Radiation Damage in Protein Crystallography: Susceptibility Study

Markus Gerstel

St Hilda's College

University of Oxford

A thesis submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy

Michaelmas 2014

Abstract

Protein structure models obtained from X-ray crystallography are subject to radiation damage. The resulting specific alterations to protein structures can be mistaken for biological features, or may obscure actual protein mechanisms, leading to misidentification or obscuration of biological insight. The radiation chemistry behind this site-specific damage is not well understood.

Radiation damage processes progress in proportion to the dose absorbed by the crystal in the diffraction experiment. Doses can be estimated using existing software, but these assume idealised experimental conditions.

To simulate complex diffraction experiments, including treatment of imperfect X-ray beam profiles and inhomogeneous dose distributions, a new program, RADDOSE-3D, was developed. RADDOSE-3D can be integrated into beamline software to provide convenient, more accurate, comparative, and publishable dose figures, also facilitating informed data collection decisions.

There is currently no method to automatically detect specific radiation damage in protein structure models in the absence of an 'undamaged' reference model. Radiation damage research therefore generally relies on detailed observation of a few model proteins.

A new metric, B_{Damage}, is designed and used to identify and quantify specific radiation damage in the first large-scale statistical survey of 2,704 published protein models, which are examined for the effects of local environments on site-specific radiation damage susceptibility. A significant positive correlation between susceptibility and solvent accessibility is identified.

Current understanding of radiation damage progression is mostly based on a few consecutive structure model 'snapshots' at coarse dose intervals. The low sampling rate considerably limits the ability to identify varying site susceptibility and its causes.

Real space electron density data are obtained for crystals of different mutants of a RhoGDI protein with very high sequence identity, to determine sensitising and stabilising factors for radiation induced structural changes. Utilising a newly developed data collection and analysis protocol, these changes could be tracked with unprecedented time resolution.

‘standing on the shoulders of giants’

Sir Isaac Newton

Acknowledgements

I gratefully acknowledge the UK's Engineering and Physical Sciences Research Council (EPSRC) for funding through a studentship, grant number EP/G03706X/1 at the Systems Biology programme at the University of Oxford's Doctoral Training Centre. Further funding was provided by the Eighth International Workshop on X-ray Damage to Biological Crystalline Samples, the Keith Prout Crystallography Fund, the Kojo Minta Memorial Fund, the National Research Council Canada, the Oxford University Department of Biochemistry Conference Fund, and the RCUK Open Access Block Grant to Oxford, which enabled me to present and publish my work.

I also thank Zygmunt S. Derewenda and his group, in particular Natalya Olekhnovich, for providing samples of three RhoGDI protein mutants, on which a central part of this study is based.

None of this work would have been possible without the guidance and support of my wonderful principal supervisor *extraordinaire*, Elspeth F. Garman. I enjoyed the company and help of our research group, Areej Abuhammad, Jonathan Brooks-Bartlett, Katharina Jungnickel, Charles Bury, and in particular Oliver Zeldin, who helped me get my foot in the door. Special thanks go to Edward Lowe for his help in the hour of greatest need.

I would also like to thank my examiners Jonathan Grimes and Martin Weik for a once-in-a-lifetime harrowing and wonderful experience, and my thesis committee and confirmation examiners Simon Newstead and Sylvia McLain as well as my transfer assessors Paul Emsley and Robert Esnouf for their comments and feedback.

This academic endeavour could not have succeeded without the inspiration, stability and support of my family, Christa, Georg, Michael, Sabine, Christoph and Lisa-Marie Gerstel. The main credit for preserving my sanity over these years has to go to Gefion Thuermer. You walked this entire journey with me, and it fills me with joy that I may now reciprocate and accompany you throughout your journey to your PhD.

Finally, I would like to thank my former Oxford MSc supervisors, Stephen Clark, now at Cambridge, and Jotun Hein, who supported my application and, together with the Oxford University Society of Change Ringers, are responsible for me returning to Oxford in the first place. And all the others who helped me on my journey in one way or the other: Elizabeth Allan · Rebecca Batters · Constanze Berg · Andrew Bissette · Joachim Bokor Anthony Bradley · John Bremridge · Johanna Burch · Tom Burnley · Armin van Buuren Young Choi · Aimee Collis · Lorraine Damerell · Charlotte Deane · Thomas Dixon Jean-Paul Ebejer · Hannah Edwards · Anselm Fliedner · Helen Ginn · Georg Hochberg Philip Hogg · Peter Holmes · James Holton · Aris Katzourakis · Olga Kuznetsova Lara Lämke · Eoin Malins · Samantha Miles · Himadri Mukhopadhyay · Marek Obermeier Nuno Oliveira · Arwen Pearson · Klaus Peukert · Andreas Popp · Gail Preston Raimond Ravelli · Julia Reda MEP · Paul Roberts · Claudia Schmidt · Veronique Schmitz Kerstin Seier · Miriam Seyffarth · Oliver Starkey · Florian Stascheck · Christoph Steltner Francesca Summers · Isaac Turner · Manfred Weiss · Adrian von Ziegler Alexander Zinser · Florian Zumkeller-Quast ...and everyone I forgot.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Radiation damage

Introduction: Radiation damage

X-ray crystallography is the current gold standard for determining the three-dimensional structures of macromolecules. By shining an intense X-ray beam on crystals of protein, protein complexes, or viruses, the limitations of optical microscopes can be circumvented, and the otherwise hidden information about their inner structure can be obtained (Bragg, 1913b).

Much of our current understanding of protein function is derived from these structure models, and their importance to the future of not just the field of structural biology, but also to that of other fields such as medicine, can hardly be overstated.

X-ray crystallography is limited by the radiation sensitivity of crystals. These are inherently fragile objects, which can disintegrate in the highly intense X-ray beams necessary for structure determination. Radiation damage in crystals may cause difficulty in obtaining a structure model at all, and limits the quality of the resulting models, but perhaps more importantly it may also compromise the biological fidelity of these models. Obviously, only reliable structure models are useful. An understanding of radiation damage mechanisms is therefore imperative, so that informed decisions regarding the design of X-ray diffraction experiments and the most efficient use of the available crystals can be made.

The effects of radiation damage in protein crystallography and the susceptibility of protein crystals to radiation damage will be explored in this study.

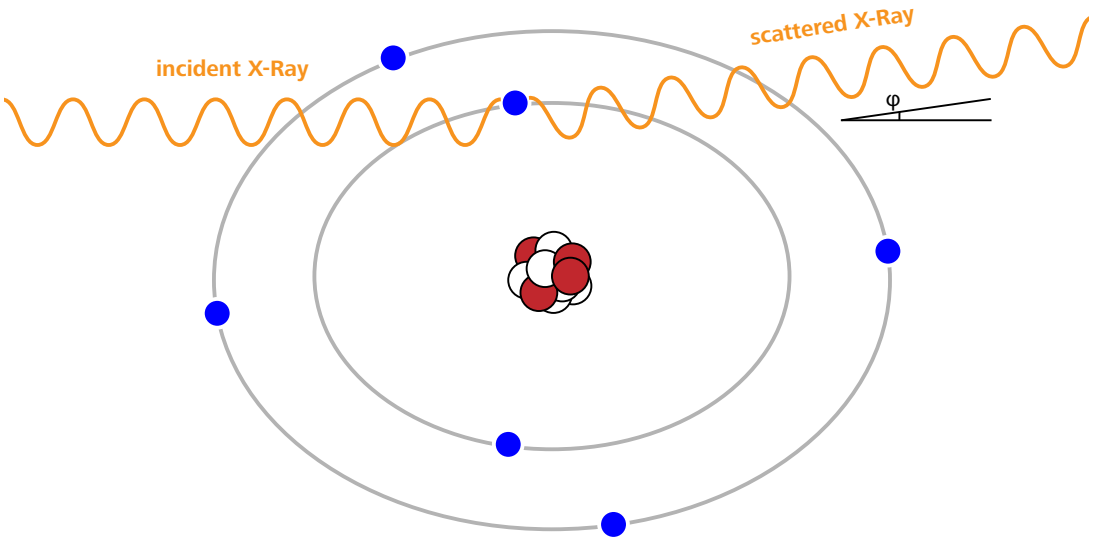


Figure 1.1: X-ray–matter interaction: elastic (Rayleigh) scattering
An incident X-ray is elastically scattered by an electron (blue) orbiting an atom's nucleus, shown as red (protons) and white (neutrons) spheres in the centre. The X-ray energy (and thus: wavelength) is conserved. The atom is not affected. The apparent scattering angle is shown here as φ. The law of reflection states that the angle of incidence is equal to the angle of reflection, meaning that the actual scattering angle is $\theta = \varphi/2$, which can then be used in Bragg's Law.

1.1 Interactions between X-rays and matter

X-ray diffraction experiments are based on the detection of Rayleigh scattering (Figure 1.1). Incident X-rays are elastically scattered by the electrons orbiting the nucleus within the specimen crystal. These scattered X-rays then give rise to a diffraction pattern through constructive and destructive interference according to Bragg's Law (Bragg, 1913a, as cited in Glazer, 2013):

$$n\lambda = 2d \sin\theta.$$

By collecting a set of these diffraction patterns for different crystal orientations, and combining these with phase information obtained either from other protein structures or by phasing experiments, to construct a three-dimensional model of the electron density distribution and thus

the locations of the atoms within the crystal (phasing is discussed in greater detail in Section 4.1.1, page 130).

Rayleigh scattering does not affect the crystal, since the incident X-ray energy is identical to the emitted X-ray energy and thus no energy is deposited in the sample[1]. There are however two further possible interactions between X-rays and matter[2] which prove to be less benign. These are Compton scattering (Figure 1.2; Compton, 1923) and the photoelectric effect (Figure 1.3; Einstein, 1905). In both cases some or all of the energy of the incident X-ray is absorbed and leads to the ejection of at least one electron at the primary interaction site. These electrons subsequently lose their energy via hundreds of further interactions within the crystal (O'Neill *et al.*, 2002). Blake & Phillips (1962) deduced from experiment that absorption of an 8 keV X-ray can disrupt 70 protein molecules and causes some disorder to a further 90[3]. The primary photoelectron causes a cascade with a range on the order of up to a few micrometres, depending on their initial energy (4 µm at 18.5 keV and 100 K; Sanishvili *et al.*, 2011). Protein crystal unit cells usually have sides smaller than 100 Å (0.01 µm), so this cascade can pass through hundreds of unit cells. The energy deposited in the protein crystal through X-ray irradiation can be expressed as a dose in gray (1 Gy = 1 $^{J}/_{kg}$). While it is impossible to measure dose it can be estimated using dedicated software and a meticulous description of all the details of the diffraction experiment[4].

In general the probability of any X-ray–matter interaction taking place is very low: As few as

[1] This is strictly incorrect: a small fraction of the X-ray energy is taken by the recoil of the atom for the conservation of momentum (Moon, 1960). Practically this has no discernible effect on either the atom or the reflected X-ray.

[2] The fourth theoretically possible interaction, pair production, is only relevant at photon energies above 1.05 MeV, far above the energy ranges relevant for conventional X-ray diffraction experiments. Pair-production may play a role in the extreme environments of diffraction experiments with X-ray Free Electron Lasers (XFEL; Ringwald, 2001). The approach in XFEL experiments is aptly called 'diffract-before-destroy', as the destruction of the sample (after diffraction) is considered part of the experiment (Neutze *et al.*, 2000).

[3] Both numbers assume a low absorbed dose up to 20 Mrad = 0.2 MGy; 1 Gy = 1 $^{J}/_{kg}$.

[4] A list of selected parameters can be found in Tables 3.1–3.3, pages 83–85, and the calculation of doses is discussed at length in Chapter 3.

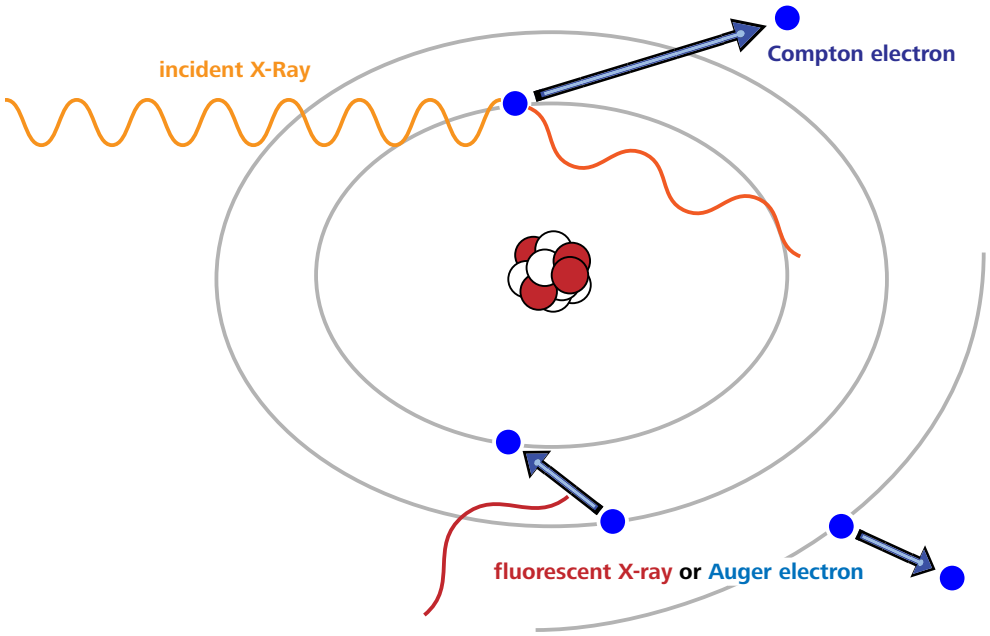


Figure 1.2: X-ray–matter interaction: inelastic (Compton) scattering

An incident X-ray is inelastically scattered by an electron. Part of the energy of the incident X-ray is used to eject an inner-shell electron of the atom, leaving an electron-hole. Due to the law of conservation of energy the partially reflected X-ray has substantially lower energy, and thus a longer wavelength. The atom is left in an unstable excited state. Eventually an outer shell electron drops onto the inner shell to fill the electron-hole. In this process potential energy is released either through the ejection of another outer shell electron (Auger electron) or by the emittance of a fluorescent X-ray.

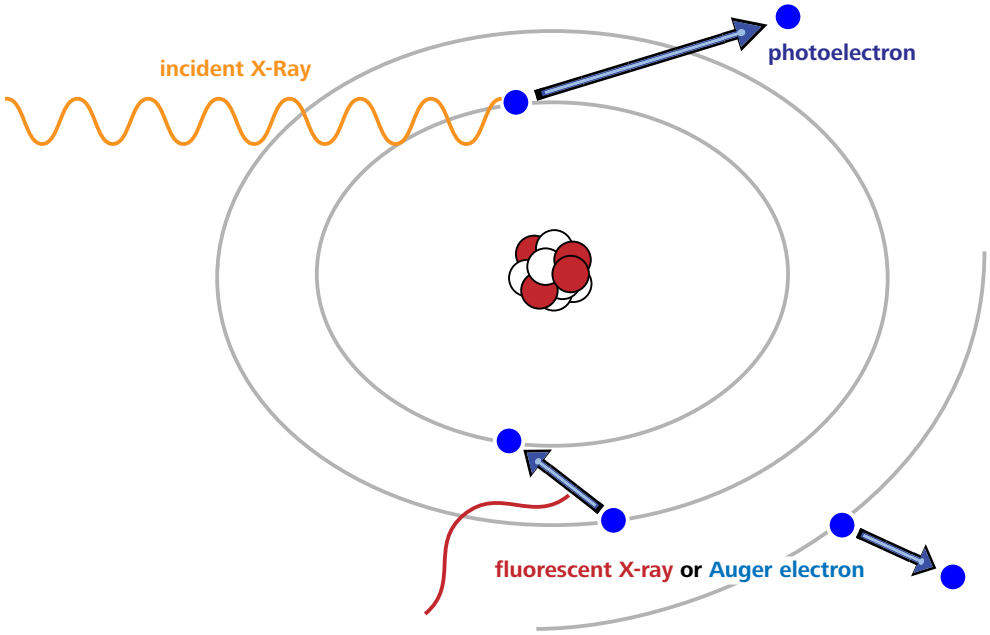


Figure 1.3: X-ray–matter interaction: the photoelectric effect
An incident X-ray is completely absorbed by an electron, which is ejected as a photoelectron. If an inner shell electron is ejected the atom is left in an unstable excited state, and can stabilise via the same decay processes as in the case of Compton scattering (Figure 1.2).

2 % of the incident photons[5] interact with a protein crystal at all. The remaining 98 % simply pass through the crystal and are then absorbed by a beam stop before they can damage the detector. Considering only the interacting photons, the probabilities of each of the three X-ray–matter interactions taking place are purely dependent on the X-ray beam energy and the atomic crystal composition, which define the overall X-ray cross section of the crystal. The effect of the X-ray beam energy can be demonstrated by calculating the X-ray cross-sections for an ordinary Chicken Egg-White Lysozyme (CEWL) protein crystal[6]. In the past, most macromolecular crystallography diffraction experiments used X-rays at the copper K_α edge, 8.04 keV (1.54 Å). At this energy only 3.1 % of the X-ray–matter interactions can be attributed to Rayleigh scattering, see Table 1.1. The vast majority of interactions (96.9 %) do not contribute to the X-ray diffraction experiment,

[5] Assuming a 100 µm thick crystal with no heavy atom content in a 1 Å (12.4 keV) X-ray beam (Garman, 2010).
[6] The crystal composition can be found in Table 1.1.

Cross-section at beam energy	Cu K_α (8.04 keV, 1.54 Å)		Se K (12.7 keV, 0.98 Å)	
Photoelectric effect	1.147 mm^{-1}	(95.5%)	0.278 mm^{-1}	(87.4%)
Compton scattering	0.017 mm^{-1}	(1.4%)	0.020 mm^{-1}	(6.3%)
Rayleigh scattering	0.037 mm^{-1}	(3.1%)	0.020 mm^{-1}	(6.3%)

Table 1.1: X-ray cross-sections of a typical Chicken Egg-White Lysozyme crystal at two different beam energies. Assumed crystal size $100\times100\times100\,\mu m^3$, unit cell a=b=77.3 Å, c=38.2 Å, α=β=γ=90°, 129 protein residues with 8 monomers per unit cell, solvent containing 1.05 M Na and 1 M Cl (composition of PDB ID 2BLX; Nanao *et al.*, 2005). Coefficients calculated with RADDOSE version 2 (Paithankar *et al.*, 2009; Murray *et al.*, 2004).

but do cause structural damage in the crystal. At higher energies (and thus shorter wavelengths) the proportion of useful Rayleigh scattering increases, with approximately twice as much Rayleigh scattering at 12.7 keV compared to 8.04 keV, as shown in Table 1.1. While the photoelectric cross-section decreases with higher X-ray energies the produced photoelectrons have a much higher energy. The diffraction intensity per unit of total absorbed energy thus increases only by approximately 10 % (Arndt, 1984). This small increase prompted a move towards higher energies (Arndt, 1984), and today beam energies near the selenium K edge (12.658 keV, 0.979 Å) have become the standard for data collection at synchrotrons (Fourme *et al.*, 2012; Carugo *et al.*, 2005).

Moving to even higher energies does not result in further improvements: at 24.8 keV (0.5 Å) the diffracted intensity per absorbed energy is increased by ∼2 % compared to 12.7 keV (Arndt, 1984). While relatively more interactions are elastic scattering events, in absolute terms only about 1/5 as many interactions happen, and thus nearly five times as many photons would be needed to result in a diffraction pattern with identical overall intensity. Other considerations, such as the availability of suitable sensitive detectors, also make the use of even higher energy beams currently impractical.

Another way of reducing the number of primary Compton- and photoelectrons produced is to reduce the concentration of heavy elements in the crystal buffer solution (Murray *et al.*, 2005). The pernicious effects of the presence of heavy atoms are now widely recognised: Heavy atoms have a disproportionately large X-ray cross section, yet, unless they are bound to the protein in ordered locations, do not contribute to the diffraction pattern at all. If heavy atoms need to be present in the crystal growth medium it is therefore advisable to back-soak any obtained crystals prior to the diffraction experiment in a solution that either does not contain heavy atoms, or contains them at a lower concentration. Holton (2009, Table 1) gives a useful table showing the dose-doubling concentrations of commonly used heavy elements at the Se edge. For example a 350 mM concentration of arsenic (which is contained in e.g. sodium cacodylate, a widely used crystallisation buffer) will halve the useful crystal lifetime in terms of dose.

1.2 Attempts to mitigate secondary radiation damage

The primary radiation damage events outlined above are a fundamental part of the X-ray diffraction experiment, and can not be avoided. There is however substantial research on the reduction of secondary radiation damage to improve the quality of protein structure models.

1.2.1 Changes of experimental temperature

By flash-cooling crystals in liquid (77 K) or gaseous (100 K) nitrogen and holding the crystal in a stream of gaseous nitrogen at 100 K throughout the diffraction experiment (Garman & Schneider, 1997), the diffusion of radiation products, such as hydroxyl or larger radicals is effectively stopped (Owen *et al.*, 2012). While overall the crystal is still damaged, the observed damage in the ordered

parts of the protein crystal is significantly reduced (Juers & Weik, 2011). At cryo-temperatures crystals can last up to 70 times as long as crystals at room temperature before significant effects of radiation damage become apparent (Nave & Garman, 2005). Experiments at even lower temperatures by Meents *et al.* (2010) reported a further reduction of radiation damage at 50 K, however Warkentin *et al.* (2012b) found no significant benefit in terms of radiation damage at an experimental temperature of 25 K compared to 100 K. Considering the associated installation and running cost of the required helium cryocooling equipment and the still questionable benefit, it is hardly surprising that 100 K is currently the preferred temperature for cryo-temperature X-ray diffraction experiments.

