB was collated. Of these, a selection was made using criteria designed to ensure a suitable sample set for the current analysis: (*a*) only MX experiments explicitly conducted at 100 K were included, (*b*) studies with reported accumulated doses <1 MGy (at highest dose dataset) were omitted altogether, and (*c*) intermediate datasets with reported total doses <1 MGy within the remaining series were excluded. A 1 MGy lower limit was imposed since the first signs of specific damage to metal-free protein crystals at 100 K (disulfide radicalization and cleavage) have been detected in $F_{\rm obs}(n) - F_{\rm obs}(1)$ maps below 1 MGy, whereas Glu/Asp decarboxylation has not (Sutton *et al.*, 2013). Datasets (apart from the first one) with doses below 1 MGy were thus deemed unsuitable for the purpose of investigating possible Tyr –OH damage events. The complete Sutton *et al.* (2013) lysozyme damage series (reported dose range 0.07–1.05 MGy) was additionally included as a control in order to justify the chosen 1 MGy lower bound. Table 1 provides a full list of the 16 damage series located in the PDB that were used in the current study.

Table 1
Selection of MX damage series (all conducted at 100 K) previously deposited in the PDB, complemented with one novel GH7 family protein damage series (deposited in conjunction with this study) and one unpublished insulin data set. See §2 of the supporting information for data collection and refinement of the GH7 series.

For series in which intermediate datasets fall below the 1 MGy selection criterion, PDB accession codes are only provided for datasets included in the current analysis.

Protein	PDB accession codes	Resolution for first dataset (Å)	Quoted total molecular weight (kDa)	Unique Tyr residues in asymmetrical unit	Publication
Acetylcholinesterase (*Torpedo californica*)	1qid, 1qie, 1qif, 1qig, 1qih, 1qii, 1qij, 1qik, 1qim	2.05	60.7	17	Weik *et al.* (2000)
Thermolysin (*Bacillus thermoproteolyticus*)	3p7p, 3p7q, 3p7r, 3p7s, 3p7t, 3p7u, 3p7v, 3p7w	2.2	34.6	28	Juers & Weik (2011)
Myrosinase (*Sinapis alba)*	1dwa, 1dwf, 1dwg, 1dwh	2.0	62.4	37	Burmeister (2000)
Malate dehydrogenase (*Haloarcula marismortui*)	2j5k, 2j5q, 2j5r	2.0	131.6	32	Fioravanti *et al.* (2007)
Chicken egg white lysozyme (*Gallus gallus*)	2ybh, 2ybi, 2ybj, 2ybl, 2ybm, 2ybn	2.0	14.6	3	De la Mora *et al.* (2011)
Chicken egg white lysozyme (*Gallus gallus*)	4h8x, 4h8y, 4h8z, 4h90, 4h91, 4h92, 4h93, 4h94, 4h9a, 4h9b, 4h9c, 4h9e, 4h9f, 4h9h, 4h9i	1.2	14.5	3	Sutton *et al.* (2013)
Pancreatic elastase (*Sus scrofa)*	3mnb, 3mnc, 3mns, 3mnx, 3mo3, 3mo6, 3mo9, 3moc	1.2	26.1	11	Petrova *et al.* (2010)
Pancreatic elastase (*Sus scrofa)*	2blo, 2blq	1.33	26.2	11	Nanao *et al.* (2005)
Insulin (*Bos taurus)*	2bn3, 2bn1	1.4	5.8	4	Nanao *et al.* (2005)
Lysozyme (*Gallus gallus)*	2blx, 2bly	1.4	14.7	3	Nanao *et al.* (2005)
Ribonuclease A (*Bos taurus)*	2blp, 2blz	1.4	13.7	6	Nanao *et al.* (2005)
Thaumatin (*Thaumatococcus daniellii)*	2blr, 2blu	1.4	22.3	8	Nanao *et al.* (2005)
Trypsin (*Bos taurus)*	2blv, 2blw	1.2	24.0	10	Nanao *et al.* (2005)
C-protein-DNA complex (*Escherichia coli*)	4x4b, 4x4c, 4x4d, 4x4e, 4x4f, 4x4g, 4x4h, 4x4i	2.8	59.7	8	Bury *et al.* (2015)
Trp RNA-binding attenuation protein (TRAP) -RNA complex (*Escherichia coli*)	5eeu, 5eev, 5eew, 5eex, 5eey, 5eez, 5ef0, 5ef1, 5ef2, 5ef3	1.98	204.1	22	Bury *et al.* (2016)
Phosphoserine aminotransferase *(Bacillus alcalophilus)*	2bhx, 2bi5, 2bi9, 2bia	1.68	81.8	28	Dubnovitsky *et al.* (2005)
Insulin (*Bos taurus)*	Undeposited damage series	1.38	5.8	4	Current publication
GH7 family cellobiohydrolase *(Daphnia pulex)*	5mcc, 5mcd, 5mce, 5mcf, 5mch, 5mci, 5mcj, 5mck, 5mcl, 5mcm, 5mcn	2.0	96.6	34	Unpublished

For each MX damage dataset series, PDB-format coordinate model and MTZ-format structure factor files were retrieved from the *PDB_REDO* online server. For each damage series, files were retrieved after initial *PDB_REDO* file curation, but prior to any further model rebuilding. At this stage, coordinate models had been subjected to either 0 or 10 cycles of rigid body refinement with *REFMAC*, dependent on whether *PDB_REDO* flagged possible $R_{\rm free}$ bias within the PDB-deposited data. This procedure was chosen to extract data as close to that used by the authors of the associated radiation damage studies as possible, whilst ensuring that all data were in a consistent format. Accordingly, for the initial structure in each damage series, with the exception of myrosinase, the recalculated $R_{\rm work}$ statistics were all reported by *PDB_REDO* to be within 2% of those reported in the originally deposited coordinate file header (and for structures where there was no *PDB_REDO* flagged potential $R_{\rm free}$ bias, $R_{\rm free}$ was also within 1% of the original reported values), see Table S1.1 of the supporting information for details. For the myrosinase damage series (Burmeister, 2000), the coordinate model following full model rebuilding with *PDB_REDO* was used in the current analysis, since the $2mF_{\rm obs} - DF_{\rm calc}$ map-model fit was poor for the original PDB-deposited 1dwa model (see §3.1 for details).

Further rigid body refinement was performed in *REFMAC* for each higher dose dataset, now using the coordinate model from the initial dataset coupled with the structure factor amplitudes corresponding to the higher dose dataset. This ensured that an identical set of atoms was modelled for each dataset within a damage series, whilst accounting for whole-molecule translation and rotation within the unit cell, as a result of any changes in unit-cell dimensions with increasing dose. The resulting coordinate models were used throughout the analysis presented here.

It must be emphasized that the reported maximum attained resolution differs between each protein system (from 1.2 to 2.8 Å). The optimum model refinement strategy (*e.g.* applicability of NCS, *B*-factor restraints) is very much dependent on resolution, and it is recognized that the selected proteins have undergone non-uniform refinement protocols prior to retrieval. Consequently, a *B*-factor-derived metric [*e.g.* relative atomic *B*-factor increase (Weik *et al.*, 2000) or *Bdamage*

(Gerstel *et al.*, 2015)] was deemed unsuitable for comparison between damage series. For each MX damage series, a per-atom quantification, $D_{\rm loss}$(atom), of the radiation-induced electron density loss derived from $F_{\rm obs}(n) - F_{\rm obs}(1)$ Fourier difference maps (n = 2,3, ...), was computed at each dose using the program *RIDL* (available upon request from the authors). A metric derived from $F_{\rm obs}(n) - F_{\rm obs}(1)$ maps depends directly on experimentally determined diffraction intensities, thus minimizing dependence on refined models. $F_{\rm obs}(n) - F_{\rm obs}(1)$ maps were generated using *FFT* (Ten Eyck, 1973) as part of the *RIDL* pipeline. Initial manual inspection of maps was performed in *COOT* to ensure that key specific damage manifestations (*e.g.* clear negative difference density surrounding disulfide bonds) were consistent with reports in the original publications pertaining to the particular damage series. Per-atom solvent accessibilities were calculated with the CCP4 program *AREAIMOL*, accounting for crystal contacts between adjacent asymmetric units, and hydrogen bonding interactions within structures were determined using *NCONT* (Winn *et al.*, 2011).

3. Results

3.1. Reanalysis of the myrosinase damage series (Burmeister, 2000)

For the first dataset in the myrosinase series (PDB accession code: 1dwa), *PDB_REDO* reported poor agreement between $R_{\rm work}/R_{\rm free}$ given by the .pdb header originally deposited in the PDB (0.165/0.179) and those recalculated directly from the deposited data post rigid body refinement but prior to model rebuilding (0.191/0.201). Visual inspection of the $2mF_{\rm obs} - DF_{\rm calc}$ map retrieved from the Electron Density Server (EDS) revealed regions of poor fits of this model for ordered solvent and side-chain conformations to density (overall map RSCC: 0.67), including several Tyr aromatic groups (notably Tyr-215). To recalculate $R_{\rm work}/R_{\rm free}$ from the original 1dwa data, *PDB_REDO* carried out a default 0-cycle *REFMAC* refinement, followed by 10-cycles of rigid body refinement (standard procedure in *PDB_REDO* when the $R_{\rm work}/R_{\rm free}$ reported in the original .pdb header are not reproducible to within 5% after the single 0-cycle *REFMAC* run). The 10-cycle rigid body refinement of 1dwa was observed to improve the fit of Tyr side-chains to density (with overall map RSCC: 0.920). Corresponding real-space EDSTATS $Z_{\rm obs}$ scores both pre- and post- further rigid body refinement are provided in Table S1.2 of the supporting information. Further full model rebuilding in *PDB_REDO* again improved the refinement statistics ($R_{\rm work}/R_{\rm free}$: 0.13/0.16 with RSCC: 0.97). The improvement in the Tyr-215 electron density fit is shown in Fig. S1.1 of the supporting information.

Consistent with Burmeister's original findings, interpretable negative difference density in $F_{\rm obs}(n) - F_{\rm obs}(1)$ maps located at the Tyr-330 –OH group was observed above (*i.e.* more negative) a -5σ map threshold for the highest dose datasets (n = 4, 5) (Fig. 1). An additional negative difference peak was coincident with the Tyr-215 –OH group (also clear above -5σ for n = 4, 5), a finding not reported in the original study (Fig. S1.2). We suggest that in that work the poor density fit for Tyr-215 resulted in no difference peak being attributed to Tyr-215. Indeed, upon our re-inspection of the $F_{\rm obs}(5) - F_{\rm obs}(1)$ map from the original data extracted from the PDB, a nearby difference peak was revealed; however, it was not directly aligned with the Tyr-215 –OH group (Fig. S1.2a). Significant density loss peaks (apparent at $> -4\sigma$) were also identifiable in the $F_{\rm obs}(5) - F_{\rm obs}(1)$ map centred around several other Tyr –OH groups (Tyr-390 above -5σ, and Tyr-400, Tyr-181 and Tyr-152 above -4.5σ). Both the relative atomic *B*-factor increase and *Bdamage* atomic displacement metrics exhibited strong agreement with these observations, with Tyr-330, Tyr-215 and Tyr-390 flagged as highly sensitive compared with other Tyr –OH groups (Fig. 2). Here the *Bdamage* metric provides a quantification of increased residue flexibility with dose, deconvoluted from atomic packing density.

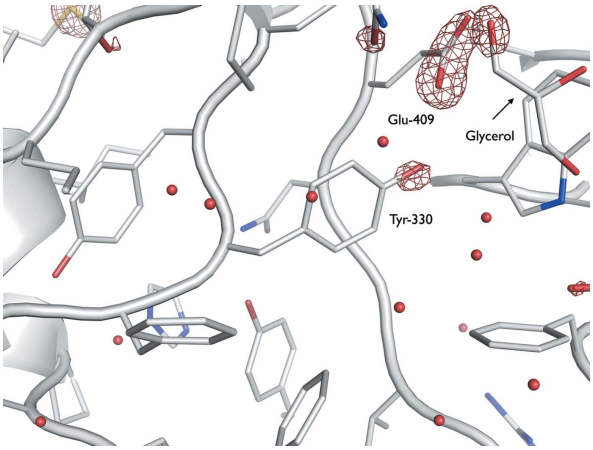


Figure 1
Specific radiation damage to Tyr-330 and Glu-409 in the myrosinase damage series. $F_{\rm obs}(5) - F_{\rm obs}(1)$ difference map, contoured at $\pm 5\sigma$ in green/red, overlaid on the initial coordinate model in white (PDB accession code: 1dwa). Coordinate model and structure factor amplitudes were retrieved from *PDB_REDO*, and $F_{\rm obs}(5) - F_{\rm obs}(1)$ maps were generated by *FFT* within the *RIDL* pipeline. Electron density loss is observable around a glycerol oxygen, presumably also due to the dislocation of the glycerol upon decarboxylation of Glu-409.

Overall, although *PDB_REDO* provided a significant improvement in $R_{\rm work}/R_{\rm free}$ statistics for the original initial myrosinase structure 1dwa, the damage signatures reported by Burmeister (2000) could be replicated.