1.2.2 Addition of protective small molecule compounds

The introduction of small molecule radioprotectant and scavenging compounds has been proposed and tested. In theory, these compounds may sweep up radiation by-products, such as hydroxyl radicals or solvated electrons, and prevent them from damaging the protein in the crystal. At room temperature 1,4-benzoquinone and ascorbate have been found to be effective scavengers, providing, respectively, a ninefold and double dose tolerance in CEWL crystals (Barker *et al.*, 2009). At cryo-temperatures (100 K) a doubling of dose tolerance has been reported for 0.5 M sodium nitrate in CEWL crystals as measured by global parameters, and a more than fivefold dose tolerance for damage to disulphide bonds (De la Mora *et al.*, 2011). The extent of the protecting effect of sodium nitrate is disputed, as another study found a doubled dose tolerance at room temperature, but no effect at cryo-temperature (Kmetko *et al.*, 2011). It would appear that the varying results in the literature are at least to some extent caused by the metrics used to judge radiation damage (Allan *et al.*, 2013; where a comprehensive review of the literature on this subject can be

found). Given that any effect within a factor of two can be ascribed to variability in the quality of protein crystals of the same type (Nowak *et al.*, 2009; Pozharski, 2012; Holton, 2009) it is still unclear whether electron scavengers may ever become generally useful in macromolecular crystallography.

1.2.3 Variation in X-ray beam energy / wavelength

Studies of radiation damage at 100 K showed radiation damage to be independent of the X-ray energy, and only to be dependent on the absorbed dose. Damage was judged by crystallographic dataset statistics and by inspection of visible damage at disulphide bridges and methionine residues for a range of X-ray beam energies between 6.5 keV and 33.0 keV (Shimizu *et al.*, 2007) and at 6.2 keV vs. 12.4 keV (Weiss *et al.*, 2005). Homer *et al.* (2011) found a weak dependence of electron density decrease with disulphide bonds (but not methionine residues) indicating stronger damage at equal doses when using an X-ray beam energy of 14 keV compared to a 9 keV beam.

The work of Shimizu *et al.* (2007) was revisited by Fourme *et al.* (2011), who argued that the quantum detection efficiency of different types of detector at different wavelengths should be taken into account. The researchers showed, based on an analysis of previously published experiments, that with an ideal detector, the data collection efficiency, a global metric defined as the amount of diffraction data of a given resolution per crystal unit-volume, is increased by a factor of $\sim$8 at 33 keV compared with at 8 keV (Fourme *et al.*, 2012). Based on Monte Carlo simulations 33.3 keV was suggested as the optimal X-ray energy for most current crystal sizes, at which the data collection efficiency reaches its maximum (Fourme *et al.*, 2012).

1.2.4 Smaller crystals, smaller beams

Monte Carlo simulations predict that small crystal samples on the order of a few micrometres may result in better data in X-ray diffraction experiments, since high-energy photoelectrons could escape from the crystal (Nave & Hill, 2005; Cowan & Nave, 2008). Smaller crystals would therefore result in an effective reduction of the absorbed dose, and thus a reduction of radiation damage. Similarly, smaller beams result in a reduction of the apparent dose, as the dose and the radiation damage are spread out into the unexposed crystal volume (Sanishvili *et al.*, 2011). The dose-reduction effect may be stronger for high beam energies, as the photoelectron range is increased. Experimentally, it has been shown that a beam energy of 18.5 keV radiation damage can be detected up to 4 µm from a 1 µm beam (Sanishvili *et al.*, 2011).

Beam sizes on the same order of magnitude as the range of the photoelectron cascade are now attainable: at the microfocus beamline ID23-2 at the European Synchrotron Radiation Facility (ESRF), the measured full-width half-maximum focus size at the sample position is smaller than 7.5 µm × 7.5 µm in normal operating conditions, with an even smaller beam focus of 6.8 µm × 3.4 µm possible depending on the operation mode of the storage ring (Flot *et al.*, 2010).

1.2.5 Spatial distribution of dose

Even with regular-sized X-ray beams the dose can be spread out into the available crystal volume by using appropriate data collection strategies. In crystallography, data are typically collected by rotating a crystal through one or more angular ranges. If the crystal is larger than the beam, previously unexposed crystal volume can be rotated into the beam during the experiment (Schulze-Briese *et al.*, 2005). The total absorbed energy, and thus dose, is then spread throughout a larger

volume. Rotation can be combined with translation along the rotation axis, which results in a helical scan (Flot *et al.*, 2010). More complex data collection strategies (e.g. offsetting the rotation and the beam axis, resulting in a toroidal shape of the exposed volume, Zeldin *et al.*, 2013a) can be devised.

There are two basic issues with all these dose spreading strategies. Firstly, the dose absorbed by a crystal can no longer be expressed as a simple number describing a homogeneous dose state. Work on this has resulted in a number of alternative dose summary metrics (Zeldin *et al.*, 2013c), including the Diffraction Weighted Dose (DWD; Zeldin *et al.*, 2013a) which aims to state the relevant absorbed dose reflected in the collected dataset rather than within the crystal at the conclusion of the diffraction experiment. The software RADDOSE-3D (Zeldin *et al.*, 2013b), presented in Chapter 3, can simulate these dose spreading strategies and calculate the summary dose metrics.

Secondly, using diffraction data obtained from a larger crystal volume may result in a comparatively worse protein structure model. The crystal arrangement has been shown to vary within one crystal, possibly even to the same extent as the variation between different crystals grown with the same recipe (Pozharski, 2012).

1.2.6 The effects of dose rate

The question of whether the dose rate (the rate at which a certain dose is deposited in the crystal) makes a difference to observed radiation damage was a contentious topic in the field for some time. It is now widely accepted that there is no significant dose rate effect at 100 K and at the flux densities currently used (10^{15} photons/second/mm^2) as shown by Sliz *et al.* (2003).

For room temperature experiments a positive dose rate effect (higher dose rates leading to a higher absorbed dose before the same level of damage is reached) has been reported by

Southworth-Davies *et al.* (2007). Warkentin *et al.* (2012a) identified a positive dose rate effect at 260 K (-13°C), but not at room temperature.

At room temperature it was suggested that radiation damage processes would progress on time scales around 1 s (Warkentin *et al.*, 2012a). Owen *et al.* (2012) have demonstrated that it is possible to outrun radiation damage when collecting data using very fast data collection. An increase in the mean lifetime of crystals of ∼45 % was reported for data collection at 25 Hz compared to below 11 Hz for 25 images from immunoglobulin γ Fc receptor IIIa crystals. Comparable results were shown on virus and membrane protein crystals in the same study, and a model explaining the observed intensity decay of these data for high frame-rate room-temperature data collection has been proposed recently (Owen *et al.*, 2014).

The logical limit to experiments trying to decrease radiation damage by increasing the dose rate are experiments involving X-ray Free Electron Lasers (XFEL). By shooting many crystals with ultra-brief (<50 fs) yet extremely bright (10^{12} photons/pulse) X-ray pulses from an XFEL (Liu *et al.*, 2013) at most one diffraction pattern from each crystal can be obtained (Neutze *et al.*, 2000), and it has been demonstrated that it is possible to collect diffraction data at room temperature in which even the highly sensitive metal cluster of photosystem II shows no signs of radiation damage (Kern *et al.*, 2013).

1.3 Global, specific, and preferential radiation damage

Fundamentally, crystallography depends on an internal ordering of the crystal. Primary damage events on the protein, and interactions of the protein with high energy electrons, radicals, or other radiation by-products negatively affect the long-range crystal order. This loss of long range

order is also called global radiation damage, and can be seen in the acquired diffraction patterns throughout the experiment, see Figure 1.4. Diffraction spots, particularly those at high resolution on the fringe of the diffraction pattern, fade and disappear. Global radiation damage limits the useful lifetime of any one crystal in an X-ray beam, and reduces the resolution and quality of the resultant structure model.

Global damage also causes an increase in the unit cell size (Ravelli & McSweeney, 2000; Murray & Garman, 2002; Ravelli *et al.*, 2002), and apparent movement or rotation of the protein within the unit cell (Ravelli & McSweeney, 2000). These radiation damage effects cause non-isomorphism in MAD (multi-wavelength anomalous dispersion) phasing experiments, and are thought to be one of the major causes of failed MAD experiments. MAD phasing is a technique to solve the phase problem in X-ray crystallography[7], and depends on measuring changes in the reflection intensities at different X-ray wavelengths. The rotation of the protein molecule by 0.5°, the displacement of the protein by 0.1 Å, or a change in the unit cell size by 0.5 % in each dimension each result in an intensity change of between 15–17 % of a 3 Å reflection (Crick & Magdoff, 1956), utterly dominating any MAD phasing signal.

In contrast to global radiation damage, which affects the quality of the overall structure model, specific damage causes detectable changes in the protein structure (Weik *et al.*, 2000; Burmeister, 2000; Weik *et al.*, 2002), such as the reduction of metallo-centres, the breaking of disulphide bonds (Figure 1.5), and the decarboxylation of aspartate and glutamate residues. Specific radiation damage can cause systematic bias in the models of susceptible substructures, such as the active site. These radiation damage induced structural changes may lead to incorrect biological conclusions on protein mechanism and function (Ravelli & Garman, 2006).

[7] The phase problem is discussed in more detail in Section 4.1.1, page 130.

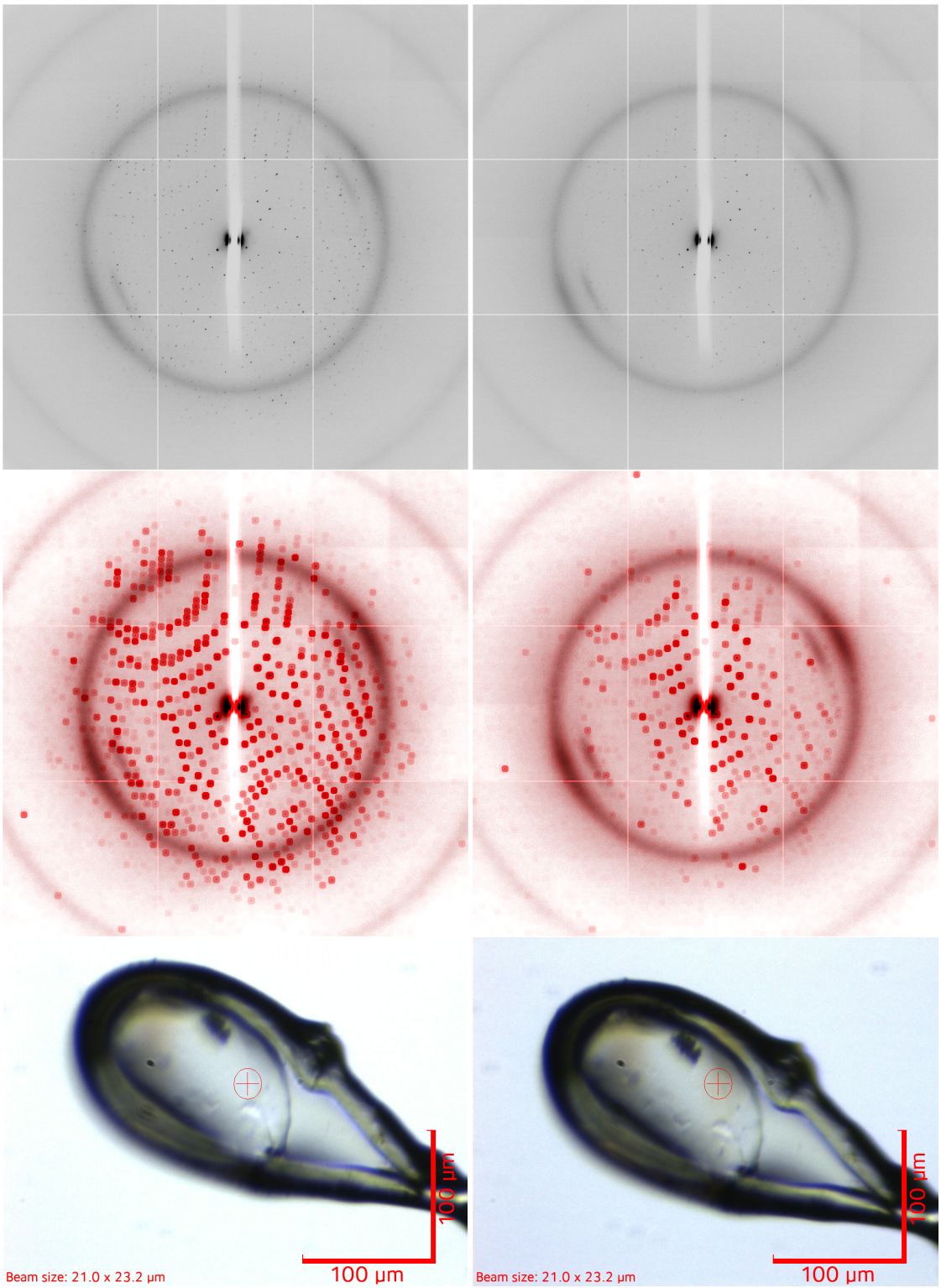


Figure 1.4: Visible effects of radiation damage during the X-ray diffraction experiment. Data are collected from a crystal. The **left** column shows the observations at the beginning of the data collection, the **right** column at the end of the data collection. **Top:** The diffraction pattern diminishes, and particularly the high resolution reflections at the outside of the detector image disappear. This is even more obvious when looking at the enhanced diffraction images in the **middle** row. By surrounding each dark pixel with a small red circle of the same intensity the loss of diffraction quality becomes clearly visible. **Bottom:** The irradiated crystal volume becomes increasingly discoloured as the experiment progresses. The marks of a previous data collection on the other half of the crystal are clearly visible.

Crystal p18 (Table 4.4, page 154) of RhoGDI EA protein, data collected at 100 K at DLS I04, 23.09.2012. Detector distance set for collection of data up to a resolution of 2 Å. Absorbed dose in the relevant region ~0 MGy in the left column (first diffraction image), >80 MGy in the right column.

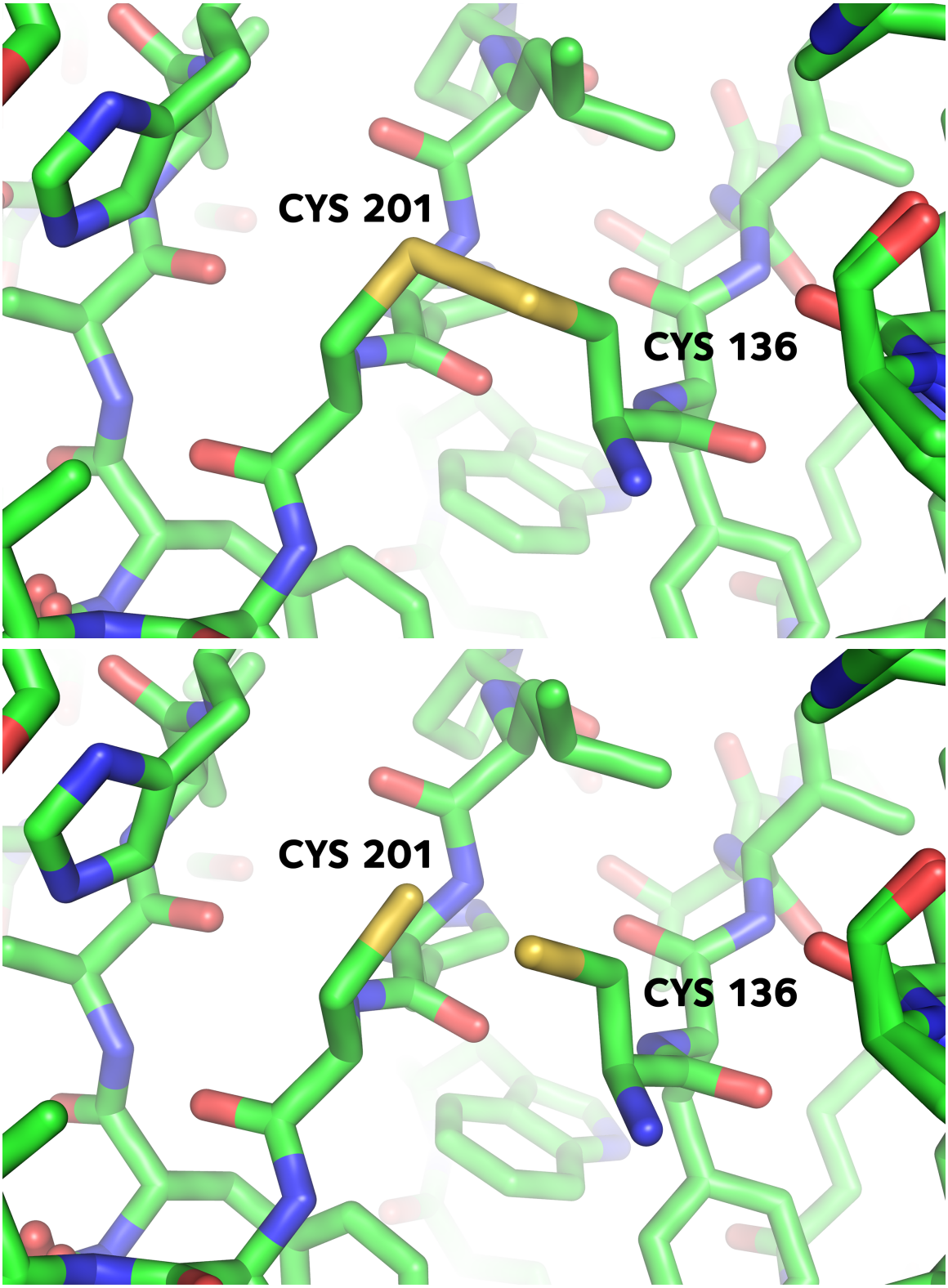


Figure 1.5: A disulphide bond (**top**) in *Sus scrofa* pancreatic elastase is broken (**bottom**) by specific radiation damage.

Data collected at 100 K with 1.20 Å and 1.82 Å resolution respectively, and published as PDB structure IDs 3MNB and 3MOC (Petrova *et al.*, 2010). Corresponding doses are given as 1.2 MGy for the top structure model and 57 MGy for the bottom structure model, calculated using RADDOSE.

The disulphide bond is probably the best studied example of specific radiation damage in macromolecular crystallography, and while ever more complete models are proposed, its exact decay mechanism is still not very well understood (Burmeister, 2000; Weik *et al.*, 2002; Petrova *et al.*, 2010; Sutton *et al.*, 2013). Less is known about the mechanisms of decay of the other affected chemical groups, although the possible radiation chemistry involved has been collated (e.g. Burmeister, 2000).

Specific radiation damage is not due to primary events alone: for example sulphur atoms have a comparatively large X-ray cross-section and therefore have a larger cross-section for the incident beam, implying that they would experience more primary radiation damage. This is true for the sulphur atoms in a disulphide bond, but also for the sulphur atoms in methionine residues. However, the latter are much less sensitive to specific radiation damage than the former. The carboxyl groups of aspartic and glutamic acid residues are composed of light atoms and have a small X-ray cross-section, yet they are more sensitive to specific radiation damage than the methionine side chain (Burmeister, 2000).

Similarly to this ordering of susceptibility for different chemical groups, there is also an ordering within each group in an individual protein structure. When there are multiple occurrences of the same sensitive motif, such as more than one disulphide bond in a protein molecule, they may not decay at equal rates. Hereafter in this thesis, this process is termed 'preferential specific damage'. The causes of this difference in susceptibility are hitherto unclear.

The question of what determines these different rates of preferential damage is the main question of the present study. While the question can be stated in a simple manner, any answer will necessarily be complex: It will depend on various radiation decay mechanisms, and different

chemical groups may be differentially susceptible for entirely different reasons.

To be able to identify and understand the radiation decay mechanisms, radiation damage observations from different sources need to be drawn together and compared. For this to be meaningful, a common baseline needs to be established: It is necessary to reliably and objectively quantify both the level of specific radiation damage that a particular chemical group has suffered, and the dose that the protein crystal sustained throughout the experiment. This study aims to open up new sources of radiation damage decay data to enable the required rigorous statistical investigation into the causes of preferential radiation damage.

1.4 Chapter overview

The work presented in this thesis brings three major contributions to the field of radiation damage research in macromolecular crystallography.

Firstly, it considers the possibility of the identification and quantification of specific radiation damage in single structure models. Over the past century there have been millions of X-ray diffraction experiments which resulted in hundreds of thousands of published protein structure models. These provide a large, yet largely untapped resource for radiation damage research in macromolecular crystallography. The aim of a regular X-ray diffraction experiment involving a new protein is to provide the most accurate structure model, and not to study its radiation damage decay behaviour. As a result, experiments are not usually set up in a way that their results can be used in radiation damage studies. In Chapter 2 a new metric, B_{Damage}, is proposed, which can be used to identify locations in a protein structure model which have been subjected to specific radiation damage, and to quantify the relative level of damage. B_{Damage} only relies on a single

structure model, and describes the first attempt in the field to carry out a large-scale statistical analysis of the radiation damage susceptibility using the vast body of published protein structures.

Secondly, there is a need for more accurate dose values. Dose is a measure of the energy deposited through irradiation in a given amount of mass. Radiation damage observations are always made at a certain level of irradiation. It is only through dose values that experiments by different researchers, or involving different crystals or beamlines, and thus any observations become comparable with one another. Dose values can also provide guidance to researchers in planning and optimising data collection strategies to obtain the best possible diffraction dataset from the available crystal volume.

The introduction of RADDOSE (Murray *et al.*, 2004; Paithankar *et al.*, 2009; Paithankar & Garman, 2010) made it possible to estimate experimental doses by imposing some simplifying assumptions. These include, for example, that the crystal is smaller than the beam so that rotation does not have to be considered, that the beam can be described by a few parameters, and that the maximum dose at any point in a crystal is the appropriate dose figure to state and on which to base further analysis. With the arrival of new microfocus beams (such as the one mentioned earlier at ESRF ID23-2), the possibility of characterising beam profiles with much more detail, and the increase of computing power and hence new opportunities to model experiments in a much more fine grained manner, these assumptions in RADDOSE are no longer appropriate. Chapter 3 describes work on three-dimensional modelling of X-ray diffraction experiments, culminating in completely new software: RADDOSE-3D. Three-dimensional modelling, new dose summary metrics, and the prospect of wiring dose modelling software directly into MX beamlines promise much more easily obtainable, yet more accurate and precise dose estimates. These estimates allow the

optimisation of data collection strategies to improve the use of available crystal volume in X-ray diffraction experiments.

The final investigation, presented in Chapter 4, concerns an extensive radiation damage progression study on three RhoGDI protein mutants with very high sequence identity. The decay of real space electron density is observed in order to identify, quantify, and understand the stabilising or sensitising effect of the local environment on protein residues. A new data collection and analysis protocol specifically designed for time-resolved radiation damage research is presented. The results draw a much clearer and higher resolved picture of how specific radiation damage progresses, and make the detailed experimental investigation into the causes and processes of specific radiation damage feasible.

Together, all these investigations aim to further the current understanding of radiation damage mechanisms in macromolecular crystallography, to facilitate the optimisation of X-ray diffraction experiments, and, through these, to improve the quality of future structure models.

Chapter 2

PDB Analysis with B_{Damage}

PDB Analysis with B_{Damage}

In this chapter a new atomic metric, B_{Damage}, is presented[1]. B_{Damage} aims to identify and quantify site-specific radiation damage in protein structures. By relying only on data available from a single structural model and not requiring a before–/after–dataset, this metric opens up the vast library of published protein structure models and thus provides a powerful tool for the statistical analysis of specific radiation damage.

The veracity of B_{Damage} is verified using a before–/after–dataset of published protein structures, before being used in a large-scale statistical survey into some possible causes for specific radiation damage, where it was successfully applied to identify a consistent positive correlation between specific damage and solvent accessibility.

2.1 Preferential specific radiation damage

Radiation damage is an integral part of macromolecular X-ray crystallography (MX). The rate of radiation damage can be reduced, for example by conducting experiments at cryotemperatures (Nave & Garman, 2005) or using scavengers (Barker *et al.*, 2009; De la Mora *et al.*, 2011; Kmetko *et al.*, 2011), but it can not be avoided.

As mentioned in the last chapter, there are two classifications for radiation damage: Global damage can be observed in the decay of the diffraction pattern, and an increase in unit cell volume and often of mosaicity. The increase in unit cell volume leads to non-isomorphism developing

[1] The work described in this chapter was presented at the 28th European Crystallography Meeting, Warwick, United Kingdom, August 2013 (Gerstel *et al.*, 2013), the 8th International Workshop on X-ray Radiation Damage to Biological Crystalline Samples, Hamburg, Germany, April 2014 (Gerstel *et al.*, 2015), and the 24th Congress and General Assembly of the International Union of Crystallography, Montreal, Canada, August 2014 (Gerstel *et al.*, 2014).

during the experiment and causes difficulties in, for example, MAD (multi-wavelength anomalous dispersion) structure determination. Meanwhile specific radiation damage manifests itself as a change in the electron density around particular protein regions in a repeatable pattern.

The work in this chapter is focused on the process and effects of specific radiation damage.

2.1.1 From specific to preferential specific radiation damage

Specific radiation damage processes at 100 K follow a certain, well established order: metallo-centres are reduced (Yano *et al.*, 2005) and disulphide bonds are radicalised (Sutton *et al.*, 2013) very early in the experiment. Then, disulphide bonds are elongated and broken (Burmeister, 2000; Ravelli & McSweeney, 2000; Weik *et al.*, 2000, 2002), glutamates and aspartates are decarboxylated (Burmeister, 2000; Ravelli & McSweeney, 2000; Weik *et al.*, 2000; Fioravanti *et al.*, 2007), tyrosine -OH groups are lost (Burmeister, 2000), the methionine C_γ-S_δ bond is cleaved, leading to a loss of the SH-CH_3 methylthio group (Ambe *et al.*, 1961; Shimazu *et al.*, 1964; Burmeister, 2000; Homer *et al.*, 2011), and covalent metal-bonds are broken (Ennifar *et al.*, 2002; Ramagopal *et al.*, 2005). Different chemical groups are therefore affected by radiation damage to different extents at different doses throughout the experiment.

The rate of specific radiation damage has been shown to depend on a number of different experimental factors. Two studies on the effect of X-ray photon energy on the rate of specific damage came to different conclusions: Shimizu *et al.* (2007) reported no change in specific damage, as determined by atomic B-factors (also known as atomic displacement parameters, or B-factors for short), at nine different X-ray energies between 6.3 and 33.0 keV. However, Homer *et al.* (2011) compared specific damage observing real space electron density and reported lower specific damage at disulphide sulphurs at an incident photon energy of 9 keV compared with 14 keV, but no

change at methionine sulphurs. Both studies were conducted using lysozyme crystals at cryo-temperatures (100 K). The difference in radiation damage susceptibility of the different chemical groups is believed to be due to their different affinity for mobile secondary electrons which are mobile even at 77 K (Jones *et al.*, 1987). After metal centres, disulphide bonds are most susceptible, as they are the most electron affinic part of the protein.

The impact of specific radiation damage goes beyond these effects on isolated residues. Radiation induced structural changes may cause the experimenter to draw incorrect conclusions from the data. For instance in the investigation of reaction pathways of functional proteins, structural changes caused by radiation damage may be mistaken as the formation of reaction intermediates, as noted by Matsui *et al.* (2002) in a study of bacteriorhodopsin. Work on photoactive yellow protein by Kort *et al.* (2004) identified a new low temperature photointermediate that was only detectable by minimum dose data collection and correction for structural changes induced by specific radiation damage, although structural data were previously available at 0.85 Å resolution (Genick *et al.*, 1998).

It has also been shown that multiple occurrences of identical chemical groups can reproducibly damage at different rates within the same protein crystal. This process is called preferential specific radiation damage. While it is obvious that the causes for preferential damage must lie with the environment of the affected protein sites, the exact causes for preferential damage are still unclear. The new metric presented in this chapter, B_{Damage}, aims to aid in the statistical analysis of preferential specific damage by allowing the identification of correlations using a much larger, representative sample of structural models, instead of relying on observations obtained by comparing multiple datasets collected from the same crystal.

2.1.2 Radiation damage between multiple datasets

The common method used to identify damage at protein sites is to obtain multiple datasets from the same crystal in different damaged states. The assumption is that any structural changes from the first model are caused by specific radiation damage. For this approach it is vital that the first dataset is obtained with the lowest possible dose, to avoid the 'contamination' of this reference model by specific radiation damage. Properly established dose values for each dataset (see Chapter 3, RADDOSE 3D, page 81) also allow the identification of the absolute susceptibility of each affected protein site as a function of dose. It is then possible to compare the susceptibilities of identical structural motifs within the same protein and correlate these with their surroundings to try and identify possible causes for preferential damage.

It has been shown by Ravelli *et al.* (2003) that the differences between two datasets caused by specific radiation damage can be strong enough to allow the experimental determination of phases. This 'Radiation-Damage-Induced Phasing' (RIP) technique is homologous to single isomorphous replacement (SIR; Crick & Magdoff, 1956), and is based on the disappearance (or smearing-out) of the electron density of sulphur atoms in the dataset corresponding to a more damaged protein state. For this, a later dataset is treated as a 'native' dataset, i.e. a dataset without heavy atoms present. An earlier dataset of the same crystal with well defined sulphur atom positions can then be understood as a 'derivative' dataset, which contains 'additional' sulphur atoms compared to the native dataset. The differences between the two datasets can provide enough information to solve the phase problem.

2.1.3 Investigations into preferential radiation damage using multiple datasets

Multiple datasets from the same crystal have also been used to investigate the causes of preferential radiation damage. The susceptibility of the three disulphide bridges in cubic insulin was investigated by Meents *et al.* (2010), who showed that the rate of damage to the solvent-exposed disulphide bridge was temperature dependent, with a fourfold increase in occupancy decay rate at 100 K compared to that at 50 K. The susceptibility of the other two disulphide bridges buried inside the protein showed only a very small temperature dependence. Petrova *et al.* (2010) measured specific damage using a decrease in atomic occupancy as a metric, and found the appearance of new disulphide cysteine rotamers in elastase crystals at temperatures of 15 K and 100 K to be directly correlated to local solvent accessibility. Using atomic B factors, Juers & Weik (2011) identified a correlation between specific damage and the distance to the nearest solvent channel in thermolysin at 160 K, but did not observe this at 100 K. A similar effect was reported by Warkentin *et al.* (2012b), who, again using atomic B factors, identified seven solvent-exposed turns of thaumatin as being the most radiation sensitive parts at experimental temperatures of 180 K and above, but not at 155 K and below.

Fioravanti *et al.* (2007) showed that radiation susceptibility does not correlate with solvent accessibility at 100 K for malate dehydrogenase, which, as a halophilic enzyme, contains 35 aspartic and 27 glutamic acid residues in each chain of 304 residues. Susceptibility was identified by peaks in the difference Fourier maps between the first and one of the two subsequent datasets, obtained after X-ray burn phases. Homer *et al.* (2011) also could not find evidence for a relationship between side-chain solvent accessibility and radiation susceptibility, determined by the electron density decay of lysozyme crystals at 100 K. Some observations indicate that acidic

residues with higher pKa are more sensitive to radiation damage, but this correlation has not yet been demonstrated (Fioravanti *et al.*, 2007). Ravelli & McSweeney (2000) could not observe a clear correlation between the susceptibility of aspartic and glutamic acid residues, and either solvent accessibility or pKa using lysozyme crystals at 100 K.

Holton (2009) suspected that preferential specific radiation damage is not caused by a single factor, but a combination of susceptibility factors, including heavy atom proximity, chemical bond strain, and electrostatic field lines.

2.1.4 Detecting damage in a single dataset

By investigating the differences between a low- and a high-dose dataset, much information can be gathered about the radiation damage decay process for a specific protein. While some radiation damage series are available (Table 2.1) these are not always suitable for radiation damage detection at an atomic level, and generally only cover a small set of well studied model proteins, such as lysozyme, insulin and elastase.

If it were possible to eliminate the need for before-after-datasets and instead identify radiation damage from a single protein structure, a much larger and much more representative set of protein structure models could be used for the investigation of radiation damage. As evidenced by the known and reproducible order of susceptibility of different protein motifs, at its core specific radiation damage is not a random process. It should therefore be possible to identify systematic changes in the electron density due to specific radiation damage. These systematic changes would manifest themselves in a changing electron density distribution, where the more damaged regions generally lose definition and are smoothed out. This effect in turn causes a change in the published protein structures which model an electron density distribution closely matching the experimental

PDB codes	Protein / Authors	Notes
2YBH,2YBI,2YBJ, 2YBL,2YBM,2YBN	Chicken Egg White Lysozyme, *Gallus gallus* De la Mora *et al.* (2011)	2.31–28.6 MGy scavenger investigation
2J5K,2J5Q,2J5R	Malate Dehydrogenase, *Haloarcula marismortui* Fioravanti *et al.* (2007)	1.2, 4.6, 8.2 MGy
1DWA,1DWF,1DWG, 1DWH,1DWI,1DWJ	Myrosinase, *Sinapis alba* Burmeister (2000)	doses unclear (flux density given)
1QID,1QIE,1QIF, 1QIG,1QIH,1QII, 1QIJ,1QIK,1QIM	Acetylcholinesterase, *Torpedo californica* Weik *et al.* (2000)	no per-atom refinement of atomic B factors
4M4F,4M4H, 4M4I,4M4J	Cu T(6) Insulin, *Bos taurus* Frankaer *et al.* (2014)	0.01, 0.06, 0.12, 0.30 MGy
3P7P,3P7Q,3P7R, 3P7S,3P7T,3P7U, 3P7V,3P7W	Thermolysin, *Bacillus thermoproteolyticus* Juers & Weik (2011)	0.1–7.1 MGy
4H8X,4H8Y,4H8Z, 4H9O,4H91,4H92, 4H93,4H94,4H9A, 4H9B,4H9C,4H9E, 4H9F,4H9H,4H9I	Chicken Egg White Lysozyme, *Gallus gallus* Sutton *et al.* (2013)	0.07, 0.14, 0.21, 0.28, 0.35, 0.42, 0.49, 0.56, 0.63, 0.70, 0.77, 0.84, 0.91, 0.98, 1.05 MGy
3MNB,3MNC,3MNS, 3MNX,3MO3,3MO6 3MO9,3MOC,3MTY 3ODF,3MUO,3MU1 3MU4,3ODD,3MU5 3MU8	Pancreatic Elastase, *Sus scrofa* Petrova *et al.* (2010)	1.2–58 MGy for the first 8 PDB IDs; further doses not calculated
4L3X,4L3Y,4L3Z	Nitrite Reductase, *Thioalkalivibrio nitratireducens* Trofimov *et al.* (2013)	doses unknown
3NSO,3NSB	Bacteriorhodopsin, *Halobacterium salinarum* Borshchevskiy *et al.* (2011)	doses unclear; 3NSB contains both low- and high-dose data
2BLO,2BLQ	Elastase, *Sus scrofa*	80 kGy–2.2 MGy
2BN3,2BN1	Insulin, *Bos taurus*	2.0 MGy–6.0 MGy
2BLX,2BLY	Lysozyme, *Gallus gallus*	0.6 MGy–3.2 MGy
2BLP,2BLZ	Ribonuclease A, *Bos taurus*	3.0 MGy–10 MGy
2BLR,2BLU	Thaumatin, *Thaumatococcus daniellii*	0.8 MGy–3.6 MGy
2BLV,2BLW	Trypsin, *Bos taurus* Nanao *et al.* (2005)	0.4 MGy–2.8 MGy

Table 2.1: Selection of low-/high-dose datasets and radiation damage series available in the Protein Data Bank (PDB).

observations.

Inspired by an earlier unpublished study by Philip Hogg and Raimond B. G. Ravelli, reported at the 5th International Workshop on Radiation Damage to Biological Crystalline Samples (3.–5. March 2008) at the Swiss Light Source, Switzerland, Turner *et al.* (2010) studied single dataset protein structures in an effort to relate the cleaving of disulphide bonds to their local environments. Continuing this work, B_{Damage} aims to combine the knowledge of the degradation state of the entire protein with the state of an individual atom, into a number expressing the location-specific electron density degradation.

2.2 Using structure models

Unlike global radiation damage, which affects the obtainable data in Fourier space, specific radiation damage directly alters the real space structure. A study of specific radiation damage therefore needs to concern itself with real space data drawn from structural model coordinate files. This section gives an overview of the available sources for these files, the data contained within them, and the usefulness of these data for this particular investigation.

2.2.1 Sources of structural models

The Protein Data Bank (PDB) is a repository containing over 100,000 macromolecular structure models, mostly obtained by X-ray crystallography or NMR spectroscopy. It is maintained by the Worldwide Protein Data Bank (Berman *et al.*, 2003, http://www.wwpdb.org/) and provides free, open access to all deposited structures.

This very modern and open research approach allowed the development of further projects

using the knowledge available in the PDB in novel and different ways. Notable examples are the PDB_REDO project (Joosten *et al.*, 2014, http://www.cmbi.ru.nl/pdb_redo/), which aims to continually improve existing protein structure models with the most up-to-date automated refining methods, and the Protein Model Portal (Haas *et al.*, 2013, http://www.proteinmodelportal.org/) and the Structural Biology Knowledgebase (Gabanyi *et al.*, 2011, http://sbkb.org) which draw together information about each protein from a curated collection of databases, including the PDB.

2.2.2 Available information in a structure model

Each entry in the PDB is uniquely and permanently identified by a sequence of four alphanumerical characters, the so-called PDB ID. Deposited PDB structure files contain one, sometimes more models consisting of coordinate records for every observed atom in the molecule. For each atom, information relating to the protein's primary sequence is stored alongside its positional model. The positional model consists of the Cartesian coordinates of the mean atom location within the unit cell, an occupancy value, and one (isotropic) or six (anisotropic) atomic B factors, also called atomic displacement parameters[2].

Atomic B factor

These atomic B factor parameters describe the reduction in scattering expected from a point source due to that source having a range of positions. They can also be understood to describe the uncertainty on the location of the atom, including that caused by thermal and crystalline disorder.

The decision of whether only one parameter per atom, or full anisotropic refinement can be used, is mostly informed by the quantity of experimental data available. A model with full

[2] Other names for the same concept, such as 'B values' and 'anisotropic displacement parameters', with the same ADP acronym, can sometimes be found in the literature.

Atomic B factor [Å^2]	Number of atoms	Percentage of atoms
B < 0	3	<0.01 %
B = 0	4,610	0.05 %
0 < B < 80	10,143,262	99.65 %
B ≥ 80	31,338	0.31 %

Table 2.2: Observed B factors in a set of 2,704 PDB structures with a resolution between 1.5 and 1.8 Å. Most observed B factors lie between 0 and 80 Å^2. Unlike atomic occupancy (Table 2.3), B factors are usually refined. One PDB structure, 3CLS, contains 3 atoms with an impossible, negative B factor.

anisotropic B-factor details has a much larger number of parameters (including the occupancy, 10 per atom, as opposed to 5 per atom). The model therefore depends on the availability of a larger number of observations, specifically unique reflections in the diffraction pattern, the number of which is itself a function of the resolution of the diffraction. Atomic B factors usually range from 0 Å^2 to 80 Å^2 (Table 2.2), and can, in the absence of translation/libration/screw groups (Schomaker & Trueblood, 1968; Painter & Merritt, 2006) or similar approaches, be refined independently. However, commonly used refining software will by default only allow a limited difference in atomic B factors of covalently bound atoms (Ed Lowe, personal communication, April 21st, 2014). An atomic B factor of 0 Å^2 represents a non-vibrating atom in an identical position in all unit cells.

Occupancy

Further, each atom in the model is annotated with an occupancy value. This value represents the probability of observing the atom at the given location in any one randomly selected unit cell. There may be a number of explanations for encountering occupancy values lower than one. For example, the atom may occupy a special position on a symmetry axis (Dauter & Jaskolski, 2010), if the atom is a metal ligand it may not be bound at each copy of the protein, or it may not be

Occupancy	Number of atoms observed	Percentage
occ < 0	5	<0.01 %
occ = 0	9,177	0.09 %
0 < occ < 0.5	68,511	0.67 %
occ = 0.5	234,192	2.30 %
0.5 < occ < 1	72,020	0.71 %
occ = 1	9,795,303	96.23 %
1 < occ	5	<0.01 %

Table 2.3: Observed occupancy values in a set of 2,704 PDB structures (see Section 2.5.1 for the full selection criteria and process).
Most atom locations have a occupancy fixed to unity. When the value is not set to unity, it is used to distinguish between two conformations. Although these conformations are probably not equally likely, a fixed value of 0.5 was chosen instead of refining the occupancy. Some PDB structures contain occupancy values outside the physical range $(0; 1]$. In particular, the structure entries 1IEO and 1WUL report occupancies above 1, and the structure 3HKW reports occupancies below 0.

bound or ordered throughout the entire experiment. The information contained in the set of diffraction images represents a time and space-average of the unit cell, so it may not be possible to determine the reason for reduced occupancy values from a single experiment. Occupancy values are regularly used to describe alternative conformations. In this case the occupancy value is understood as representing the prevalence of a specific conformation. Conformations involve a number of atoms and the occupancies of the participating atoms are usually set to identical values; they should in any case not be understood as independent values.

If multiple conformations are marked using occupancy values then it is not necessarily the case that the sum of the occupancy values for any given atom reaches 1, but any value above 1 can be considered an unphysical model. Unfortunately not all protein structure models observe this simple rule. Due to a combination of erroneous refinement of occupancy values, a lack of understanding of the concept of occupancy on behalf of the experimenter, or a lack of oversight of the automated processes, occupancy values outside the meaningful region $(0; 1]$ can be found in the PDB (Table 2.3).

The distribution of observed occupancy values indicates a deeper problem in using these values for radiation damage detection: in the majority of cases the occupancy is fixed to unity. In other cases, the model creators mostly use it to mark two alternative conformations with occupancy values, but prefer 'clean' values such as 0.5 each. Occupancy values in PDB structures are usually not refined, and therefore only carry limited information. This makes the occupancy value unsuitable for radiation damage identification purposes.