3.2. Investigation of radiation-induced trends between proteins

In the publications relating to previous radiation damage investigations, dose values have been independently calculated using a variety of tools [*e.g.* *RADDOSE* v1 (Murray *et al.*, 2004), v2 (Paithankar *et al.*, 2009) and v3 (Paithankar & Garman, 2010), and *RADDOSE-3D* (Zeldin *et al.*, 2013*a*)]. As such, direct comparison of raw $D_{\rm loss}$ values between protein

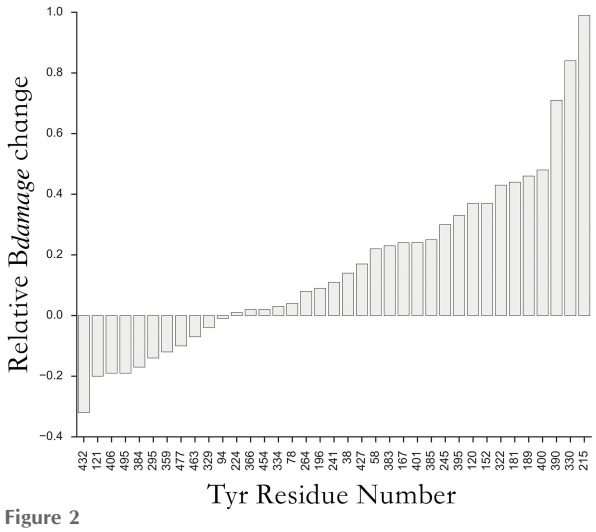


Figure 2
Bdamage change for each Tyr –O atom in the highest dose myrosinase structure (PDB accession code: 1dwi) relative to the 1dwa structure. For each atom a, the relative change is calculated as $[Bdamage(a^{1dwi}) - Bdamage(a^{1dwa})]/Bdamage(a^{1dwa})$. To account for non-unity occupancies in the originally deposited data, all atomic occupancies in the 1dwa and 1dwi coordinate models were set to 1 and a further round of isotropic B-factor refinement was performed in *phenix.refine* prior to calculating *Bdamage*.

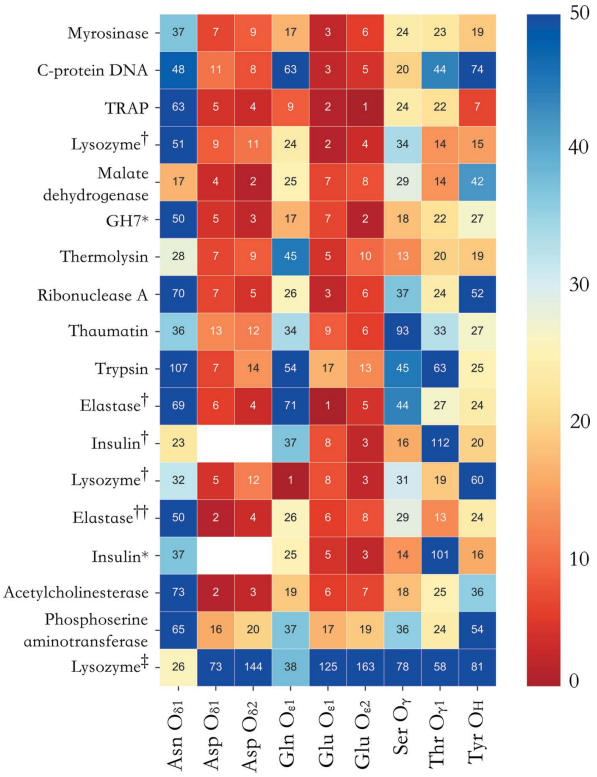


Figure 3
Density loss ranking for each damage series for distinct side-chain oxygen atom types occurring within each protein structure. Blank squares represent atom types not present within a given protein. The ranking scheme starts at 0 for the most damaged atom species within each protein, thus a large ranking number indicates low D_{loss}. For clarity, each square is coloured by rank, with the corresponding colour key provided to the right of the grid. Superscripted proteins correspond to the following studies: †Nanao *et al.* (2005), ††Petrova *et al.* (2010), ‡Sutton *et al.* (2013). *Data collected from GH7 and insulin crystals (distinct from *PDB_REDO* collated series) are additionally included as part of the test set (see Table 1 for details).

crystals *reported* to be at identical doses is highly sensitive to miscalibration between dose scales. No attempt has been made here to recalculate doses for the damage series, and all doses are as reported within the original publications. A discrete ranking scheme has been implemented to mitigate uncertainties when comparing density map-derived metric values between independent protein damage series, since experimental procedures and refinement strategies are also not conserved amongst them.

For a damage series consisting of a single low- and high-dose dataset, D_{loss} values correspond to the single $F_{obs}(2) - F_{obs}(1)$ map, and each atom type A (*e.g.* the set of Tyr –OH oxygen atoms) is ranked with respect to the mean D_{loss} values attained by that atom type (rank = $R_A \geq 0$). For damage series containing multiple higher dose datasets, atom type ranks have been determined per-dataset ($R_{A,i}$ for $i = 2, \ldots, n$) and the products of these per-dataset ranks ($R_{A,2} \times R_{A,3} \times \ldots \times R_{A,n}$) are then ranked to provide an across-dataset rank for each atom type.

Fig. 3 illustrates the resulting ranks of radiation-induced electron density loss for different side-chain oxygen moieties in the damage series investigated. Only the oxygen atoms in the structures have been compared with Tyr –OH oxygen, to maintain a constant total number of electrons per compared atom. In a protein at a specific dose, neglecting variations in thermal motions between different side-chains, two oxygen atoms O_A and O_B should exhibit identical D_{loss}(atom) values if the yields of O_A and O_B cleavage events throughout the protein crystal are the same. Whereas clear Glu and Asp decarboxylations are indicated by uniformly high Glu-$O_{\varepsilon 1}$/$O_{\varepsilon 2}$ and Asp-$O_{\delta 1}$/$O_{\delta 2}$ ranks, Tyr –O is consistently ranked lower, and of a similar order as seen in the control group of Gln, Asn, Ser and Thr O atoms for each protein structure, which are not suspected to be cleaved. Indeed in 14/18 damage series, at least one out of the control group has a rank higher than that of Tyr –O. In particular, for the myrosinase series (Burmeister, 2000), Gln-$O_{\varepsilon 1}$ receives a higher ranking than Tyr –O.

In phosphoserine aminotransferase, Glu and Asp are ranked as less susceptible than typically observed, with the highest rankings instead attained by atoms within the lysine-pyridoxal-50-phosphate Schiff base present within the protein (data not shown). Indeed, the Schiff base active site was reported in the original publication (Dubnovitsky *et al.*, 2005) to undergo severe radiation-induced relaxation of chemical strain, with pronounced reorientation of the Schiff base cofactor. The poor rankings (>73, Fig. 3) for Glu and Asp in lysozyme (Sutton *et al.*, 2013) are consistent with the low maximum dose (1.05 MGy) reported for this damage series.

Whereas disulfide bond breakage is reported in this study, the dose is below the onset of Glu/Asp decarboxylation for this protein.

Overall it is evident that despite unavoidable variability between the included protein systems (*e.g.* differing maximum resolution, crystallization buffers), the implemented $D_{\rm loss}$ ranking scheme is capable of detecting specific damage trends throughout the investigated proteins. Regardless of the feasibility of a radiation-induced Tyr –OH damage event within crystalline protein at 100 K, Tyr –OH density loss is in general significantly lower than for Glu and Asp carboxylates for doses >1 MGy, at which decarboxylation events were reported.

3.3. A hypothesis for the observed Tyr electron density loss

Here the distribution of $D_{\rm loss}$ values (so-called *damage signatures*) for topologically similar residue types have been directly compared. These are shown in Fig. 4 for the highest dose datasets for myrosinase, Trp RNA-binding attenuation protein (TRAP), malate dehydrogenase (Fioravanti *et al.*, 2007) and the GH7 protein, all of which contain a statistically sufficient sample size of atoms for histogram analysis (each >62 kDa). For each residue, the damage signature plot contains $D_{\rm loss}$ information for all non-hydrogen atoms of that residue at a specific dose. In Fig. 4, at each dose, $D_{\rm loss}$ values have been normalized to the average $D_{\rm loss}$ value attained at that particular dose by the set of C_α backbone atoms in the structure. This quantifies per-atom density loss relative to a known radiation-insensitive set of atoms. As a result, signatures are observed to be typically unimodal distributions, with undamaged atoms attaining near zero values and right-tails indicating the presence of radiation-sensitive atom types within a residue. For each residue pairing, the Kolmogorov–Smirnov (KS) test has been employed as a non-parametric statistic to test the null hypothesis that the $D_{\rm loss}$ values of both residue types have been drawn from the same distribution. The KS-test statistic thus provides a quantifier for the similarity between the $D_{\rm loss}$ signatures for two residue types (Fig. 5).

For each protein, Asp and Asn exhibited qualitatively different behaviour, with the extended right-hand tail of the Asp distribution indicative of radiation-induced carboxylate loss, in contrast to the radiation-insensitive amine functional group of Asn (Figs. 4*b*, 4*d*, 4*f*, 4*h*). Conversely, Phe and Tyr exhibited qualitatively indistinguishable $D_{\rm loss}$ signatures (Figs. 4*a*, 4*c*, 4*e*, 4*g*). Ile and Leu also showed qualitatively indistinguishable $D_{\rm loss}$ behaviour (Fig. S1.3), as expected, since both have similar aliphatic side-chain topologies that are

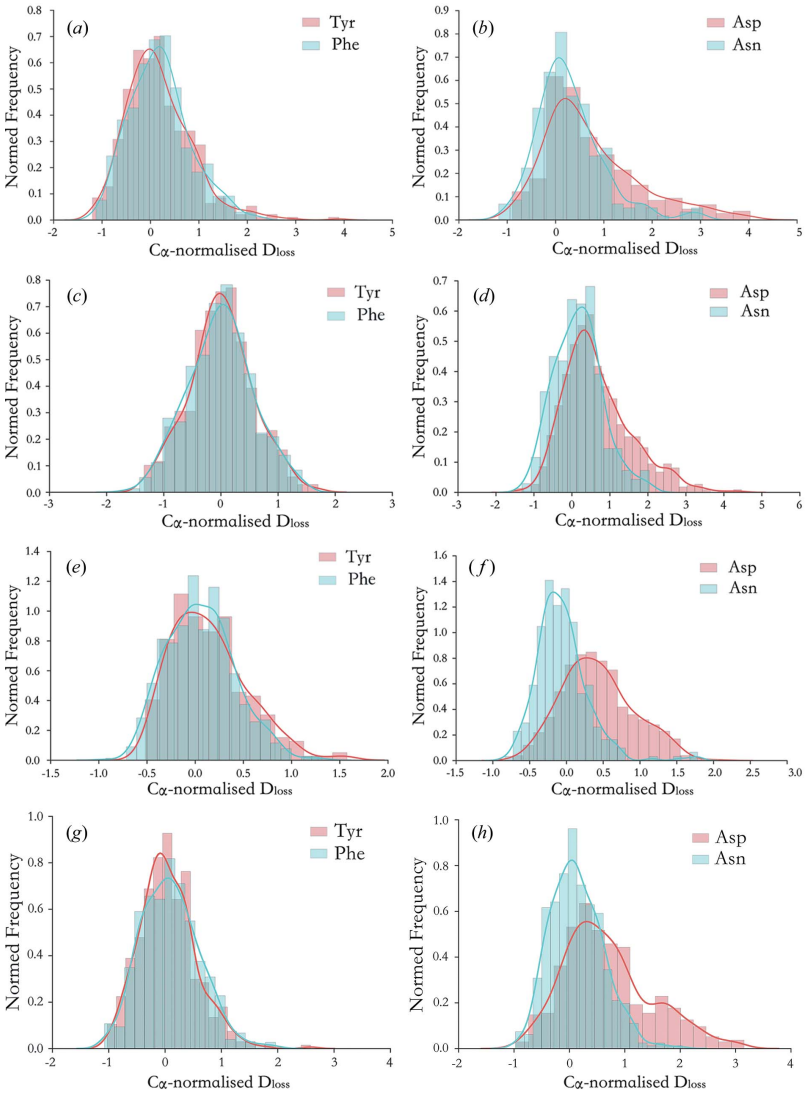


Figure 4
C_α-normalized $D_{\rm loss}$ damage signature histogram plots over all atoms of specific residue types for (*a*, *b*) myrosinase, (*c*, *d*) malate dehydrogenase, (*e*, *f*) TRAP and (*g*, *h*) GH7. The normalization scheme is: C_α-normalized $D_{\rm loss}$(atom) = $[D_{\rm loss}({\rm atom}) - \mathbb{E}_{C_\alpha}(D_{\rm loss})]/|\mathbb{E}_{C_\alpha}(D_{\rm loss})|$, where $\mathbb{E}_{C_\alpha}(D_{\rm loss})$ is the average $D_{\rm loss}$ calculated over the set of C_α atoms. Two residue types are directly compared in each plot: (*a*, *c*, *e*, *g*) Tyr and Phe, and (*b*, *d*, *f*, *h*) Asp and Asn. Gaussian kernel density estimates, which are non-parametric approximations of the probability density function underlying each $D_{\rm loss}$ distribution, are overlaid on the histogram plots. Highest dose datasets for each series are displayed; other doses within each series exhibit qualitatively similar behaviour.

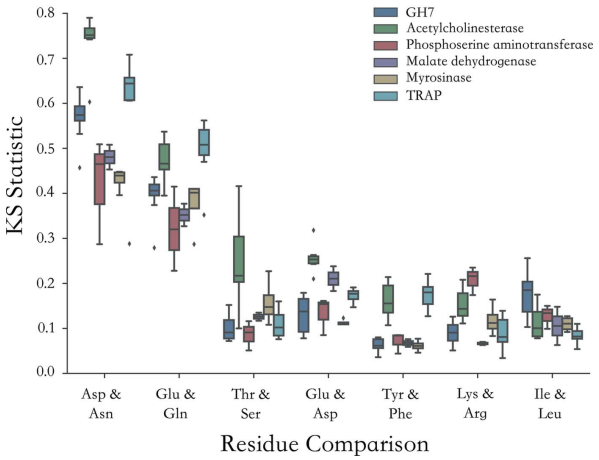


Figure 5
Kolmogorov–Smirnov (KS) test statistics comparing similarity of different residue types for large protein structures (>60 kDa). For each damage series, box plots illustrate the distribution in KS statistic values across all the datasets in that series. A higher KS statistic indicates less similarity between the distribution of $D_{\rm loss}$ values for a specified residue pairing.