Basis for radiation damage detection

Given a PDB file, the only suitable starting point for an atomic metric describing the specific radiation damage that has been suffered by the structure are the atomic B factors. In this study only non-hydrogen atoms and isotropic atomic B factors were considered. The more informative anisotropic atomic B factors were not used since, due to the large number of degrees of freedom and the associated requirement for high resolution data, there are relatively few deposited structures which contain them.

2.2.3 Structure model deficiencies

To compare observations concerning radiation induced structural damage between experiments, the decrease in electron density is routinely measured against the absorbed dose (absorbed energy per unit mass; 1 Gy = 1 $^{J}/_{kg}$). While refined protein structure models from X-ray determinations are now deposited in the Protein Data Bank (PDB; Berman *et al.*, 2003) together with the corresponding structure factors, they are not usually accompanied by dose values. Therefore users of these protein structure models have generally little to no information available on the degree of specific radiation damage that might have been suffered. One reason for this is that it is impossi-

ble for the experimenter to directly and quantitatively measure the effective dose absorbed by a crystal. Experimental proxies such as relative isotropic B factors (B_{rel}) (Kmetko *et al.*, 2006) or the ratio of summed mean intensities I_D/I_1 of successive data sets (Garman, 2010) can be used to track the progression of damage in reciprocal space within a certain range, but these do not allow comparisons between experiments. Other directly or indirectly observable experimental proxies, such as the unit cell size and mosaicity increase do not relate to absorbed dose in a systematic manner (Murray & Garman, 2002; Ravelli *et al.*, 2002).

Even when dose values are available, these are usually only reported in the publications accompanying new X-ray structure depositions, and they are not systematically reported within the PDB depositions themselves. They are therefore not generally available for experimenters working with many PDB structures. For this reason it is also unknown how many of the deposited PDB structures were determined from X-ray diffraction experiments exceeding the recommended maximum dose limit of 30 MGy (Owen *et al.*, 2006), and whether they are compromised by radiation induced structural damage, and if so, to what extent. This information, if available, might help those who use PDB structure models derived from X-ray crystallography to avoid treating the deposited models as 'truth written in stone' (Pozharski *et al.*, 2013).

2.3 Introducing B_{Damage}

The electron density model based on a dataset obtained from a crystal subjected to radiation damage is generally distributed over a larger volume compared to one based on a dataset obtained from a pristine crystal. This 'smearing' of electron density is reflected in the structure model by an increase in the atomic B factors.

However radiation damage is only one component influencing the final atomic B factors. Mobility and intrinsic disorder of a particular substructure provides a much stronger contribution to the B factor value. Thus atoms on the surface of macromolecules generally have a higher B factor than buried atoms that are held in place by neighbouring residues. Weiss (2007) has shown that, given information about the 'buriedness'[3] of atoms, it is possible to predict the atomic B factors with some reliability. If 'buriedness' provides a useful predictor of this strong contribution, it may be possible to investigate the effects of radiation damage using a metric based on atomic B factors but corrected by the 'buriedness'-effect. This metric is B_{Damage}.

B_{Damage} is a per-atom, scalar metric, defined as the isotropic B factor of the atom divided by the average B factor for all protein atoms in a similar environment. Using B_{Damage} allows the identification of locations in a protein structure model that have been affected by specific radiation damage, and the quantification of that effect. In contrast to other available methods, B_{Damage} does not require both a before- and after-damage dataset or model, but can be applied to any single existing protein structure model.

The process of calculating B_{Damage} can be split into three steps, which are explained throughout this section: the determination of the packing density of an atom (Section 2.3.1), the identification of similar regions within the same protein (Section 2.3.2), and finally the calculation of the B_{Damage} value (Section 2.3.3). The methods and difficulties of calculating B_{Damage} in actual protein structure models are discussed in Section 2.3.4.

[3] The 'buriedness' was expressed via a packing density metric, in this case atomic contact numbers. See Section 2.3.1.

2.3.1 Packing density determination

B_{Damage} relates the B factor of an atom to those of atoms in a similar environment, with the similarity based on packing density. There are multiple possible ways to define packing density. Any definition of packing density can be understood as a function $P : A \to \mathbb{R}$ that assigns a packing density value $P(a)$ to each atom a in the set A of all atoms within the protein. As presented here, B_{Damage} is only defined for non-hydrogen atoms, and hydrogen atoms are not considered in the calculations.

The three most well known packing density functions are the Ooi number (Nishikawa & Ooi, 1986), the cx ratio (Pintar *et al.*, 2002) and the atomic contact number (ACN; Weiss, 2007). The Ooi number of an atom a is defined as the number of C_α atoms around a sphere with a fixed radius r around the C_α of the amino acid residue containing the atom a. The CX[4] algorithm provides a measure of protrusion, and the cx ratio is defined as the ratio between the unoccupied and occupied volume within a sphere with radius r around the atom a. The atomic contact number corresponds to the number of atoms around a sphere with radius r around the atom a.

All three metrics (Ooi, CX, ACN) are based on what lies within a sphere of a certain radius (sphere of interest). A small radius up-weights local differences; a larger radius shifts the focus towards global protein properties. Radii between 6–8 Å are typically used (Halle, 2002; Nishikawa & Ooi, 1980; Weiss, 2007). At 8 Å only those residues directly in contact with a given residue are included (Nishikawa & Ooi, 1986). Pintar *et al.* (2002) recommend 10 Å, although radii up to 14–18 Å have also been chosen, particularly when describing global structural features (Nishikawa & Ooi, 1986).

[4] Pintar *et al.* (2002) named the individual ratios 'cx' in lower case letters, and the algorithm itself 'CX' in upper case letters.

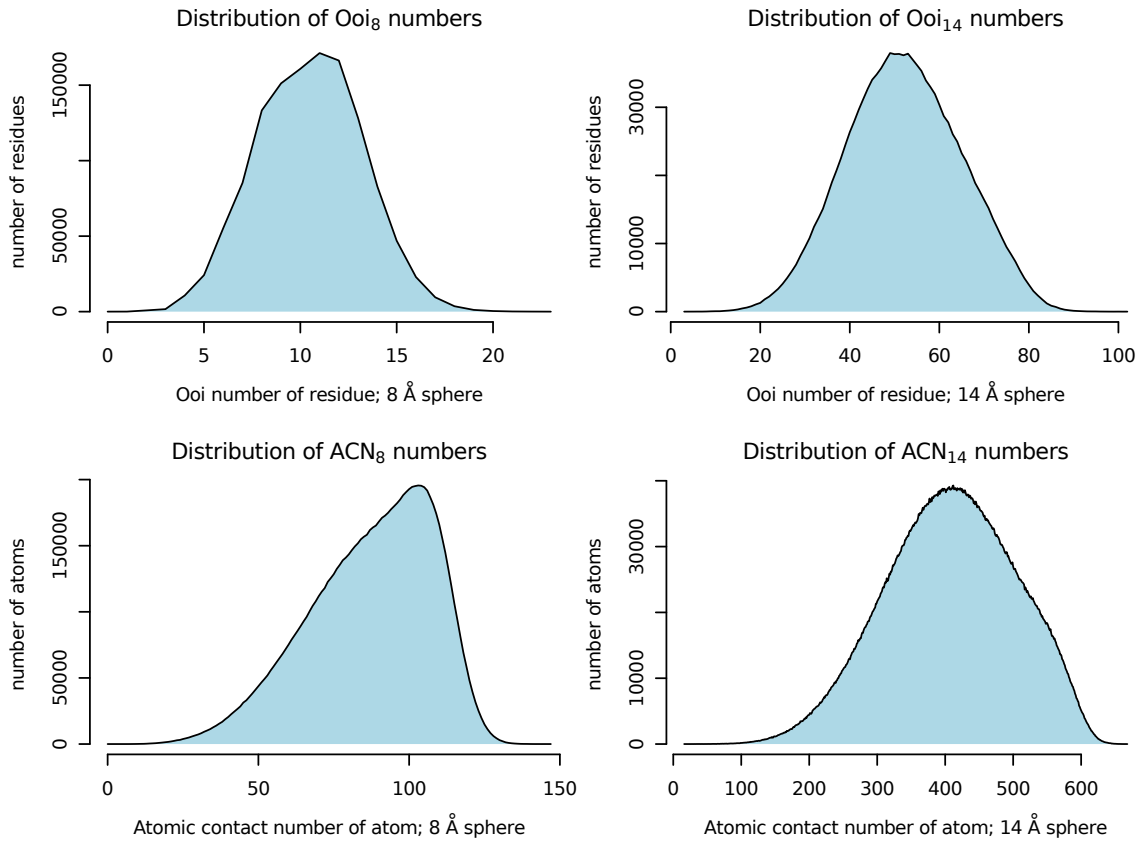


Figure 2.1: Distribution of Ooi numbers and atomic contact numbers (ACN) for radii of 8 Å and 14 Å for the 2,704 selected PDB structures. Note that both Ooi numbers and ACNs are discrete metrics; they can only take integer values.

If the radius is too small, then the packing information is strongly influenced by the surroundings of the local residue, and the possible spread in numerical values for the packing density metric is low. In the work presented here, the packing density metric is only used to provide a partitioning of the protein into volumes of similar packing density. A low spread in values increases the influence of stochastic effects and noise: thus a suitable packing density should have a minimum radius of 8 Å. With an increasing radius, the spread in the numerical values of Ooi and atomic contact numbers increases (Figure 2.1). However, if the radius is too large, then the effects of secondary structure are smoothed over and the packing density information is blurred throughout the protein. Again this means that the partitioning is adversely affected.

The cx ratio of a protein atom usually lies between 0 and 15 (Pintar *et al.*, 2002). CX, in its basic form, makes the assumption that every atom takes up the same volume (20.1 $Å^3$; Richards, 1974). If, by this definition, the same volume is simultaneously occupied by n atoms, it is counted n times. When the atom occupying the volume lies outside the sphere of interest, its volume is not considered, even if its volume intersects within the sphere of interest. Similarly, the volume of any atom that itself lies within the sphere of interest is counted completely, even if its volume partially lies outside the sphere of interest.

The Ooi number is based on the sphere around the residue's C_α, therefore all atoms within the residue have the same Ooi number. This is not a desirable property for a packing density metric to be used in B_{Damage}. While it may not be a problem for small amino acids such as alanine or glycine, the ends of side chains of larger amino acids, particularly lysine, can reside in very different packing environments compared to those of the main chain carbon atoms. The Ooi number also fails to take the side chains of other residues into account: A densely packed environment containing only glycine residues will always result in higher Ooi numbers than an equally densely packed environment containing only larger amino acid residues, as it measures the density of C_αs.

As presented here, B_{Damage} relies on atomic contact numbers. In contrast to Ooi numbers, atomic contact numbers take all non-hydrogen atoms into account. Additionally the sphere of interest is based on each atom itself, rather than being representative of that residue. When calculating the packing density, the neighbouring copies of the protein via crystallographic symmetry operators have to be taken into account, since radiation damage occurs in the context of a protein crystal, not just on an isolated molecule. The calculation of atomic contact numbers is straightforward and can be easily implemented within the database framework used in this inves-

tigation (Section 2.3.4) and extended for the symmetry atoms, which is an advantage over the CX algorithm.

Typical Ooi and ACN packing densities for different secondary structures are shown in Table 2.4. With a 14 Å radius the different secondary structure elements exhibit similar intrinsic differences between their distributions as with smaller radii: e.g. atoms in π-helices have high minimum and mean packing densities with a large variance, while coils, 3-10-helices, and turns have the lowest mean packing densities but with a large variance. Thus packing density information is not yet blurred throughout the protein with a radius of 14 Å.

2.3.2 Partitioning a protein: similarity

B_{Damage} relates the B factor of an atom to those of atoms in a similar environment, with the similarity based on packing density. With the packing density defined, the next issue is the definition of *similarity*.

The simplest possible definition of similarity is equality: All (non-hydrogen) atoms a with the same packing density metric n can be grouped together into a packing density environment E_n so that $E_n := \{a \in A \mid P(a) = n\}$, with P being the packing density function. It is obvious that the set of all non-empty E_n now forms a partition of the set of all protein atoms A.

However, as has been shown above, especially when using larger radii the number of possible packing density metrics n increases. With a sphere of interest with a radius of 14 Å, ACNs can lie between ~10 and 600 (Figure 2.1). It follows that within any one protein, many atomic contact numbers are only seen once. The packing density environments E_n will therefore only contain one atom. This effectively results in the atom being compared to itself for a B_{Damage} value, which, obviously, is meaningless.

Density metric, radius	SS	Min	Mean	Max	Var	SS	Min	Mean	Max	Var
Ooi, 8 Å	b	4	11.5	21	6.1	H	4	10.5	22	5.5
	C	1	9.7	22	8.5	I	4	12.4	20	6.9
	E	4	12.0	23	5.1	T	8	9.6	22	8.1
	G	4	9.5	21	7.1					
Ooi, 13 Å	b	10	43.4	72	89.5	H	8	42.9	80	103.6
	C	3	39.7	78	106.2	I	22	48.8	71	136.1
	E	11	47.0	84	93.6	T	5	38.6	82	106.8
	G	8	39.6	75	104.2					
Ooi, 14 Å	b	11	53.1	90	132.1	H	8	52.6	100	148.6
	C	3	48.9	95	153.9	I	29	59.9	82	211.8
	E	12	58.0	102	141.3	T	5	47.8	96	154.3
	G	9	49.1	96	150.4					
ACN, 8 Å	b	13	92.3	141	333.9	H	8	89.2	147	360.8
	C	0	81.1	145	444.8	I	22	98.8	141	337.2
	E	8	96.5	145	294.2	T	8	79.9	144	447.6
	G	8	82.2	141	424.6					
ACN, 13 Å	b	72	340.6	531	5,245.9	H	46	336.2	542	5,757.5
	C	11	313.7	541	6,032.7	I	158	385.7	517	7,705.1
	E	46	366.8	547	5,191.2	T	35	306.7	540	6,229.6
	G	48	314.5	539	6,074.3					
ACN, 13.5 Å	b	74	376.7	589	6,416.0	H	50	372.4	600	6,973.5
	C	14	348.1	597	7,293.6	I	173	428.4	570	9,721.7
	E	53	405.5	601	6,377.9	T	37	340.8	597	7,541.5
	G	51	349.3	602	7,341.6					
ACN, 14 Å	b	80	415.3	654	7,790.0	H	52	410.9	666	8,366.2
	C	16	385.1	658	8,757.2	I	190	474.0	633	12,043.5
	E	56	446.8	668	7,762.1	T	38	377.4	666	9,056.6
	G	57	386.4	659	8,809.8					

Table 2.4: Range, mean and variance of different packing density metrics (Ooi and ACN) with different radii applied to the 2,704 selected PDB structures. Both ACN and Ooi metrics are counts, so minimum and maximum values are integers. Mean and variance are given to 1 dp. The seven secondary structure (SS) labels are determined by STRIDE (Frishman & Argos, 1995): alpha helix (H; n=3,323,956), 3-10 helix (G; n=428,747), π-helix (I; n=1,515), extended conformation (E; n=2,351,831), isolated bridge (b; n=120,856), turn (T; n=1,859,115) and coil (C; n=1,679,478).

To resolve this problem of low sample numbers, it is necessary to combine neighbouring, and therefore similar, packing density environments. This can be achieved with a similarity function $sim(m) : \mathbb{R} \to \mathcal{P}(\mathbb{R})$, which for each packing density m returns a set of similar packing densities. These sets of packing density environments E_n are combined into larger sets of similar packing density environments $S_m := \bigcup_{n \in sim(m)} E_n$. Each similar packing density environment S_m therefore contains all atoms that have a packing density *similar* to m.

There are multiple ways to specify such a similarity function. One possible definition is 'similarity by binning', i.e. $sim(m) := \left\{ x \in \mathbb{R} \;\middle|\; \left\lfloor \frac{x}{z} \right\rfloor = \left\lfloor \frac{m}{z} \right\rfloor \right\}$ with a suitable bin size $z > 0$. In this case, with $z = 10$, atoms with contact numbers 145 and 153 would be treated as dissimilar. Similarity can also be defined by a fixed range, for example $sim(m) := \{x \in \mathbb{R} \mid m - \epsilon \leq x \leq m + \epsilon\}$, $\epsilon \geq 0$. Here, with a range of $\epsilon = 10$, atoms with contact numbers 145 and 153 would be treated as similar. Further possible definitions include flexible ranges and bin sizes, for example based on distribution quantiles.

For the results presented here, B_{Damage} was used with the former definition of similarity, based on fixed bins of size 10. This definition allows efficient assignment of atoms to similarity sets, as it can be carried out independently for each atom within the protein. A ribonuclease A structure (2BLP) with its atoms coloured by their respective packing environment is shown in Figure 2.2. During the evaluation of the B_{Damage} metric for this work, different bin sizes (for a 14 Å radius: 10 and 40) were tested as well as similarity by fixed ranges, as described above. No particular advantage or disadvantage was found when using either method.

Any similarity function and its parameters need to be chosen depending on the range and shape of the underlying packing density function: As the range of the packing density distribu-

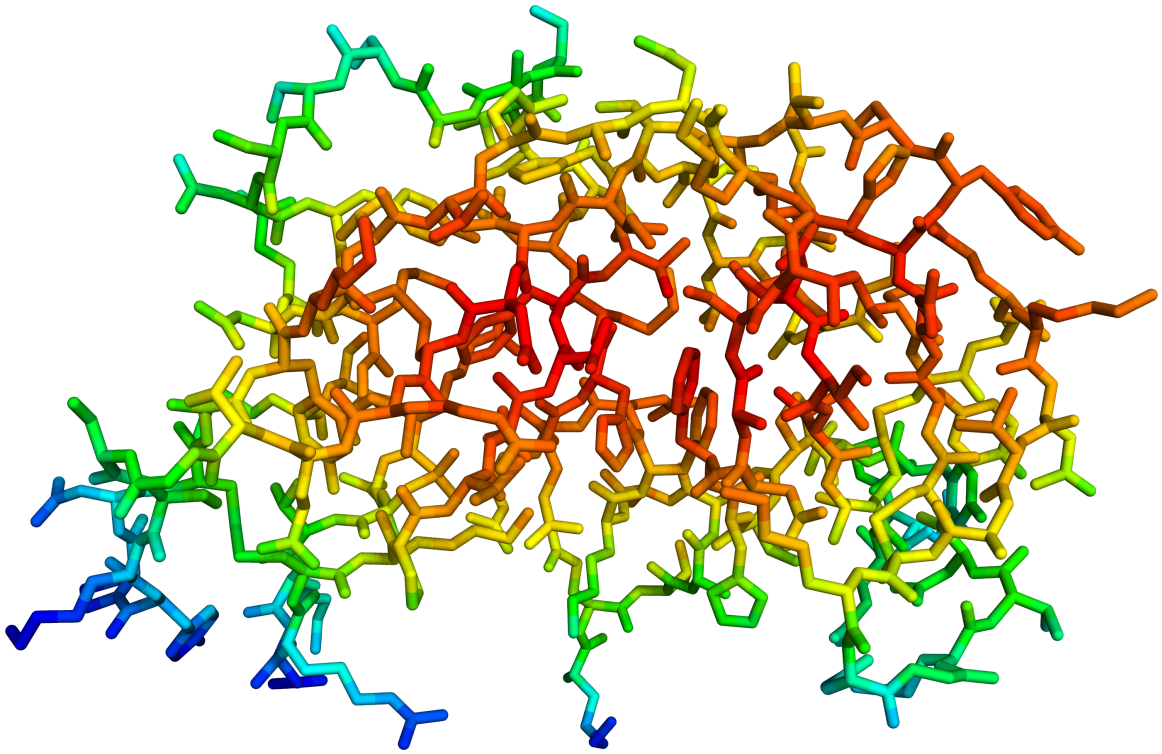


Figure 2.2: Structure model of ribonuclease A (2BLP) coloured by packing density. Regions with similar packing density, based on ACNs with a 14 Å sphere and a fixed bin size of 10, are shown in the same colour. Densely packed regions in the centre of the protein are shown in red, solvent-exposed surface regions in blue. Crystal contacts to neighbouring protein symmetry copies (not shown) are formed around the front and top right corner of the protein. These areas are thus shown to be in a high packing density environment.

tion increases (for example with an increasing radius of the sphere of interest), the bin size or range must increase as well. Otherwise the number of samples in each similar packing density environment S_m decreases, which may again lead to unreliable results.

2.3.3 Calculation of B_{Damage}

With the definition of *similar environments* established, B_{Damage} can now be calculated:

$$B_{\text{Damage}}(a) := \frac{B(a)}{\overline{B}(S_{P(a)})}$$

The B_{Damage} of any atom a is defined as its atomic B factor $B(a)$ divided by the average atomic B factor $(\overline{B})$ of all atoms in a similar environment (S) to that of a, based on its packing density $P(a)$.

Since atomic B factors are always positive[5], B_{Damage} is positive. If the similarity function used is symmetric, as both of the examples given above are, it follows from the definition of B_{Damage} that the average B_{Damage} over the entire protein is always 1.