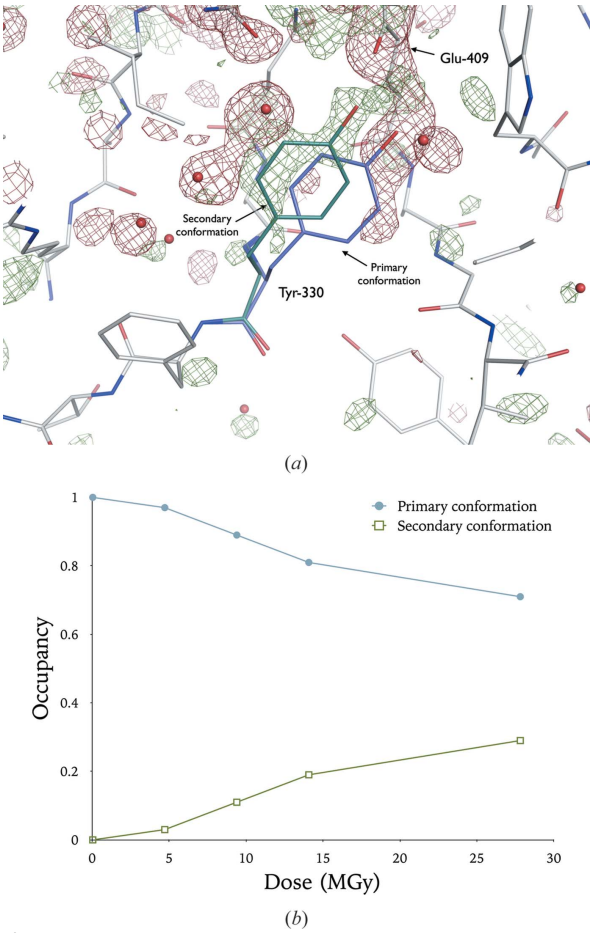


Figure 6
(*a*) Specific radiation damage to the myrosinase crystal as reported by Burmeister (2000). $F_{\rm obs}(4) - F_{\rm obs}(1)$ difference map, contoured at $\pm 3\sigma$, overlaid on the initial coordinate model (PDB accession code: 1dwa). A secondary Tyr-330 side-chain conformation (occupancy: 50%) has been built into the positive $F_{\rm obs}(4) - F_{\rm obs}(1)$ difference density manually using *COOT*. (*b*) Occupancy of primary and secondary Tyr-330 conformation after occupancy refinement (in *phenix.refine*, with atomic *B*-factors held constant) of the model pertaining to each successive dataset in the damage series. Doses were originally quoted in units of photons mm^{-2}; diffraction weighted doses (DWD) (Zeldin, Brockhauser *et al.*, 2013) have been calculated in *RADDOSE-3D* (Zeldin, Gerstel *et al.*, 2013) for this damage series using crystal composition (heavy atom content, crystal size) and beam characteristics (energy, flux, exposure time and collimation) as supplied by Burmeister (2000).

believed to be relatively radiation-insensitive in protein crystals at 100 K, even at doses of several MGy.

As can be seen in Fig. 5, the Tyr–Phe KS-test statistic was consistently low and of the same order as the Leu–Ile control across the range of investigated proteins. The radiation effects to Phe and Tyr were thus deemed to be highly similar in protein crystals at 100 K in the MGy dose regime, with comparable spatial displacement of the aromatic rings. It is also noted that Thr and Ser exhibit comparable $D_{\rm loss}$ distributions, indicating that the presence of an additional methyl in Thr does not affect the damage to these groups.

To confirm whether the Tyr –OH density loss observations in myrosinase were consistent with full aromatic ring displacement, a secondary Tyr-330 aromatic ring conformation (initially with 50% occupancy) was built into the 1dwa coordinate model. With increasing dose, there is an observed shift in the relative occupancies of the two Tyr-330 conformations (Fig. 6). The growth in secondary conformation occupancy is consistent with partial reorientation of the Tyr-330 ring with dose. However, interestingly, despite the clear presence of a secondary conformation within the $F_{\rm obs}(4) - F_{\rm obs}(1)$ map at a low σ-level (Fig. 6*a*), for all doses within the myrosinase damage series, the formation of an unmodelled alternative Tyr-330 conformation was not detected using *RINGER* at low $2mF_{\rm obs} - DF_{\rm calc}$ map σ-thresholds (above a default $+0.3\sigma$ cutoff).

3.4. Factors affecting the Tyr –OH group

To examine the dependence of Tyr –OH $D_{\rm loss}$ on availability of proximal hydrogen bond interactions to acidic residues, for each structure, Tyr residues were grouped into two subsets: those predicted to interact with Glu/Asp carboxylates at distances under 4 Å (between the Tyr –OH oxygen and the carboxylate oxygens), and all other Tyr residues. Fig. 7 illustrates the results for two selected proteins, with the other systems presented in Fig. S1.4. Seven damage series contained a distinct subset of Tyr –OH atoms with $D_{\rm loss}$ exceeding that attained by a group of non-hydrogen bond-forming residues, whereas for six series no difference was observed between the two Tyr subgroups. In one protein [elastase (Petrova *et al.*, 2010)], no Tyr residues were predicted to form such hydrogen bonds. Only for two structures were the lowest $D_{\rm loss}$ values assigned to a subset of hydrogen bonded Tyr –OH groups [lysozyme (De La Mora *et al.*, 2011) and insulin (Nanao *et al.*,

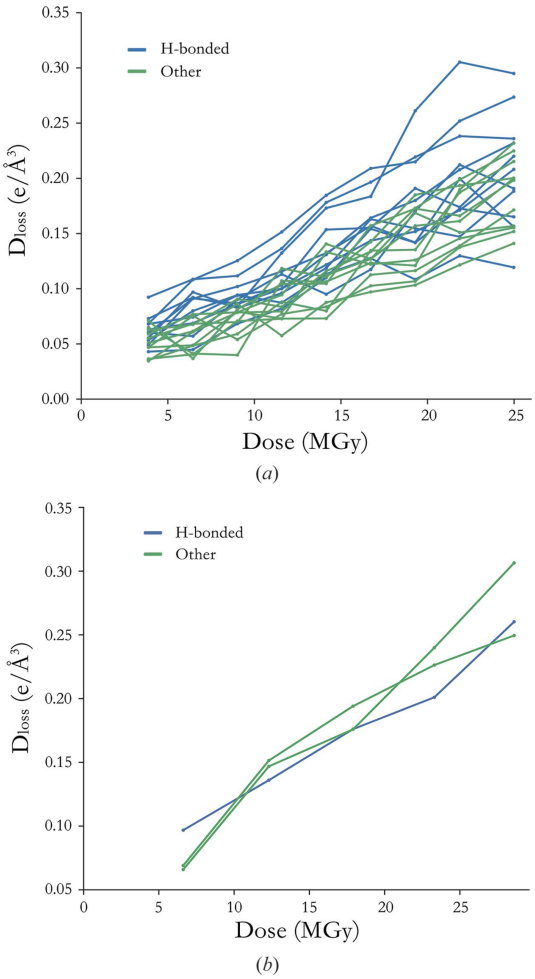


Figure 7
D_{loss}(atom) metric as a function of dose for Tyr –OH atoms in (*a*) TRAP and (*b*) lysozyme (De la Mora *et al.*, 2011). Tyr –OH residues exhibiting hydrogen bond interactions to Glu or Asp carboxyl groups are coloured blue. Plots for other tested protein series are provided in Fig. S1.4.

2005)]. However, these were sampled over three and four Tyr groups, respectively, and are thus not considered statistically significant. Consequently, whereas Tyr –OH hydrogen bonding interactions with proximal Glu/Asp residues promoted radiation-induced Tyr –OH electron density loss in a subset of tested proteins (in strong support of prior anecdotal evidence), this was not uniformly reported for all proteins. Moreover, only for myrosinase (Burmeister, 2000) and thermolysin (Juers & Weik, 2011) were the collective group of hydrogen bond-forming Tyr –OH atoms statistically distinguishable from the remaining Tyr –OH groups (at α = 5% significance, Hotelling T^2 test).

The loss of a direct Tyr–Glu/Asp hydrogen bond may not in itself be sufficient for increased Tyr ring flexibility, and undoubtedly other factors will influence the Tyr local environment: the identity of the remaining radicalized Glu/Asp species post decarboxylation, the proximity of additional hydrogen bond acceptors, and aromatic ring stacking interactions. Currently the relative importance of these factors is challenging to deconvolute and quantitate. To parametrize the reduction in Tyr steric hindrance following Glu/Asp decarboxylation, the solvent accessible area for each Tyr aromatic ring was calculated, after first removing all Glu and Asp carboxylates from each protein coordinate model. Poor correlation was observed between Tyr –OH D_{loss} and the resulting solvent accessibility for each damage series (Fig. 8*a*). It has been suggested that disulfide bond cleavage may induce changes to local protein backbone conformation and Tyr side-chain orientation, but, for the subset of disulfide containing proteins (Fig. 8*b*), no correlation was detected between the Tyr –OH D_{loss} metric and the distance (Å) to disulfide bonds.

4. Discussion

The characterization of radiation-mediated cleavage events in protein crystals remains an essential component of structure solution by crystallographic methods, with undetected specific damage currently being particularly detrimental to the success of modern time-resolved investigations aiming to unravel protein kinetics. The correct interpretation of electron density around each Tyr –OH moiety is often vital to ensure valid descriptions of the intramolecular interactions underpinning protein function. Tyrosine can play a characteristic role in signal transduction pathways, in which signalling specificity is dependent on direct enzyme-mediated group substitution (acylation, phosphorylation, sulfation) of the –OH functional group. In metalloproteins, tyrosine has also frequently been reported in the coordination sphere of the bound metal, with the Tyr phenolic oxygen directly acting as a donor to the metal cofactor.

The large body of radiation chemistry studies on both aqueous and solid state protein samples has been consulted to suggest feasible mechanisms for tyrosine-related damage events to crystalline proteins held at 100 K. Low-energy solvated electrons (~1–10 eV) and positive holes are thought to be the dominant reactive species able to migrate at 100 K (Jones *et al.*, 1987), with less reactive hydrogen atoms also predicted to be mobile but produced in far lower yield than electrons upon water radiolysis (Spinks & Woods, 1990). The hydroxyl radical, known to be highly reactive from room-temperature protein and nucleic acid aqueous studies, is thought to be immobilized within the bulk solvent constituents of the crystal at 100 K (Owen *et al.*, 2012). However, for Tyr –OH removal, no such electron- or hole-driven mechanisms have been presented in the radiation chemistry literature, and the tyrosyl radical produced following deprotonation of the oxidized residue is deemed to form a stable redox couple with tyrosine. Indeed, in a neutral disulfide, the typical S—S bond strength is ~55 kcal mol^{-1} (Benson, 1978). However, in the reduced species (the radical anion) that strength is lowered to ~23 kcal mol^{-1} [derived from our quantum chemical calculations, as performed by Bury *et al.* (2016)]. By way of contrast in a phenol (such as tyrosine), we calculate that the C—O

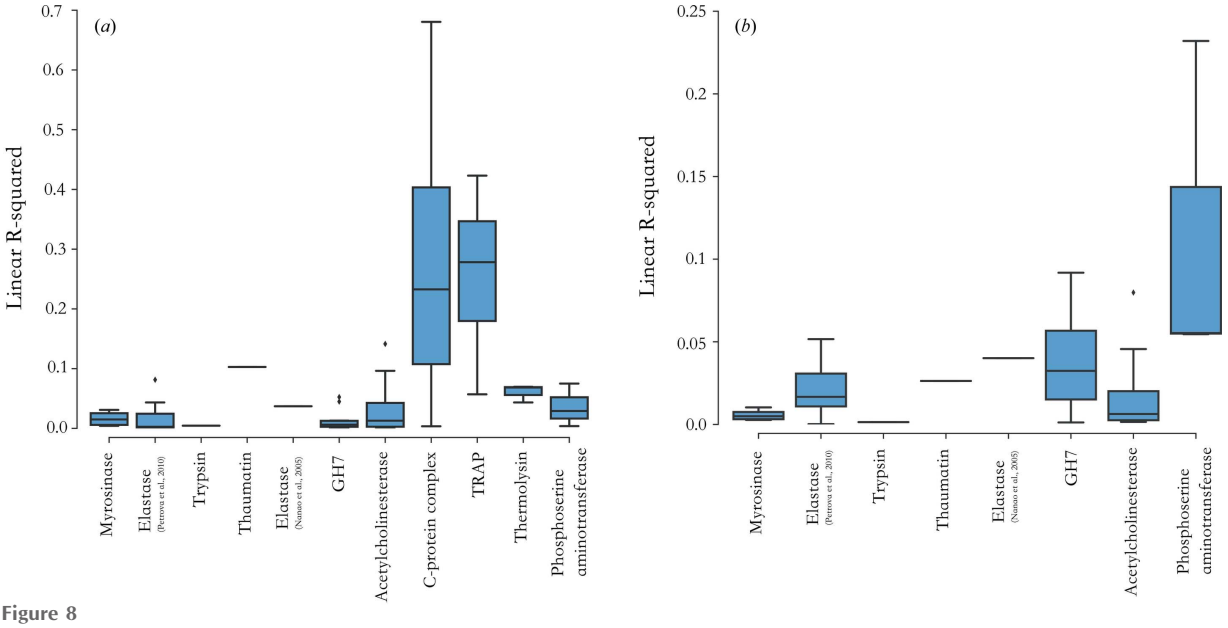


Figure 8
Linear *R*-squared coefficients calculated for each MX damage series for correlation between Tyr –OH D_{loss} and (*a*) solvent accessibility and (*b*) distance to closest disulfide bond. Box plots represent distribution in *R*-squared values for series containing multiple datasets. Only structures containing eight or more Tyr residues per asymmetric unit are included, and in (*b*) only structures containing disulfide bonds are included.

bond strength is approximately 110 kcal mol^{-1} and it is not significantly different in the oxidized species.