2.3.4 Calculating B_{Damage} for structure models

For the work presented here, a large number of B_{Damage} values was calculated. Throughout the project over 4,000 PDB structures were analysed using many different definitions of B_{Damage}, involving different packing density metrics (Ooi numbers and ACNs, each with radii ranging from 4 to 14 Å) and similarity functions (both binning and fixed range with multiple parameter val-

[5] Some entries in the PDB do have negative or zero atomic B factors (Table 2.2), neither of which are realistic values. The latter implies an electron density singularity at a point in real space, which is impossible given thermal motion, even at 100 K, and the fact that structural models, if obtained by X-ray crystallography, result from a time and space average electron density distribution, and that electrons are, by their very nature, not stationary. Negative values imply either a lack of understanding of structural model parameters or the improper use of crystallographic refinement software, and definitely a lack of quality control of the resultant structural model.

ues). To facilitate an analysis on this scale, a software pipeline (Figure 2.3) centred around a relational database design was developed and implemented using MySQL (http://www.mysql.org/). Within the database, the structure models can be efficiently stored alongside the different B_{Damage} values and further physico-chemical annotations. Using a database instead of a file-based approach allowed the use of powerful database queries to run complex calculations involving many structures without the user having to spend a lot of time writing code involving file operations. Databases are designed for handling large amounts of data, which proved very useful for this project: The database scaled to a peak size of 406 GB without any noticeable performance degradation. The database was connected to a newly developed web service (Gerstel *et al.*, 2015; https://github.com/Garmangroup/RServer) which can generate publication-quality figures directly from the current state of the database using R script (R Development Core Team, 2011) and the R-MySQL module.

PDB structure models are available in three file formats: as plain text file, mmCIF and PDBML/XML. The newer file formats, mmCIF and PDBML/XML, hold data in a more structured way. However, in this project the plain text format files were used, as this is still the most widely supported format. These files adhere, in general, to a common file format specification. Unfortunately the standard is not particularly precise on the more intricate details, particularly the semantics of alternative conformations. Therefore some PDB files do contain custom annotations or idiosyncrasies, preventing the use of some programs which are otherwise perfectly capable of reading PDB files. To minimise problems with software reading PDB files, all files were processed with the program PDBCUR (Winn *et al.*, 2011) which was also set to remove hydrogen atoms, anisotropic B factor information (leaving the separately specified isotropic B factor) and zero occupancy atoms.

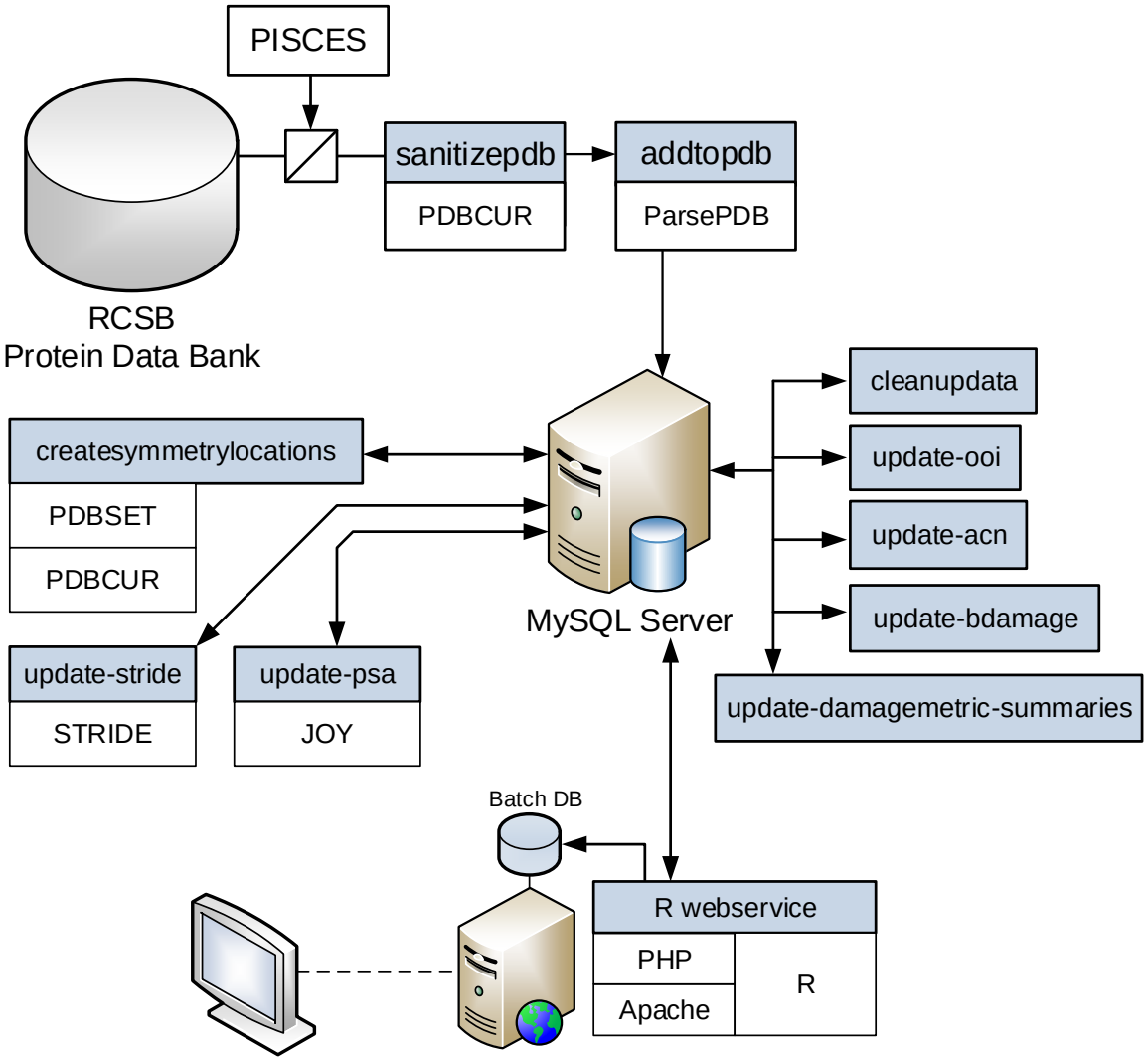


Figure 2.3: The software and data used to calculate B_{Damage} and correlations. Software marked on a blue background was developed specifically for this project. The database layout is shown in Appendix A (page 211).

PDBCUR replaced multiple conformations, where available, by the highest occupancy conformation. If this left more than one conformation, the first conformation was selected.

To populate the database, these cleaned PDB files were parsed (interpreted) using a modified version of ParsePDB (Bulheller & Hirst, 2009). During the project two mistakes were discovered in ParsePDB: when set to discard alternative atom locations ParsePDB would erroneously skip the atom if its alternative positions were not specified in order. ParsePDB was also unable to correctly read atomic B factors above 100 Å^2; it lost the first digit due to a mistake in the PDB format implementation. The authors were notified and provided with a fix in both cases.

The investigation in this project strictly concerns the radiation damage to proteins within a protein crystal. Metrics, such as packing density, therefore need to include protein atoms of neighbouring protein molecules, stacked against the original one, as shown, for example, in Figure 2.4. To this effect, protein structures were placed in their crystallographic context by calculating the position of all symmetry related atoms in a three-step process.

Firstly, the software PDBCUR (Winn *et al.*, 2011) was used to fill the unit cell with symmetry related protein copies (keyword GENUNIT). PDBCUR produced a possibly unexpected packing result, shown in Figure 2.5: each generated protein copy was placed correctly within the unit cell, however the generated copies were not placed within the same unit cell.

Secondly, to obtain the expected complete packing, the unit cell was translated to obtain the atomic coordinates for all non-hydrogen atoms within the 26 spatially neighbouring unit cells[6]. This step also ensured that relevant copies of the protein in neighbouring unit cells were taken into account.

[6]Imagine a perfect cube (unit cell) on a flat surface. 8 cubes of equal size can be placed around it so that each cube is in contact with the original one (at positions to the left, right, in front and behind, and at the diagonal positions). In three dimensions two additional layers of 9 cubes each can be placed above and below. $9 + 8 + 9 = 26$.

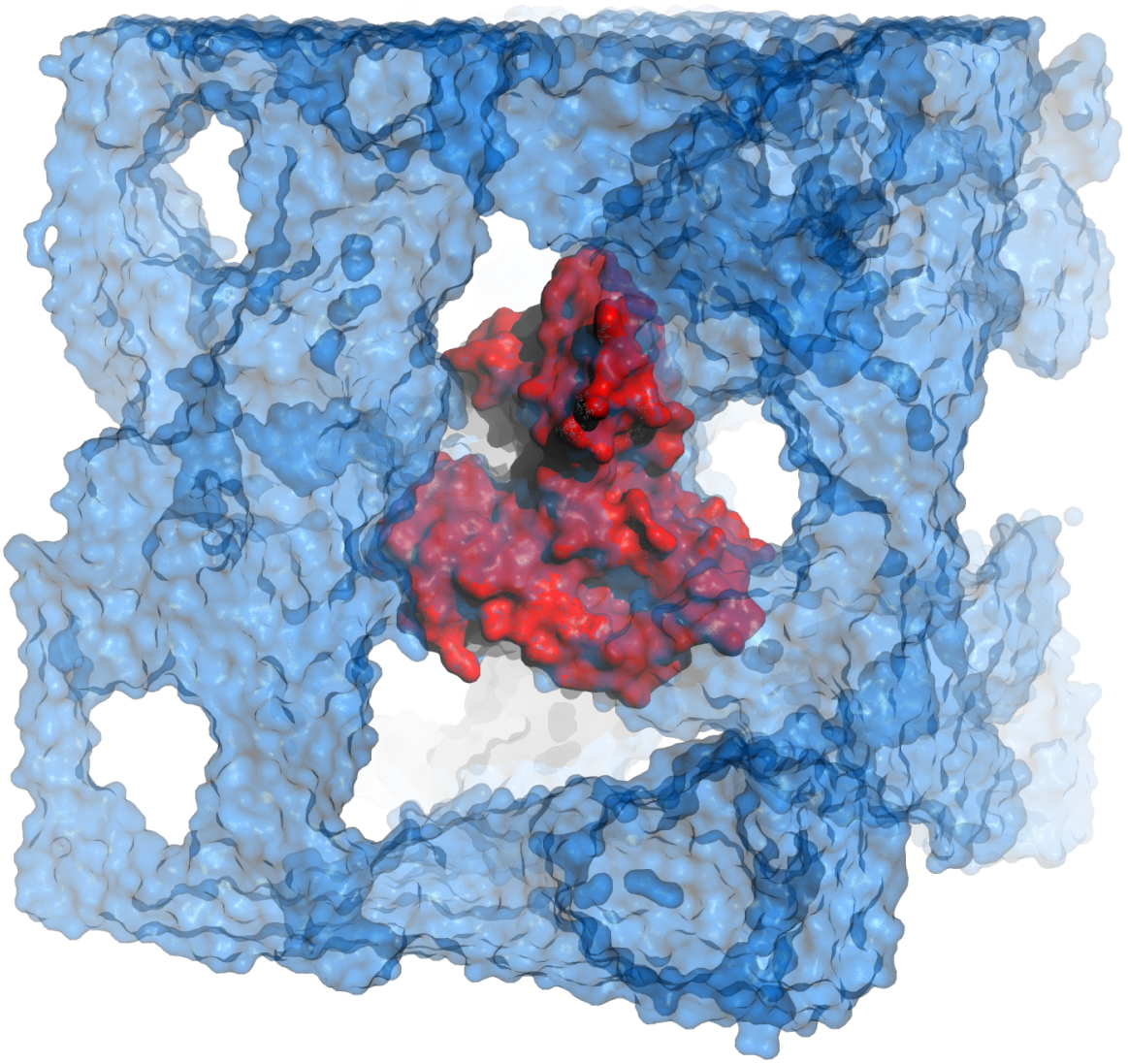


Figure 2.4: Arrangement of *Homo sapiens* RhoGDI protein (E154A, E155A mutant, PDB ID 2JHU; Cooper *et al.*, 2007) within a protein crystal. One complete protein copy is shown in red. A 14 Å bounding box (not shown) is drawn around the original protein, and all symmetry copies (blue) of the original protein lying within this box are identified. Solvent channels are visible in the resulting structure, appearing as carved out regions, and illustrate the sponge-like nature of protein crystals.
The coordinates of all blue atoms are stored in the database and can then be used to quickly calculate the ACNs for each red atom with varying radii of up to 14 Å.

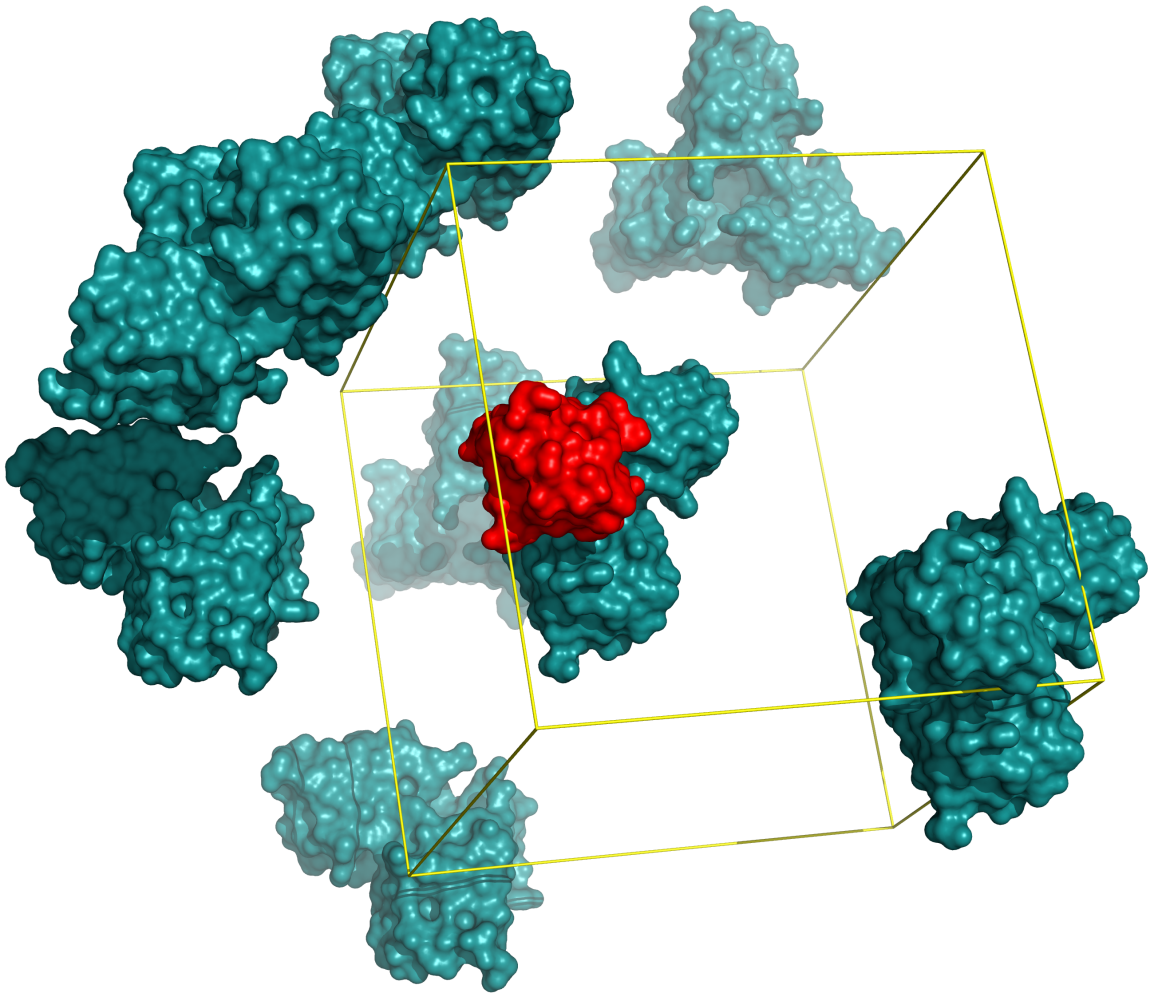


Figure 2.5: Unit cell arrangement of *Bos taurus* insulin protein (PDB ID 2BN3, space group $I\,2_1\,3$; Nanao *et al.*, 2005) as given by PDBCUR with the GENUNIT command. The original protein is shown in red, its copies in teal. Each protein copy is placed at the correct position, but not necessarily within the same unit cell (shown in yellow).

Thirdly, to reduce the number of atoms for the following calculations, a bounding box was placed around the original protein structure with PDBSET (Winn *et al.*, 2011) and extended by 14 Å (the selected maximum radius for any packing density calculation) in every direction. Only those identified symmetry-related atoms that lay within this bounding box were retained and stored in the database.

The correctness of these symmetry operations was manually verified on 5 PDB entries with different lattice types. Using PyMOL (http://www.pymol.org) the contents of the bounding box were visually inspected. Solvent channels should be visible as empty space surrounding a solidly stacked protein volume, as illustrated in Figure 2.4. The structure was inspected for clashes between original atoms and symmetry-related copies, as well as between symmetry-related copies themselves. As a second verification step, symmetry locations were obtained for a short sequence of the original protein. These locations were then visualised in WinCoot (Emsley *et al.*, 2010) and overlaid with the symmetry copies as calculated by WinCoot, Figure 2.6. The atom locations were then carefully inspected to ensure they matched up exactly.

With the symmetry atom locations established, the desired packing density metric (or metrics) and, finally, the B_{Damage} values can be calculated easily, following the method outlined above.

2.4 Using B_{Damage}

In this section B_{Damage} will be validated using a set of protein structure models obtained from low- and high-dose experiments. This set will be introduced in Section 2.4.1. Within that test set, the distribution of B_{Damage} values will be characterised and compared with the distribution of atomic B factors for entire protein structures as well as for selected known susceptible regions

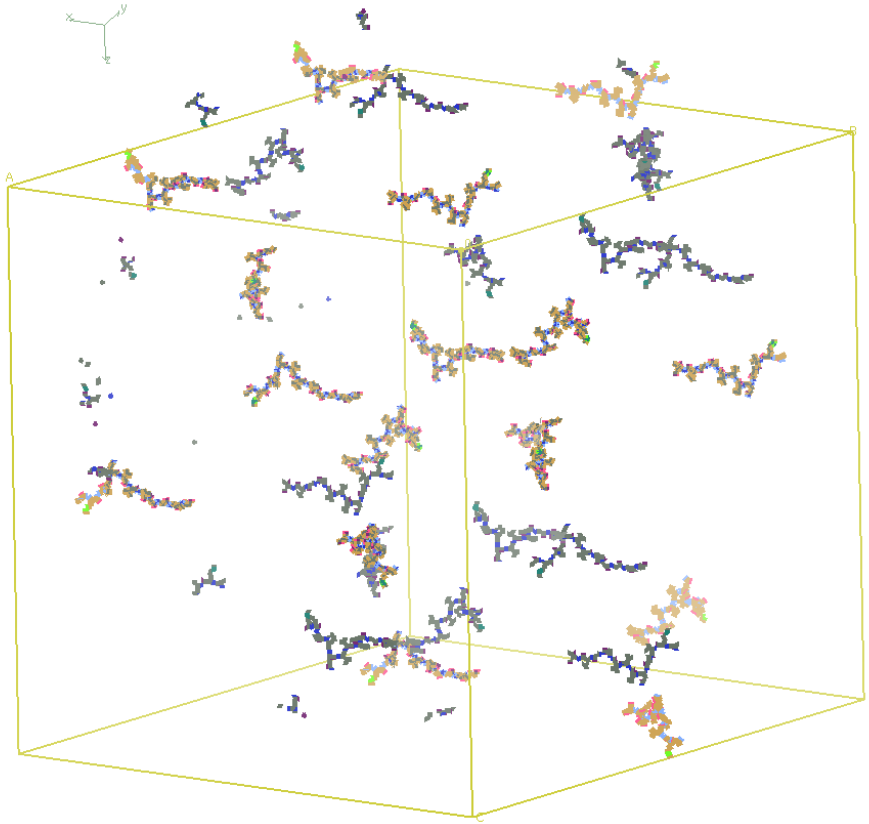


Figure 2.6: Validation of the calculation of the positions of symmetry-related copies of a short sequence of protein using WinCoot (Emsley *et al.*, 2010). The locations of the symmetry-related copies of the first twelve residues of a protein (here: *HOMO SAPIENS* RhoGDI mutant 2JHU, Cooper *et al.*, 2007) were calculated (brown) and compared with those calculated by WinCoot (blue) using its `Draw`/`Symmetry` feature. Visual inspection shows that the positions coincide perfectly.

(Section 2.4.2). The effects of a higher dose-difference in diffraction experiments on B_{Damage} will be explored using a pair of high dose-difference protein structure models and a pair with a low dose-difference (Section 2.4.3). The changes in B_{Damage} values between the low- and high-dose datasets will be investigated for specific amino acid residue atoms (Section 2.4.4). Finally correlations between B_{Damage} and solvent accessibility (Section 2.4.5), and B_{Damage} and packing density (Section 2.4.6) will be examined.

2.4.1 Selection of test-set proteins

To validate B_{Damage} it is necessary to test whether B_{Damage} values correlate with established findings regarding specific radiation damage. To this end, B_{Damage} was applied to a set of twelve deposited structures obtained from six different proteins by Nanao *et al.* (2005) at 100 K (Table 2.5). In the original work these structures were used for an investigation into radiation-damage induced phasing (RIP; Ravelli *et al.*, 2003). For each crystal, a complete low-dose (attenuated beam) 'before' dataset was obtained, followed by exposure of the crystal to the unattenuated X-ray beam ('burn') and the subsequent collection of a second low-dose complete 'after' dataset. With the exception of the two insulin datasets, all datasets were taken with a dose lower than that of the X-ray 'burn' phase (see Table 2.5). For five of the six proteins, the specific radiation damage induced by the unattenuated X-ray 'burn' caused sufficient movement or dispersion of cysteine disulphide sulphurs between the 'before' and 'after' dataset to allow *de novo* phasing by the RIP method. In the sixth case, ribonuclease A, the before and after structures were obtained by using a known RIP substructure.