Re-investigation of the original myrosinase MX damage series data independently reproduced the observations of Tyr-330 –OH group electron density loss in $F_{obs}(n) - F_{obs}(1)$ difference maps. In the Burmeister damage study, rigid body and overall anisotropic *B*-factor refinement (but no positional refinement) was performed starting from a 1.2 Å myrosinase model derived from a different crystal used for another experiment (Burmeister *et al.*, 2000). We suggest that this refinement strategy may account for the low density map-model RSCC calculated for the original data (0.67). However, whereas full re-refinement and model re-building with *PDB_REDO* resulted in a significant increase in RSCC, negative difference peaks surrounding the Tyr-330 –OH group were still detectable above noise in $F_{obs}(n) - F_{obs}(1)$ maps. *PDB_REDO* re-refinement also provided an improved fit to $2mF_{obs} - DF_{calc}$ map density for several Tyr aromatic rings (*e.g.* Tyr-215), and a pronounced negative difference density (comparable with that at Tyr-330) became interpretable around the Tyr-215 –OH group; a finding not originally reported by Burmeister (2000). Tyr-215 and Tyr-330 occupy similar hydrated pockets in myrosinase, with both being involved in hydrogen bond interactions (to Glu-153 and Glu-409, respectively), and both being within 5 Å of an ordered glycerol molecule. As such, the similar –OH electron density loss for Tyr-215 and Tyr-330 is perhaps not surprising.

The D_{loss} per-atom quantification of radiation-induced electron density loss has provided an invaluable tool to permit objective investigation of Tyr –OH density loss events in further protein damage series collected at 100 K. The available MX damage series deposited in the PDB gave a wide range of independently refined coordinate models corresponding to proteins of varying size (5.7–204 kDa), each with unique nuances of buffer composition (potential free radical scavengers) and macromolecular composition (monomeric or multimeric protein, presence of nucleic acid, glycans and metal cofactors). No restriction on resolution or crystallization cocktail composition has been enforced in the current sampling, and, accordingly, no attempt has been made to quantitatively compare raw D_{loss} values between proteins. However, as a testament to the strength of the D_{loss} metric, clear trends have emerged, with radiation-induced Glu and Asp decarboxylation highlighted to be prevalent across the range of protein systems in our study at doses >1 MGy. Prominent Tyr –OH density loss did not appear to be a widespread feature of general protein structures, a finding that is consistent with the high energetic barrier for Tyr phenolic C—O bond scission. It is thus arguable whether tyrosine should be included with Cys, Glu, Met and Asp as a radiation-sensitive locus.

The evidence suggests that (*a*) the Tyr and Phe aromatic rings undergo comparable positional displacement with increasing dose, and (*b*) Tyr –OH electron density loss is correlated with adjacent Tyr-C_ζ density loss [for structures in which high Tyr –OH D_{loss}(atom) values have been reported here: myrosinase and TRAP with linear R^2 at 0.72 and 0.58 for the respective highest dose datasets, see §3 of the supporting information]. Overall these observations support a model of full aromatic ring displacement without the necessity of bond cleavage. Indeed, through occupancy refinement of multiple Tyr side-chain conformations, the original Burmeister (2000)

results (Tyr-330) have been shown to be highly consistent with this postulate, since as the occupancy of the original conformation decreases, the population of alternative positions increases (Fig. 6*b*).

There has been anecdotal evidence to suggest a relationship between radiation-induced Tyr –OH electron density disordering and the presence of hydrogen bond interactions to Glu/Asp carboxyl side-groups. Indeed, in TRAP (Bury *et al.*, 2016), a macromolecule with 11-fold rotational self-symmetry, 11 symmetrically related Tyr copies all hydrogen bond to nearby Glu residues, and all exhibit $F_{obs}(n) - F_{obs}(1)$ map peaks that are distinguishable above noise.

In the current study, whereas select Tyr residues forming hydrogen bond interactions to proximal Glu/Asp carboxylates exhibited comparably high D_{loss} metric values (relative to non hydrogen bond-forming Tyr –OH groups), this behaviour was restricted to a subset of proteins (myrosinase, TRAP, thermolysin, insulin), and was not conserved amongst all such Tyr residues for these proteins. We suggest that a multitude of compounding factors associated with a hydrogen bonding interaction must play a role in the density loss rate for each individual Tyr –OH group (*e.g.* readiness of Glu/Asp decarboxylation, lability of the Tyr aromatic group, steric hindrance both prior and post decarboxylation). However, no quantitative correlation could be established for these individual factors. For example, considering the two most susceptible Tyr –OH groups in myrosinase (Burmeister, 2000), following loss of interacting Glu $–CO_2$ moieties, a negligible change in overall aromatic ring solvent accessible area (used as a measure for steric hindrance) was detected for Tyr-330, whereas an increase of 1.36 Å^2 was observed for Tyr-215.

Nonetheless, the top three Tyr residues in myrosinase flagged by the *Bdamage* metric (Tyr-215, Tyr-330, Tyr-390) do all exhibit hydrogen bonding interactions to glutamates. We suggest that the high relative *Bdamage* change values for the active site residues Tyr-215 and Tyr-330 are a direct indication of relaxation of the strained active site with accumulated dose. This finding is in close accord with the consensus that active sites are particularly prone to specific damage events (Holton, 2009; Dubnovitsky *et al.*, 2005). For the full set of myrosinase tyrosine –OH oxygen atoms, a strong positive correlation was found between D_{loss} and change in *Bdamage* between the first and higher dose datasets (Fig. S1.5, linear $R^2 > 0.7$ for all higher dose datasets). Overall, the *potential flexibility* of each Tyr appears to directly correlate with electron density loss around the –OH group. The radiation-induced conformational shift for Tyr-330 is pronounced since this residue is purposefully labile in order to reorientate to accept glucose into the binding pocket. However, such a tyrosine conformational shift is an indirect effect of nearby radiation chemistry (*e.g.* disulfide bond cleavage, decarboxylation), and thus should not be considered a feature exclusive to tyrosines.

5. Related literature

The following references, not cited in the main body of the paper, have been cited in the supporting information: Chen *et al.* (2010); Fuentes-Montero *et al.* (2014); McCoy *et al.* (2007); Southworth-Davies *et al.* (2007).