The 'before' and 'after' datasets provide a useful low-dose/high-dose comparison for the six model proteins. The low-dose 'before' datasets are of course not zero-dose datasets, so it can

Protein	Elastase		Insulin		Lysozyme	
Dose	low	high	low	high	low	high
PDB ID	2BLO	2BLQ	2BN3	2BN1	2BLX	2BLY
Space group	P 2_1 2_1 2_1		I 2_1 3		P 4_3 2_1 2	
Unit-cell parameters (Å)	$a = 49.77$ $b = 57.62$ $c = 74.08$	$a = 49.79$ $b = 57.63$ $c = 74.09$	$a = 78.05$	$a = 78.08$	$a = 77.33$ $c = 38.16$	$a = 77.35$ $c = 38.19$
Dose (Gy) per data set	8×10^4	8×10^4	2×10^6	2×10^6	6×10^5	6×10^5
X-ray burn	2×10^6		2×10^6		2×10^6	
CYS residues	8		6		8	
# disulphides	4		3		4	
ASP/GLU res.	8 / 4		— / 4		7 / 2	
TYR/MET res.	11 / 2		4 / —		3 / 2	

Protein	Ribonuclease A		Thaumatin		Trypsin	
Dose	low	high	low	high	low	high
PDB ID	2BLP	2BLZ	2BLR	2BLU	2BLV	2BLW
Space group	P 3_2 2 1		P 4_1 2_1 2		P 2_1 2_1 2_1	
Unit-cell parameters (Å)	$a = 64.13$ $c = 63.64$	$a = 64.14$ $c = 63.74$	$a = 57.78$ $c = 150.08$	$a = 57.90$ $c = 150.36$	$a = 54.16$ $b = 58.25$ $c = 66.58$	$a = 54.18$ $b = 58.28$ $c = 66.60$
Dose (Gy) per data set	3×10^6	3×10^6	8×10^5	8×10^5	4×10^5	4×10^5
X-ray burn	4×10^6		2×10^6		2×10^6	
CYS residues	8		16		12	
# disulphides	4		8		6	
ASP/GLU res.	5 / 5		12 / 6		6 / 4	
TYR/MET res.	6 / 4		8 / 1		10 / 2	

Table 2.5: Low- and high-dose structure information for six crystals of different proteins from Nanao *et al.* (2005).

be expected that these data will give structures that already contain specific radiation damage. However, these datasets are sufficiently different to allow phasing with RIP. It is therefore possible to investigate the specific radiation damage development in further detail. In fact, this is one of the key reasons why, out of all the possible protein structure models (Table 2.1, page 31) the Nanao *et al.* (2005) set was selected to validate B_{Damage}: the set consists of six dissimilar proteins, each of which undergoes enough specific radiation damage to be processed by RIP, all data were processed using the same[7] protocols, using the same software, and by the same people, which should help to minimise systematic protein-to-protein variation and noise.

The expectation is that residues and atoms susceptible to early radiation damage show a clear increase in B_{Damage} values between the two datasets. Residues and atoms that are not affected by radiation damage by the end of the second dataset should show no increase in B_{Damage}. Similarly, no increase is expected for protein regions that are so radiation damage sensitive that damage is expected to have happened within the first dataset.

2.4.2 Distribution of B_{Damage} values

The atomic B factor and B_{Damage} distributions for the six low-dose and the six high-dose datasets are shown in Figure 2.7.

The effects of specific radiation damage on the distribution of B factors and B_{Damage} were explored by comparing the low-dose and high-dose Nanao *et al.* (2005) datasets using Welch two sample t-tests. The changes in the distributions for the entire structure, excluding hydrogen atoms, were subtle (Figure 2.7, top): B factors slightly shifted towards lower values for the high-dose dataset (not significant, $p = 0.188$). The overall distribution of B_{Damage} has, by definition, a

[7] in the case of ribonuclease A, reasonably similar

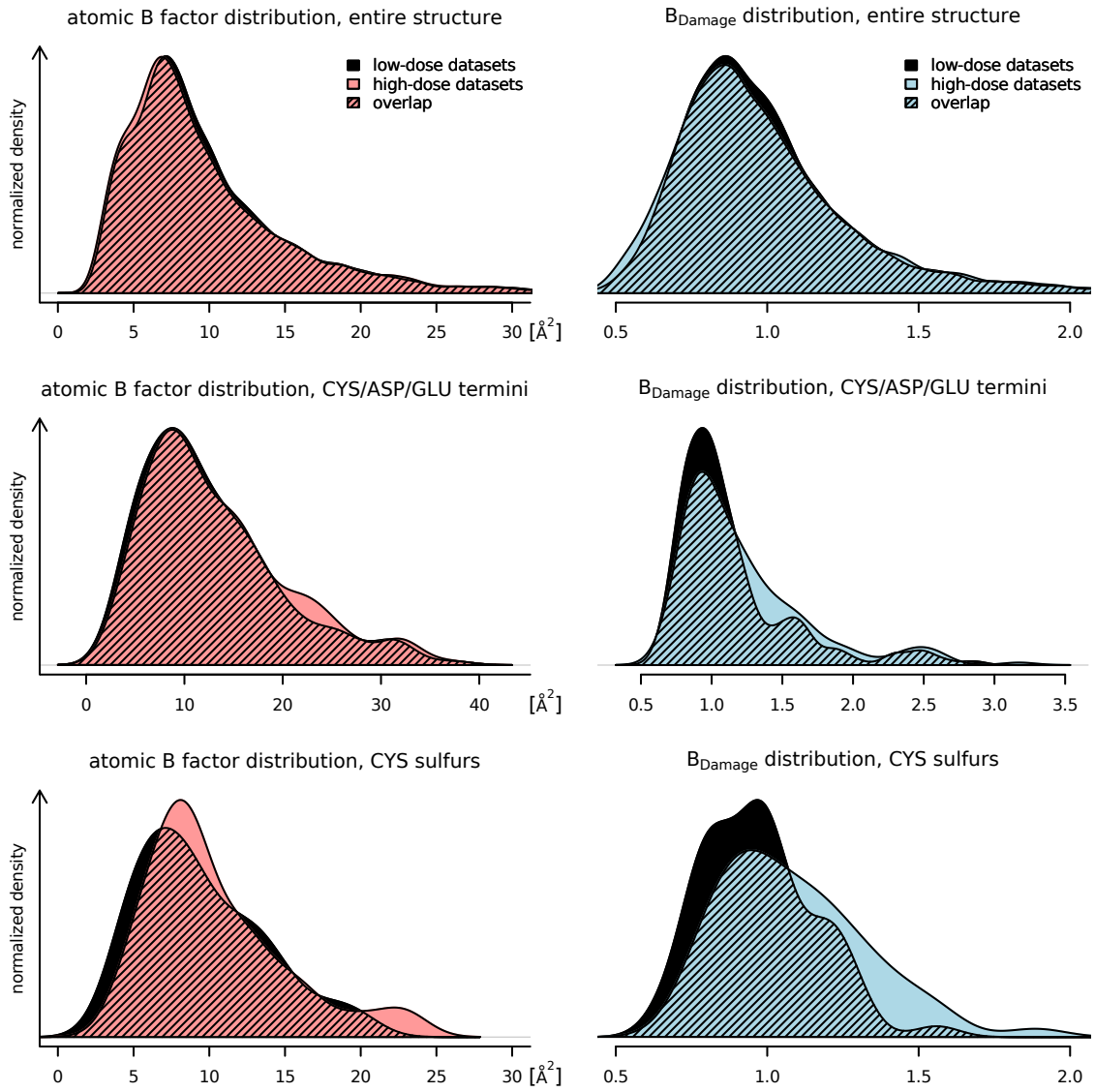


Figure 2.7: Distributions of atomic B factors (left) and the B_{Damage} metric (right) for the low-dose (black) and high-dose (colour) datasets of the six proteins from Nanao *et al.* (2005).
Top: The distribution for all atoms shows that both B factor and B_{Damage} have unimodal, slightly right-skewed distributions. B_{Damage} has a mode below unity and by definition a fixed mean of 1. The differences between the low- and high-dose datasets are minimal: Radiation damage causes the atomic B factors to decrease, and the spread of B_{Damage} to increase slightly.
Middle: Only B factors and B_{Damage} metrics of CYS S_γ, ASP O_δ, and GLU O_ϵ atoms are shown. These atoms are at the end of residue side-chains known to be susceptible to specific radiation damage. Both the atomic B factors and B_{Damage} show higher values for the high-dose datasets. The change in B_{Damage} is marked and separates specifically susceptible atoms from the bulk of the protein.
Bottom: These distributions only show cysteine sulphurs. The atomic B factors tend to increase, but do not behave consistently. B_{Damage} shows a consistent, strong shift towards higher values, indicating that the B factor of the sulphur atoms increases much faster than the B factor of other atoms in similar packing densities.
All distributions are smoothed using a kernel density estimator (R Development Core Team, 2011).

fixed mean at 1.0, thus no movement of the mean can be observed ($p = 1$). Using Levene's test (Brown & Forsythe, 1974) an increase in variance for the high-dose dataset and thus a 'flattening' of the distribution was seen ($p = 3 \times 10^{-4}$).

When considering only the side chain terminal atoms of amino acid residues known to be susceptible to specific radiation damage (CYS S_γ, ASP O_δ, and GLU O_ϵ atoms; Figure 2.7, middle) a clear difference in the change of distributions was apparent. A small increase in the number of atoms with B factors between 20 Å and 25 Å was visible, but no shift of the distribution could be observed ($p = 0.351$). For B_{Damage} there is some evidence for a shift towards higher values ($p = 0.058$), indicating that the increase in B factor for these atoms was larger than that expected from atoms in similar packing densities.

Looking only at cysteine sulphurs, an even clearer picture emerged (Figure 2.7, bottom). The atomic B factors tended to increase, yet the overall change in the distribution of B factors was insignificant ($p = 0.257$). The change in B_{Damage}, however, clearly showed that cysteine sulphurs were affected by the increasing dose ($p = 0.003$).

These results indicate that B_{Damage} is a sensitive metric that distinguishes between the more stable regions of the protein and those regions sustaining specific radiation damage.

2.4.3 Correlation of B_{Damage} with dose

If B_{Damage} is able to identify specific radiation damage, it should show larger changes for atoms which have been exposed to a higher dose. Here the effects of the X-ray burn on atomic B factors and B_{Damage} of the highly susceptible GLU O_ϵ atoms were compared. In the reference dataset, the ribonuclease A crystal was subjected to an X-ray burn, with an effective dose difference between the two datasets of 4 MGy, whereas the five other proteins only suffered an X-ray burn giving a

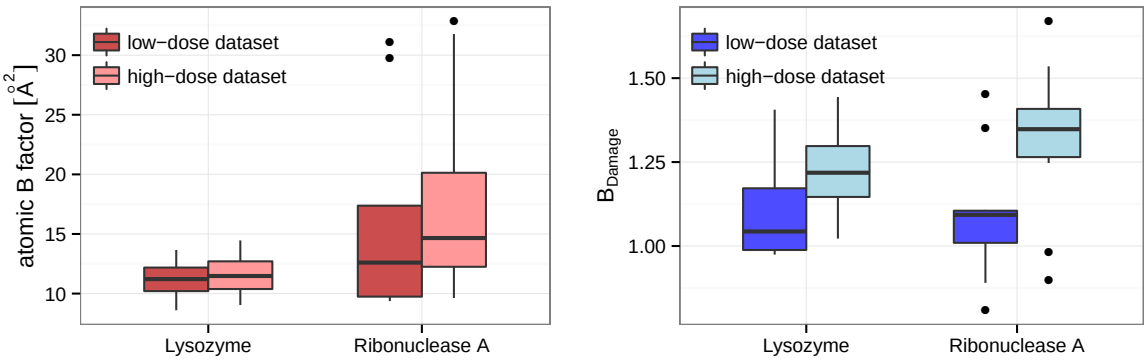


Figure 2.8: Dose-dependent changes in the distributions of atomic B factors and B_{Damage} values of GLU O_ϵ for two Nanao *et al.* (2005) set proteins.
Left: Lysozyme (left) was subjected to a 2 MGy X-ray burn, ribonuclease A (right) to a 4 MGy burn. The increase in B factors is barely visible and within a margin of error for lysozyme. The increase for ribonuclease A is more visible. The coloured box of the boxplot indicates the interquartile range. The black line represents the range, excluding outliers. The latter are shown as black dots.
Right: B_{Damage} shows a significant change in the GLU O_ϵ, indicating that the electron density around these atoms degrades more rapidly than for other comparable areas of the protein. The difference in B_{Damage} is larger for ribonuclease A (right), consistent with its absorption of a significantly higher dose.

difference in dose of 2 MGy (Table 2.5).

As can be seen in Figure 2.8, atomic B factors indicate no visible change for GLU O_ϵ atoms in lysozyme and little change in ribonuclease A. In contrast, the B_{Damage} metric achieves a much clearer separation between the low-dose and high-dose datasets, once again suggesting its reliability as a damage metric. Similar results can be observed for the ASP O_δ atoms (Figure 2.9).

ASN O_δ atoms can serve as a control. They occupy positions at a similar distance from the main chain as ASP O_δ atoms, but are not affected by radiation damage due to a comparative lack of charge and reactivity. No change in B_{Damage} is observed for ASN O_δ atoms. This indicates that B_{Damage} agrees with experimentally observed radiation damage behaviour.

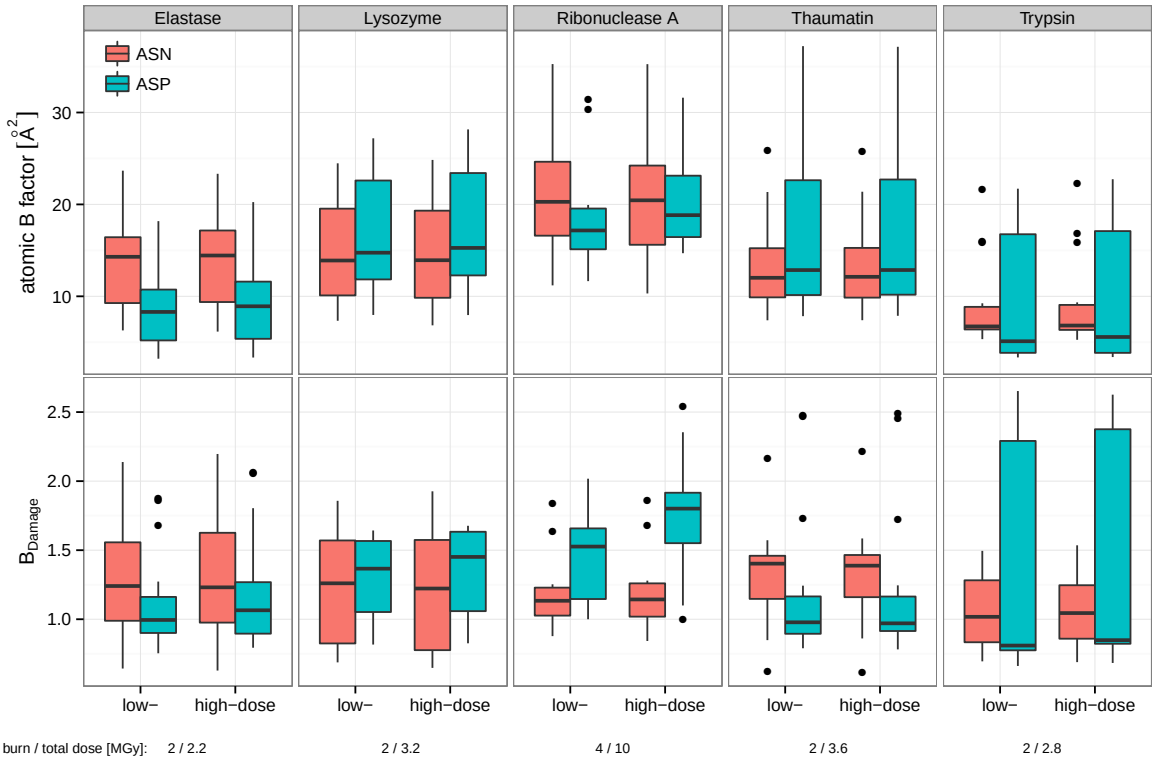


Figure 2.9: B factor **(above)** and B_{Damage} **(below)** of ASP and ASN O_δ atoms of five of the six Nanao *et al.* (2005) proteins in the low and high dose states. The deterioration of the ASP side chain termini of ribonuclease A cannot be observed from the B factor alone, but is indicated by B_{Damage}. The ASN oxygens remain unaffected. Insulin is not shown, as it does not contain ASP residues.

2.4.4 Differential B_{Damage} observations across proteins and residues

The differences in the median B_{Damage} for the terminal atoms of known specific radiation damage sensitive residues and control residues for all six protein models of (Nanao *et al.*, 2005) are shown in Figure 2.10. Since only ribonuclease A was subjected to a more intense X-ray burn, more signs of advanced specific radiation damage can be expected. As before, there is a distinct increase in B_{Damage} between the low- and high-dose state for the oxygens of the aspartic acid and glutamic acid carboxylates in the ribonuclease A model. This indicates that these groups suffered specific damage above the background rate of damage between the low- and high-dose dataset. The asparagine and glutamine residue oxygens show no significant change in B_{Damage}, which is consistent with the established knowledge that these amino acid residues only become susceptible to specific radiation damage later at higher doses (Juers & Weik, 2011). B_{Damage} also increases in the ribonuclease A model at methionine C_{ϵ}s, which only suffer bond cleavage at higher doses. The relevant median atomic B factors do not show this significant change for either aspartic or glutamic acid, or for methionine residue atoms (data not shown). The other five proteins in the Nanao *et al.* (2005) set do not show this marked B_{Damage} behaviour at aspartic acid, glutamic acid or methionine residues due to the lower absorbed doses of the 'burn'. With the exception of elastase, there is no change in B_{Damage} for tyrosine oxygens, indicating that at the present dose levels tyrosine residues have not yet suffered damage.

2.4.5 Correlation with solvent accessibility

Solvent accessibility information was obtained from the software PSA (Mizuguchi *et al.*, 1998), which calculates the relative solvent-accessible area of residues by rolling a water sphere with a

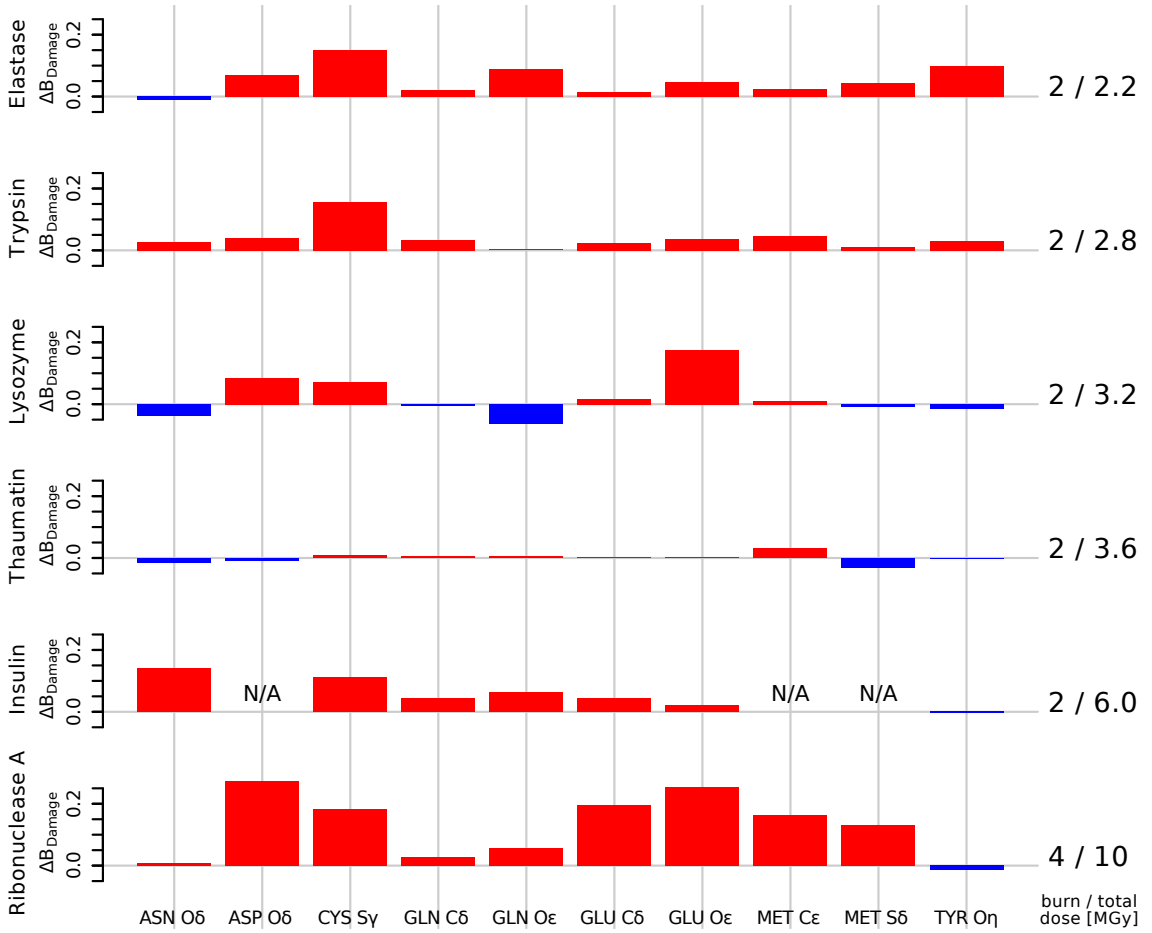


Figure 2.10: The change in the median B_{Damage} is shown for 10 groups of atoms of the six Nanao *et al.* (2005) protein structures before and after an X-ray burn. Insulin does not have any ASP or MET residues.

Different proteins show different B_{Damage} patterns. B_{Damage} tends to increase for most atoms shown, including ASP O_δ, CYS S_γ, GLU C_δ and O_ϵ, MET C_ϵ and S_δ. This indicates that these atoms damage faster than others within the same protein, and correlates well with the known specific damage decay pattern of amino acid residues.

The B_{Damage} increase in ASP O_δ is always stronger than that in ASN O_δ, which is not affected by specific radiation damage. Similarly, GLN residues are generally less affected than GLU residues.

With the possible exception of elastase, little change is observed in the TYR O_η hydroxyl group. This is not unexpected, since this group is only affected by specific radiation damage at higher doses.

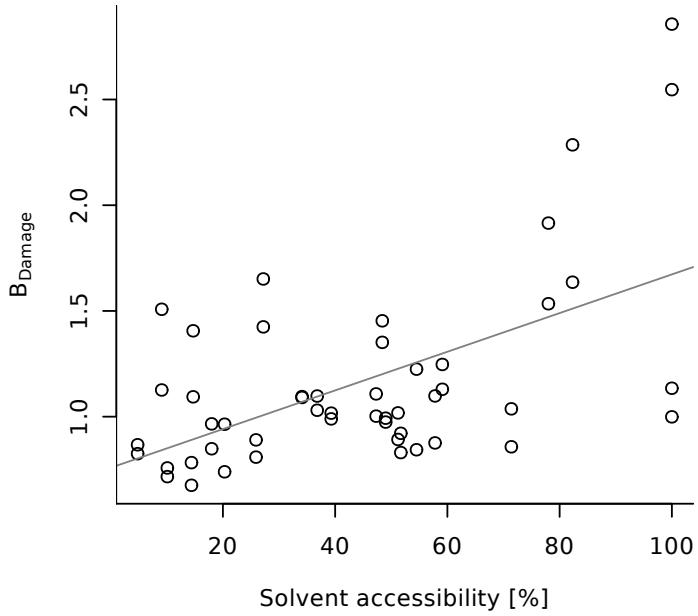


Figure 2.11: The B_{Damage} values of GLU O_ϵs of the six proteins in the Nanao *et al.* (2005) low-dose dataset are positively correlated with their solvent accessibility. The solvent accessibility is given as a percentage of the residue surface that is accessible to solvent as calculated by the program PSA (Mizuguchi *et al.*, 1998). A list of correlations between the solvent accessibility and other atoms across the datasets is given in Table 2.6.

radius of 1.4 Å over the van der Waals radii of the atoms. Correlations can be obtained for atoms of each residue by plotting B_{Damage} against the calculated solvent accessibility (Figure 2.11).

B_{Damage} has a uni-modal, slightly right-skewed distribution with its mode below 1 and a mean value of 1 (Figure 2.7, page 57). B_{Damage} cannot take negative values, yet its distribution can still be assumed to be approximately normal due to its small standard deviation, thus allowing the use of simple linear regression to analyse the relationship between B_{Damage} and physicochemical parameters of selected residues and atoms.

The low- and high-damage reference sets show significant correlation between the per-residue solvent accessibility and B_{Damage} for the entire protein as well as for whole ASP, CYS, GLU and TYR residues ($p < 0.001$). There is some evidence of correlation with the radiation damage prone termini of GLU (Figure 2.11), ASP, CYS, and TYR side chains, and MET residues (Table 2.6;

Residue / Atoms	Nanao *et al.* proteins low-dose	high-dose	2074 PDB structures
ARG N_ϵ	0.901*	0.915*	0.574*
ARG C_ζ	0.432	0.458	0.615*
ARG N_η	0.700*	0.710*	0.643*
ASP O_δ	0.670*	0.687*	0.259*
CYS S_γ	0.550*	0.676*	0.286*
MET S_δ	2.041	3.326	0.438*
MET C_ϵ	3.266	5.807	0.338*
GLN N_ϵ	1.445*	1.453*	0.432*
GLN O_ϵ	0.921*	0.989*	0.393*
GLU O_ϵ	0.914*	0.933*	0.309*
TRP N_ϵ	0.898*	1.124*	0.260*
TYR O_η	0.667*	0.681*	0.305*

Table 2.6: Slopes of simple linear regression models predicting B_{Damage} from solvent accessibility for selected atoms of the Nanao *et al.* (2005) datasets and the PDB dataset. The intercept values are not shown. Significant correlations with $p < 0.05$ are marked with an asterisk (linear correlation coefficient, $t(n-2)$ distribution).

$p < 0.05$). The data hint at a strong correlation for the MET C_ϵ, but this is not significant[8] ($p > 0.1$; $n = 11$).

2.4.6 Correlation with packing density

In contrast, the overall correlation between B_{Damage} and the ACNs of all non-hydrogen atoms is zero. This property directly follows from the definition of B_{Damage}, since it is a metric normalised by the average B factor of atoms with similar atomic contact numbers.

2.5 PDB investigations

The investigation described above into the relationships between B_{Damage} and physicochemical parameters of selected residues and atoms is fundamentally limited by the small selection of proteins considered. As previously shown, there is considerable variation in the effects measured on

[8] This is not a contradiction. In other words: there appears to be a large effect of solvent accessibility on the B_{Damage} values, however there is little evidence for this in the observed data.

single proteins (Figure 2.10, page 62). To obtain generalised conclusions on these relationships a much larger, representative sample of available protein structures must be considered. Only then might it be possible to identify statistically relevant relationships and radiation damage effects.

2.5.1 Selection of the PDB sample set

The PDB as the main source of protein structure models has been introduced in Section 2.2.1 (Page 32). It is neither reasonable nor desirable to use the entire PDB for this kind of statistical analysis for a number of reasons. Its large size, currently 102,720 structures[9], makes a full statistical analysis difficult and increases the time- and resource-requirements. A statistical analysis of all structure models would be strongly influenced by a small number of model proteins (e.g. lysozyme) that have a disproportionally large number of depositions. The PDB also includes depositions of structure models from wildly different backgrounds obtained by different experimental techniques. Extremely high resolution structures such as the 48-residue crambin at 0.48 Å (PDB ID: 3NIR) are stored alongside a 70 Å model of a 26-mer myosin crossbridge found in insect flight muscles (PDB ID: 1M8Q). A careful selection of an appropriate, representative subset of the PDB is required for a reliable statistical analysis based on PDB structures.

Using the advanced search function of the PDB website, a set of 11,836 PDB structures was identified based on the following filter criteria: all structures were to be solved by X-ray crystallography with a resolution between 1.5 Å and 1.8 Å, have a refined crystallographic R value $R_{work} \leq 20\%$ and contain at least one protein chain with a sequence length between 100 and 1,000 residues. This limitation to structures of similar resolution is of particular relevance when analysing atomic B factor distributions, as the error on atomic B factors is known to correlate

[9] reported on http://www.pdb.org as of the 19th August 2014

Protein structures		Steps in the selection process
11,836		Initial selection by PDB search
– 31	(0.3%)	Could not be processed by PDBCUR (Winn *et al.*, 2011)
– 23	(0.2%)	...by ParsePDB (Bulheller & Hirst, 2009) – File contains multiple models
– 6	(0.1%)	...by ParsePDB (Bulheller & Hirst, 2009) – File curation errors
– 4	(0.0%)	Curating errors regarding disulphide bond declarations
– 65	(0.5%)	Could not be processed by STRIDE (Frishman & Argos, 1995)
– 0	(0.0%)	...by JOY/PSA (Mizuguchi *et al.*, 1998)
= 11,707	(98.9%)	Structures successfully processed
2,704		Non-redundant structures culled by PISCES (Wang & Dunbrack, 2003)

Table 2.7: Number of PDB structures in different stages of the selection process

with resolution (Carugo & Argos, 1999, cited in Weiss, 2007). Although each PDB entry conforms to the standardised PDB file syntax, not every PDB entry can be processed without errors by every program. This is caused by semantic ambiguities in the file format standard. These allow the model builders a high degree of flexibility when describing their models, but can cause issues when the processing software tries to automatically interpret the files. To minimise the impact of these interpretation errors on the findings in this study, any PDB entry that could not be processed by any relevant software package was removed from the set (Table 2.7). To ensure that the results were not skewed due to the high multiplicity of certain model proteins in the PDB, a non-redundant list of proteins was culled (selected and retained) from the remaining 11,707 PDB entries using PISCES (Wang & Dunbrack, 2003) to apply a sequence identity cutoff of $\leq 25\%$. A final, non-redundant set of 2,704 PDB entries was thus obtained. This set is given in full in Appendix B (page 212).

2.5.2 Primary sequence

To investigate the effects of immediate sequence neighbours, B_{Damage} values can be stratified for each R_N-R_R-R_C residue triplet (the neighbouring residue R_N on the N-terminal side, the residue R_R

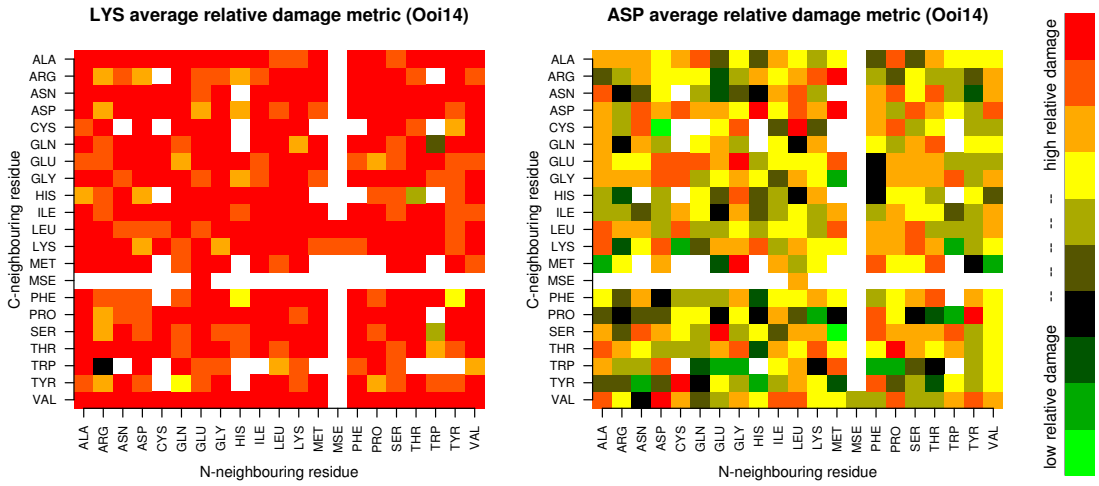


Figure 2.12: Mean B_{Damage} for primary sequence triplets in a selection of PDB proteins. Shown are early results for primary sequence triplets with LYS (**left**) and ASP (**centre**) as the central residue. For each triplet the decile of the mean B_{Damage} of all observations is shown colour-coded; red denotes very high, yellow high, black average and greens low deciles (**right**). Triplets with less than 20 observations are not shown. Note that for these results a variant of B_{Damage} based on Ooi numbers with a 14 Å radius was used.

itself, and the C-terminal side neighbour residue R_C).

For each residue the average B_{Damage} value for all its atoms was calculated. Results from identical R_N-R_R-R_C sequence neighbourhoods were again averaged together, and displayed in one colour-coded matrix for each amino acid R_R, with red colours for high and green for low B_{Damage} values. Example plots from a preliminary stage of the investigation[10] are shown in Figure 2.12. It is visible from these plots that the central residue has a large effect on the B_{Damage} value, e.g. the lysine plot (Figure 2.12, left) mostly consists of the red colour, indicating high B_{Damage} values, whereas the plot for aspartic acid (Figure 2.12, right) mostly features lower B_{Damage} values. The latter figure also indicates that B_{Damage} does not appear to behave symmetrically: For example aspartic acid in GLY-ASP-MET and GLY-ASP-TRP triplets has low B_{Damage} values, but relatively high values in MET-ASP-GLY and TRP-ASP-GLY triplets.

[10] For which a variant of B_{Damage} based on Ooi numbers with a 14 Å radius was used on a smaller selection of PDB entries.

It became clear that the identification of any patterns from these data would be very difficult. This is partly due to the fact that, based on 20 amino acids (discounting selenomethionine), $20 \times 20 \times 20 = 8,000$ distinct amino acid triplets can be observed. Using a standard $p < 0.05$ hypothesis test, one would therefore expect the identification of a significant effect in 400 triplets, when, in fact, there may not be any effect present (type I error, false positives). Ideally only parts of the residue such as the side chain termini should be considered, further restricting the number of observations and thus increasing noise.

Even with a large set of PDB entries it is therefore necessary to start the kind of investigation carried out here with a small number of reasonable hypotheses obtained by other means. Research into the effects of primary sequence neighbours was therefore postponed until such hypotheses become available.

2.5.3 Secondary structure

Secondary structure information was obtained using STRIDE (Frishman & Argos, 1995). STRIDE assigns one of seven secondary structure motifs to each residue, depending on the hydrogen bond energy and backbone torsion angle information. 33.2% of all residues within the PDB subset were marked alpha helix, 23.4% extended conformation, 19.9% turn, 17.9% coil, 4.3% 3-10 helix, 1.2% isolated bridge, and 0.016% π-helix.

Applying Levene's test with median centres (Brown & Forsythe, 1974) for equal variance on a sample of 1,500 atoms from each of the seven secondary structure labels indicates that the seven distributions have different variances (heteroscedasticity; $p < 1.1 \times 10^{-7}$). However, visual inspection of the distribution of B_{Damage} values in the PDB subset indicates that, with the exception of π-helix-regions, B_{Damage} is independent of the secondary structure as determined by

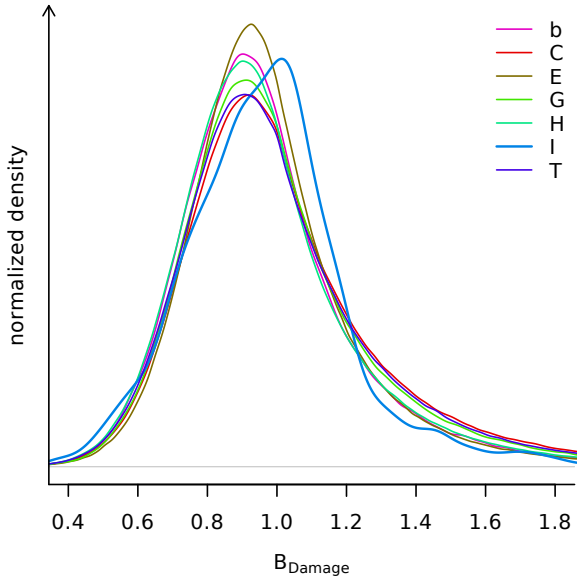


Figure 2.13: Distribution of B_{Damage} values per protein secondary structure motif. The seven secondary structure labels are: alpha helix (H), 3-10 helix (G), π-helix (I), extended conformation (E), isolated bridge (b), turn (T) and coil (C, none of the above). Distributions have been smoothed using a kernel density estimator (R Development Core Team, 2011).

STRIDE (Figure 2.13). It is possible that the different π-helix region behaviour may be caused by sampling effects: these regions are much rarer than any other secondary structure and account for only 0.016% of all residues in the PDB subset.

From these results it can be concluded that the secondary protein structure in general is not a relevant factor in the distribution of specific radiation damage.

2.5.4 Disulphide bond types

In many MX studies, see Garman (2010) for a review, it has been observed that disulphides are most susceptible to damage, so these were the first type of bond to be investigated. A preliminary investigation by Hogg & Ravelli (unpublished) suggested that there might be differences in susceptibility to radiation damage between the different disulphide bond groups. According to the signs of the five dihedral angles (Figure 2.14) a disulphide bond can be classified as one of three

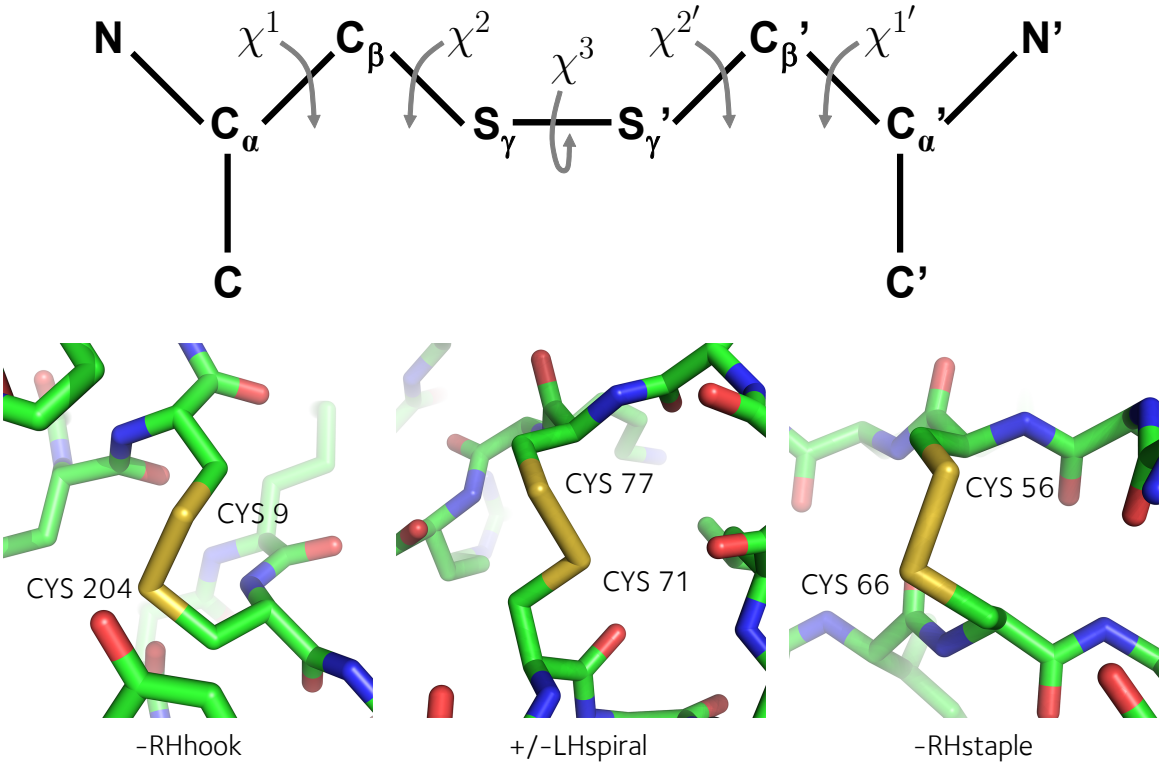


Figure 2.14: Classification of CYS-CYS disulphide bonds according to dihedral bond angles.
Top: A disulphide bond can be described by 5 dihedral angles (Table 2.8), which can be defined from either side.
Bottom: The three basic groups, spiral, hook and staple are determined by the three central angles χ^2, χ^3 and $\chi^{2'}$ and are named after the motifs they resemble. From each group one disulphide bond is shown as an example (Thaumatin, PDB ID 2BLR, Nanao *et al.*, 2005). By combining its basic group with its handedness and remaining angles, a disulphide bond can be classified as one of 20 bond types (Table 2.8).

different groups: spiral, hook or staple (Table 2.8). Together with their handedness and the signs of the χ^1 and $\chi^{1'}$ angles, this information can be used to label each disulphide bond as one out of a total of 20 different disulphide bond types (Schmidt *et al.*, 2006).

These disulphide bond types are closely related to the function of the bond within the protein: catalytic disulphide bonds, which can mediate chemical reactions, usually have a +/-RHHook arrangement (Schmidt *et al.*, 2006). All protein function regulating (allosteric) disulphide bonds are -RHStaple type bonds (Schmidt *et al.*, 2006). The remaining disulphide bonds are considered to

sign of dihedral angle χ^1	χ^2	χ^3	$\chi^{2'}$	$\chi^{1'}$	handedness & group	disulphide bond type
–				–	left handed Spiral	-LHSpiral
–				+	left handed Spiral	+/-LHSpiral
+	–	–	–	–	left handed Spiral	+/-LHSpiral
+				+	left handed Spiral	+LHSpiral
–				–	left handed Hook	-LHHook
–				+	left handed Hook	-/+LHHook
+	–	–	+	–	left handed Hook	+/-LHHook
+				+	left handed Hook	+LHHook
–				–	right handed Staple	-RHStaple
–				+	right handed Staple	+/-RHStaple
+	–	+	–	–	right handed Staple	+/-RHStaple
+				+	right handed Staple	+RHStaple
–				–	right handed Hook	-RHHook
–				+	right handed Hook	-/+RHHook
+	–	+	+	–	right handed Hook	+/-RHHook
+				+	right handed Hook	+RHHook
–				–	left handed Hook	-LHHook
–				+	left handed Hook	+/-LHHook
+	+	–	–	–	left handed Hook	-/+LHHook
+				+	left handed Hook	+LHHook
–				–	left handed Staple	-LHStaple
–				+	left handed Staple	+/-LHStaple
+	+	–	+	–	left handed Staple	+/-LHStaple
+				+	left handed Staple	+LHStaple
–				–	right handed Hook	-RHHook
–				+	right handed Hook	+/-RHHook
+	+	+	–	–	right handed Hook	-/+RHHook
+				+	right handed Hook	+RHHook
–				–	right handed Spiral	-RHSpiral
–				+	right handed Spiral	+/-RHSpiral
+	+	+	+	–	right handed Spiral	+/-RHSpiral
+				+	right handed Spiral	+RHSpiral

Table 2.8: Disulphide bond categorisation by the sign of the dihedral angles (Schmidt *et al.*, 2006).