Acknowledgements

We thank John McGeehan for kindly supplying us with the GH7-family protein crystal and Jonathan Brooks-Bartlett for giving us access to his complete insulin damage series data (prior to deposition). We gratefully acknowledge the UK Engineering and Physical Sciences Research Council for studentship funding in the Systems Biology Programme of the University of Oxford Doctoral Training Centre (CSB). IC is supported by the US Department of Energy Office of Science, Office of Basic Energy Sciences under Award Number DE-FC02-04ER15533.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst.* D**66**, 213–221.
Benson, S. W. (1978). *Chem. Rev.* **78**, 23–35.
Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.
Burmeister, W. P., Cottaz, S., Driguez, H., Iori, R., Palmieri, S. & Henrissat, B. (1997). *Structure*, **5**, 663–676.
Burmeister, W. P., Cottaz, S., Rollin, P., Vasella, A. & Henrissat, B. (2000). *J. Biol. Chem.* **275**, 39385–39393.
Bury, C. S., Garman, E. F., Ginn, H. M., Ravelli, R. B. G., Carmichael, I., Kneale, G. & McGeehan, J. E. (2015). *J. Synchrotron Rad.* **22**, 213–224.
Bury, C. S., McGeehan, J. E., Antson, A. A., Carmichael, I., Gerstel, M., Shevtsov, M. B. & Garman, E. F. (2016). *Acta Cryst.* D**72**, 648–657.
Chen, V. B., Arendall, W. B. 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D , 12–21.
Chrysochoos, J. (1968). *Radiat. Res.* **33**, 465–479.
Corbett, M. C., Latimer, M. J., Poulos, T. L., Sevrioukova, I. F., Hodgson, K. O. & Hedman, B. (2007). *Acta Cryst.* D**63**, 951–960.
De la Mora, E., Carmichael, I. & Garman, E. F. (2011). *J. Synchrotron Rad.* **18**, 346–357.
Dubnovitsky, A. P., Ravelli, R. B., Popov, A. N. & Papageorgiou, A. C. (2005). *Protein Sci.* **14**, 1498–1507.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Fasanella, E. I. & Gordy, W. (1969). *Proc. Natl Acad. Sci. USA*, **62**, 299–304.
Fioravanti, E., Vellieux, F. M. D., Amara, P., Madern, D. & Weik, M. (2007). *J. Synchrotron Rad.* **14**, 84–91.
Fuentes-Montero, L., Parkhurst, J., Winter, G., Waterman, D., Gildea, R., Brewster, A., Hattne, J., Sauter, N. & Evans, G. (2014). *Acta Cryst.* A**70**, C1440.
Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.
Garman, E. F. & Weik, M. (2015). *J. Synchrotron Rad.* **22**, 195–200.
Gerstel, M., Deane, C. M. & Garman, E. F. (2015). *J. Synchrotron Rad.* **22**, 201–212.
Henderson, R. (1990). *Proc. Biol. Sci.* **241**, 6–8.
Holton, J. M. (2009). *J. Synchrotron Rad.* **16**, 133–142.
Houée-Levin, C., Bobrowski, K., Horakova, L., Karademir, B., Schöneich, C., Davies, M. J. & Spickett, C. M. (2015). *Free Radical Res.* **49**, 347–373.
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. & Starodub, D. (2009). *J. Electron Spectrosc. Relat. Phenom.* **170**, 4–12.
Jones, G. D., Lea, J. S., Symons, M. C. & Taiwo, F. A. (1987). *Nature (London)*, **330**, 772–773.
Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). *IUCrJ*, **1**, 213–220.
Juers, D. H. & Weik, M. (2011). *J. Synchrotron Rad.* **18**, 329–337.
Kaila, V. R. I., Verkhovsky, M. I. & Wikström, M. (2010). *Chem. Rev.* **110**, 7062–7081.
Kopoldová, J., Liebster, J., Gross, E. & Kopoldova, J. (1967). *Radiat. Res.* **30**, 261–274.
Lang, P. T., Ng, H. L., Fraser, J. S., Corn, J. E., Echols, N., Sales, M., Holton, J. M. & Alber, T. (2010). *Protein Sci.* **19**, 1420–1431.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Murray, J. & Garman, E. (2002). *J. Synchrotron Rad.* **9**, 347–354.
Murray, J. W., Garman, E. F. & Ravelli, R. B. G. (2004). *J. Appl. Cryst.* **37**, 513–522.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. (2005). *Acta Cryst.* D**61**, 1227–1237.
Owen, R. L., Axford, D., Nettleship, J. E., Owens, R. J., Robinson, J. I., Morgan, A. W., Doré, A. S., Lebon, G., Tate, C. G., Fry, E. E., Ren, J., Stuart, D. I. & Evans, G. (2012). *Acta Cryst.* D**68**, 810–818.
Owen, R. L., Rudino-Pinera, E. & Garman, E. F. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 4912–4917.
Owen, R. L. & Sherrell, D. A. (2016). *Acta Cryst.* D**72**, 388–394.
Paithankar, K. S. & Garman, E. F. (2010). *Acta Cryst.* D**66**, 381–388.
Paithankar, K. S., Owen, R. L. & Garman, E. F. (2009). *J. Synchrotron Rad.* **16**, 152–162.
Petrova, T., Ginell, S., Mitschler, A., Kim, Y., Lunin, V. Y., Joachimiak, G., Cousido-Siah, A., Hazemann, I., Podjarny, A., Lazarski, K. & Joachimiak, A. (2010). *Acta Cryst.* D**66**, 1075–1091.
Rappaport, F. & Diner, B. A. (2008). *Coord. Chem. Rev.* **252**, 259–272.
Ravelli, R. B. & McSweeney, S. M. (2000). *Structure*, **8**, 315–328.
Schöneich, C. (2005). *Biochim. Biophys. Acta*, **1703**, 111–119.
Solar, S., Solar, W. & Getoff, N. (1984). *J. Phys. Chem.* **88**, 2091–2095.
Southworth-Davies, R. J., Medina, M. A., Carmichael, I. & Garman, E. F. (2007). *Structure*, **15**, 1531–1541.
Spinks, J. W. T. & Woods, R. J. (1990). *Introduction to Radiation Chemistry*, 3rd ed. New York: Wiley.
Stubbe, J., Nocera, D. G., Yee, C. S. & Chang, M. C. Y. (2003). *Chem. Rev.* **103**, 2167–2202.
Sutton, K. A., Black, P. J., Mercer, K. R., Garman, E. F., Owen, R. L., Snell, E. H. & Bernhard, W. A. (2013). *Acta Cryst.* D**69**, 2381–2394.
Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.
Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.
Tsai, A. L. & Kulmacz, R. J. (2010). *Arch. Biochem. Biophys.* **493**, 103–124.
Weik, M., Bergès, J., Raves, M. L., Gros, P., McSweeney, S., Silman, I., Sussman, J. L., Houée-Levin, C. & Ravelli, R. B. G. (2002). *J. Synchrotron Rad.* **9**, 342–346.
Weik, M., Ravelli, R. B., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 623–628.
Whittaker, J. W. (2003). *Chem. Rev.* **103**, 2347–2364.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Cryst.* D**67**, 235–242.
Wisniowski, P., Carmichael, I., Fessenden, R. W. & Hug, G. L. (2002). *J. Phys. Chem. A*, **106**, 4573–4580.
Zeldin, O. B., Brockhauser, S., Bremridge, J., Holton, J. M. & Garman, E. F. (2013). *Proc. Natl. Acad. Sci. USA*, **110**, 20551–20556.
Zeldin, O. B., Gerstel, M. & Garman, E. F. (2013). *J. Appl. Cryst.* **46**, 1225–1230.
Zubavichus, Y., Zharnikov, M., Shaporenko, A., Fuchs, O., Weinhardt, L., Heske, C., Umbach, E., Denlinger, J. D. & Grunze, M. (2004). *J. Phys. Chem. A*, **108**, 4557–4565.