The fundamental bond type is defined by the 3 centre angles χ^2, χ^3 and $\chi^{2'}$. Due to symmetry, the 8 possible combinations of angle signs reduce to 3 fundamental bond groups (spiral, hook, and staple) together with a left or right handedness. The handedness is determined by the sign of the central angle χ^3.

The group combined with the handedness and the signs of the outermost angles χ^1 and $\chi^{1'}$ results in a total of 20 different disulphide bond types.

The definitions of the disulphide angles are shown in Figure 2.14.

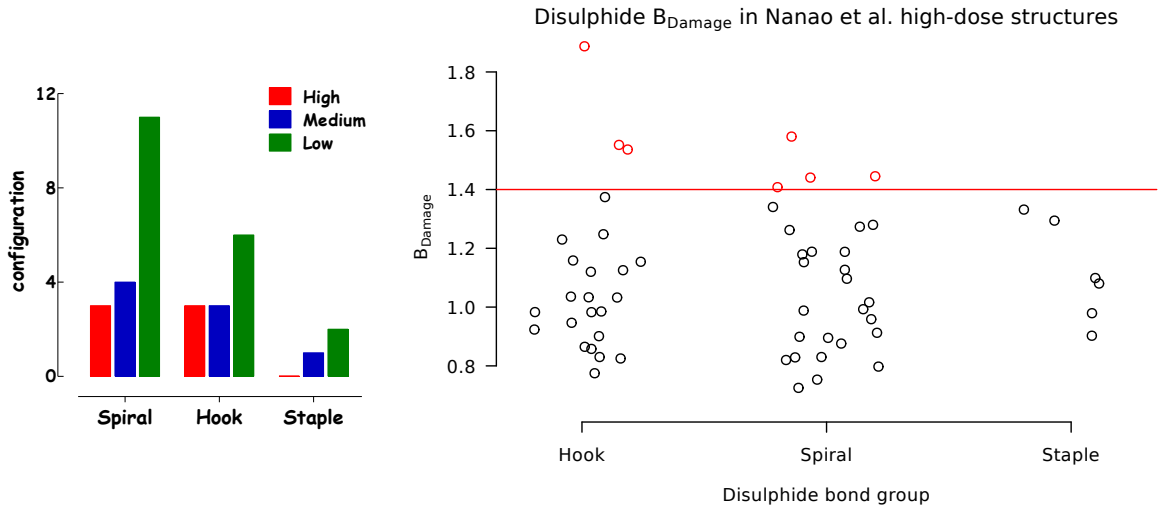


Figure 2.15: **Left:** Figure from Hogg & Ravelli (unpublished) on the susceptibility of sulphurs in different disulphide bond groups for the 6 Nanao *et al.* (2005) proteins and a 7th protein, identified by manual inspection. Hook and spiral disulphide bond groups (Figure 2.14) are found to be the most susceptible. **Right:** Distribution of B_{Damage} for sulphurs in different disulphide bond groups for the Nanao *et al.* (2005) high-dose dataset. The most damaged disulphide bonds, shown in red, again belong to the hook or spiral group.

mainly stabilise the protein structure.

Hogg & Ravelli (unpublished) investigated the disulphide radiation damage susceptibility of the six Nanao *et al.* (2005) protein structures and a seventh protein (*TORPEDO CALIFORNICA* acetylcholinesterase; Weik *et al.*, 2000). They ranked the disulphides according to observations of specific damage and found that the more damaged disulphide bonds belonged to the spiral and hook group (see Figure 2.15, left). A similar pattern can be found by plotting the B_{Damage} values of the disulphide sulphurs of the six Nanao *et al.* (2005) high-dose protein structures stratified by their disulphide bond group (see Figure 2.15, right).

Using PERL scripts, all disulphide bonds in the 2,704 selected PDB proteins were identified in the database. To this effect the dihedral angles that make up the disulphide bond were calculated from the known atomic positions. In total 2,964 disulphide bonds were found, specifically 1,070 hook, 1,424 spiral, and 470 staple type bonds.

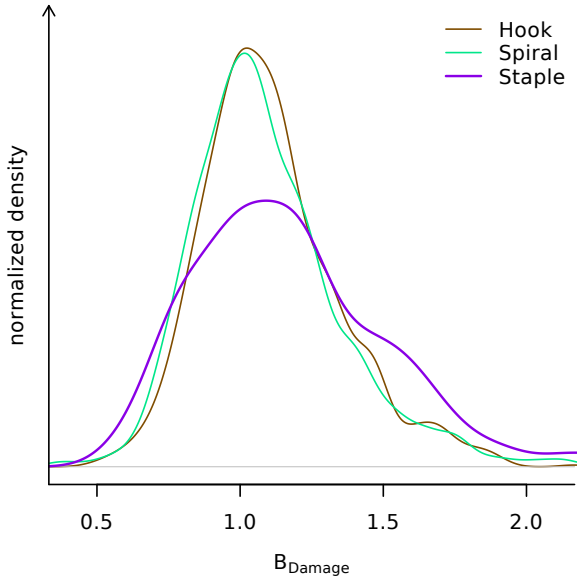


Figure 2.16: Distribution of B_{Damage} values of disulphide sulphurs stratified by disulphide bond group. Distributions have been smoothed using a kernel density estimator.

The three different basic disulphide bond groups show different B_{Damage} distributions in the PDB subset (Figure 2.16). The B_{Damage} distribution of staple disulphide bonds deviates from the other two, indicating there is an underlying difference between them. The heteroscedasticity of the B_{Damage} distributions can be confirmed using Levene's test with median centres (Brown & Forsythe, 1974) ($p < 3 \times 10^{-16}$).

While examination with ANOVA is precluded by the heteroscedasticity, the fact that the different disulphide bond groups show a range of variances indicates that there may be a systematic difference in the damage behaviour of the different disulphide bond groups. Staple disulphide bonds show a tendency towards higher B_{Damage} values. Further analysis of the 20 disulphide types is limited by the large difference in sample sizes (e.g. only two disulphide bonds of type +LHStaple are observed, but 790 of type -LHSpiral).

In contrast to the B_{Damage} values, the three corresponding distributions of the atomic B factors do look similar, and are not heteroscedastic ($p > 0.05$, not shown). This indicates that the

difference between the three disulphide bond types is not manifested simply by an intrinsic bond property, such as a (possible) increased flexibility of staple type bonds. The difference only becomes measurable in B_{Damage}.

2.5.5 Solvent accessibility

As before (Section 2.4.5, page 61), the solvent accessibility was determined for each residue, and the correlations between solvent accessibility and B_{Damage} was investigated.

For the PDB subset, strong evidence for correlations between the residue solvent accessibility and B_{Damage} for all ASP, CYS, GLU, MET, and TYR residues and their side chain termini was found (Table 2.6, page 64). However, the observed correlations are consistently weaker in the PDB subset than in either of the low- or high-dose Nanao *et al.* (2005) sets.

2.6 Discussion

The data presented here show an evaluation of the B_{Damage} metric using the six pairs of datasets from Nanao *et al.* (2005) and an investigation into B_{Damage} correlations in a broad sample of deposited PDB structure models.

2.6.1 Nanao dataset

Atomic B factors contain a strong packing density dependent component, which dominates information useful for assessing specific damage. B_{Damage} highlights areas of specific damage, which are not visible from using atomic B factors alone. Specifically damaged residues, such as the glutamic acid residues in lysozyme and ribonuclease A, can be identified. The preferential damage between residues of the same type can be compared using B_{Damage} values. The effectiveness of

the B_{Damage} metric was shown by applying it to a set of proteins (Nanao *et al.*, 2005) and comparing the results to expected and established protein specific damage patterns. The applicability of B_{Damage} was demonstrated by investigating the changes in specific damage with increasing dose.

Care must be taken not to generalise the relationship of B_{Damage} and dose from the observations on this small set of proteins. While all six of the Nanao *et al.* (2005) proteins were subjected to different doses, only the ribonuclease A protein crystal was subjected to a more intense X-ray burn. Aside from this being only a single sample, Nanao *et al.* (2005) indicate that the specific damage pattern of ribonuclease A may be atypical. They found that its disulphide bonds were not the most susceptible sites within the protein, but rather the most susceptible residues were found on the protein surface and play a role in crystal packing. This, however, is consistent with the observation that the exact decay patterns vary considerably between different proteins.

2.6.2 PDB dataset

B_{Damage} provides additional new insight over atomic B factors for the representative PDB subset: different disulphide bond groups show different B_{Damage} distributions, indicating that staple disulphide bonds may be preferentially sensitive to specific radiation damage, or may even undergo a different specific damage process from other disulphide bonds.

In contrast, the protein secondary structure does not significantly affect the distribution of B_{Damage}. This implies that the secondary structure may not be a relevant factor in preferential specific damage processes.

The PDB dataset does provide strong evidence for a positive correlation between B_{Damage} and solvent accessibility for all the residues and residue termini known to be particularly susceptible to specific damage. This is consistent with the results of Petrova *et al.* (2010) as well as

Warkentin *et al.* (2012b). These groups based their damage metric on atomic B factors or, the intrinsically linked, refined occupancy values with the atomic B factors held at a fixed value. These results are in contrast to those of Fioravanti *et al.* (2007) who used peaks in the electron density difference map to identify sites of damage, and Homer *et al.* (2011) who investigated real space electron density decay in subsequent data sets. This indicates that the difference between methods for defining 'damage' could significantly affect the conclusions that are drawn. From these results it can be concluded that B_{Damage} appears to be a useful indicator of the radiation damage susceptibility of residues.

2.6.3 Metric stability under parameter variation

To test the stability of the B_{Damage} metric and the reliability of the results, all the tests presented in this chapter, including those on the Nanao *et al.* (2005) datasets, were repeated with atomic contact number radii of 8 Å, 13 Å and 13.5 Å (data not shown), leading to comparable outcomes.

2.7 B_{Damage} – a new metric

In this chapter the theoretical background of B_{Damage} was presented. B_{Damage} was then validated using a before/after-set of protein structures with known radiation damage progression. Finally it was used on a large set of protein structure models to statistically analyse the properties of specific radiation damage. The programs and the database layout used for calculating B_{Damage} are available from the author upon request.

B_{Damage} is a new metric that can be calculated for a single, refined PDB structure model and provides per-atom information on relative specific damage. B_{Damage} highlights areas of the protein

where radiation mediated structural changes may have taken place. It can be used by a crystallographer inspecting a single structure for such changes, as well as for statistical investigations providing insight into the causes of preferential specific radiation damage decay, and may ultimately aid the understanding of specific damage mechanisms.

B_{Damage} is particularly useful for large scale statistical investigations, as it does not depend on human intervention or classification of damage levels, for example via the observation of electron densities, and only requires a single protein structure coordinate model.

Since B_{Damage} can be calculated from a single PDB file it can be used by the crystallographer during and after refinement to identify candidate locations for radiation damage mediated structural changes in the protein.

The automated identification and quantification of specific radiation damage over a number of PDB structures allows the statistical investigation of preferential specific damage, allowing the use of PDB models that previously were not suitable for radiation damage research. Ultimately it may be possible to use a structure-wide B_{Damage} derivate as an overall quality indicator for specific damage, which could be used alongside other established tests during structure deposition (Read *et al.*, 2011) to ensure the accuracy and veracity of published protein structure models.

2.7.1 Limitations

B_{Damage} depends, by its nature, on high quality PDB structure models. It should only be applied to PDB files containing atomic B factors that were refined per-atom. In its given definition B_{Damage} does not allow for the use of multiple models and partial occupancies. Unmodeled regions of the protein are particularly problematic, as they affect both the average atomic B factor calculations and the packing density calculations. B_{Damage} does not currently make use of the information

available in anisotropic atomic B factors. A protein-wide variant of B_{Damage}, rather than its current description at an atomic level, might be used to indicate the level of structural damage sustained during the diffraction experiment, and could be used as a quality control measure upon PDB structure deposition.

2.7.2 Future investigations

The PDB survey did not take into account that the diffraction data underlying the PDB structure models were taken at different experimental temperatures. Of the 2,704 selected structure models, 2,123 (78.5%) were obtained using diffraction data collected at cryotemperatures (100 K), 61 (2.3%) at room temperature ($\leq$ 123 K), and 48 (1.8%) at temperatures inbetween. The remaining 472 (17.5%) structure models had either no indication of experimental temperature or specified clearly unrealistic values.

There are also a number of possible areas where B_{Damage} can be refined. One possible example would be an atom type normalisation, in which the B_{Damage} value of each atom would be normalised against a global or protein-wide average B factor of atoms of this type. This would counter the tendency of, for example, sulphur atoms to exhibit higher B_{Damage} values due to their higher atomic B factor.

An atom-specific normalisation would also allow direct comparisons of B_{Damage} values between atoms. It may then be possible to investigate high B_{Damage} peaks within a number of structures to see how well B_{Damage} correlates with visible structural artefacts.

Chapter 3

Dose estimation using RADDOSE-3D

Dose estimation using **RADDOSE-3D**

This chapter describes the development of the software RADDOSE-3D, its use case and design rationale.

RADDOSE-3D was initially developed in collaboration with Oliver Zeldin. I designed the overall code framework, the object hierarchy, and the object interaction model including the support for different crystal and beam types, input and output routines, and elements required for batch processing in automated environments such as the RADDOSE-3D web service.

In this chapter, the concept of dose and the need for dose estimation software in macromolecular crystallography (MX) are established and the existing versions of the RADDOSE software are described, including their shortcomings and the reasons for developing RADDOSE-3D.

Section 3.2 describes the RADDOSE-3D architecture from a more technical perspective and explains underlying design decisions.

In the following Section, 3.3, an account of the development process and performance optimisation carried out on RADDOSE-3D is given, which ensure that RADDOSE-3D provides fast and accurate results.

Section 3.4 details the current limitations of RADDOSE-3D and provides ideas for future extensions and possible uses of RADDOSE-3D.

The final Section, 3.5, describes how RADDOSE-3D can be obtained. Users running many simulations can download RADDOSE-3D to their machines. Occasional users may find the RADDOSE-3D web interface useful. Professional users can compile their own versions of RADDOSE-3D, and extend and modify it for their own purposes.

3.1 Dose estimation in macromolecular X-ray crystallography

The metric central to any crystallographic experiment involving radiation damage is dose. Dose, measured in gray (1 Gy = 1 J/kg), describes the energy absorbed by a crystal. The deposition of energy is mainly caused by the absorption of incident X-rays via the photoelectric effect (Figure 1.3, page 7), and the partial absorption of energy via inelastic scattering events (Compton scattering, Figure 1.2, page 6). These events can both introduce a systematic bias in the protein conformation within the crystal (specific radiation damage) and destroy the long-range crystal order and therefore cause the decay of the diffraction pattern (global radiation damage). In contrast to 'exposure time' or 'diffraction images taken' dose is a transferable unit, allowing the direct comparison of experimental results across different beamlines and environments (Blake & Phillips, 1962). Dose therefore forms the baseline against which all radiation damage processes should be described.

There are two dose limits used as guidelines in MX diffraction experiments: The Henderson limit, 2×10^7 Gy, was predicted by Henderson (1990) using known limits from electron microscopy. According to Henderson, the total intensity of the diffraction pattern of a protein crystal would be expected to be halved below or by this dose.

The second limit, the Garman limit, 3×10^7 Gy (Owen *et al.*, 2006), indicates that diffraction data collected beyond it is likely to result in protein structures with compromised biological fidelity. This limit was derived from experimental X-ray damage observations at 100 K and the insight that below a total diffraction intensity of $0.7 \times I_0$ (I_0 being the total initial diffraction intensity) the electron density maps show significant damage.

Predominant **crystal** parameters affecting dose absorption and distribution	
composition	The presence of heavy atoms disproportionately increases the X-ray absorption cross section and therefore leads to an increase in dose (Murray & Garman, 2002; Holton, 2009). Packing within the unit cell determines the crystal solvent content and therefore the concentration of protein- or ligand-bound heavy atoms.
shape & size	For crystals larger than the beam, parts of the crystal are not exposed. Conversely, for crystals smaller than the beam, parts of the beam do not intersect the crystal. With very small or thin crystals, fluorescent photons may escape from the crystal, reducing the absorbed dose. The X-ray beam is attenuated by the crystal, leading to an uneven dose distribution within the crystal along its depth.
solvent	The solvent composition affects X-ray absorption. When crystals are mounted on a loop, they are held in place by the surface tension of mother liquor within the loop. The amount of mother liquor may vary and cause the enclosure of the crystal in a blob of mother liquor with varying depth, affecting local beam attenuation.
temperature	Experiments at cryotemperature may be affected by ice formation during sample preparation or handling. Deposited ice attenuates the main beam. As a weak side effect it leads to cones of scattered, secondary X-rays that illuminate parts of the crystal outside the main beam. At room temperature ionisation products such as small free radicals can spread the absorbed energy throughout the crystal.

Table 3.1: A selection of crystal parameters affecting dose absorption and distribution in an MX experiment.

Both limits describe an upper bound for cryo-crystallographic experiments. There is in fact no guarantee that a particular protein crystal will not decay faster and stop diffracting before either limit is reached (Murray *et al.*, 2005). However even in those circumstances, the behaviour of one protein crystal can be predictive for the behaviour of another crystal of the same protein, even when using a different beam intensity and with varying experimental conditions (Bourenkov & Popov, 2010). There is therefore good reason for an experimentalist to know the doses involved in their beamline experiments.

Predominant **beam** parameters affecting dose absorption and distribution	
size	Together with the crystal size and the data collection strategy the (x,y) extent of the beam determines the crystal and solvent volume exposed to X-ray irradiation throughout the experiment. Photoelectrons can escape the beam path, which can, especially with very small beams, lead to a reduced effective dose.
flux	The dose absorbed by the crystal directly depends on the beam flux (photons per second). While an experiment with doubled flux and halved time results in the same dose distribution, there is some evidence that specific radiation damage is dose rate dependent (Leiros *et al.*, 2006).
energy / wavelength	Photon-matter interaction probabilities are dependent on the beam energy or wavelength. No X-ray beam is perfectly monochromatic, so even under non-Laue conditions the beam energy should ideally be modelled not as a fixed quantity, but as a distribution (Rack *et al.*, 2010).
profile	The X-ray beam is usually assumed to have a flat profile with a uniform flux throughout the beam extent, or alternatively modelled as a two-dimensional Gaussian profile with the maximum in the beam centre. Both are inadequate approximations for real beam profiles, as shown by for example by Zeldin *et al.* (2013a, Fig. S1) for beamline ID14-4 at the European Synchrotron Radiation Facility. Ideally the beam profile should be imaged and tracked on the day of the experiment using the actual beam conditioning and data collection settings.
divergence	No beam is perfectly parallel. The beam divergence or convergence affects which parts of the crystal are actually exposed at any point in time. By treating the beam as parallel a small error may be introduced for in-house sources. This effect is negligible for the near-parallel beams of synchrotron sources.
time	None of these parameters is fixed over time. Modern synchrotron sources will follow a continuous injection protocol leading to very small variance in flux over time. Optics elements may be subjected to heating by the beam causing thermal drift of the shown parameters. On the extreme end of the X-ray diffraction experimental spectrum, the wavelength distribution of X-ray free electron lasers varies significantly from shot to shot due to the nature of the self-amplification process (Saldin *et al.*, 2010).

Table 3.2: A selection of beam parameters affecting dose absorption and distribution in an MX experiment.

Predominant experimental parameters affecting dose absorption and distribution	
Data collection strategy	
rotation	For beams smaller than the crystal or beams that do not have uniform flux the rotation of the crystal (change of ϕ, χ, ω angles over time) determines the resulting dose distribution within the crystal. Offsetting the rotation axis from the beam axis results in a toroid shaped exposure pattern (Zeldin *et al.*, 2013a).
translation	Any translation during exposure, e.g. running a helical scan or a translational scan (Zeldin *et al.*, 2013c), will strongly affect the dose distribution and any dose summary metrics.
exposure time	The absorbed dose is directly dependent on the fluence (flux integrated over exposure time), see *flux* in Table 3.2.
Others	
humidity / atmosphere	X-ray diffraction experiments are not typically performed *in vacuo*. The X-ray beam is therefore attenuated by the air between the X-ray source and the sample. Both the length of the beam path and the composition of the air (Owen *et al.*, 2009) affects the effective attenuation.

Table 3.3: Further parameters affecting dose absorption and distribution in an MX experiment.

However, it is impossible to measure dose directly. The energy absorbed by a crystal depends on a large number of factors, not all of which can be controlled by the experimenter. Some experimental parameters, such as the crystal temperature, have no effect on the total absorbed dose, but may affect the distribution of dose within the crystal. A list of these factors currently thought to be most important is given Tables 3.1 (related to properties of the crystal), 3.2 (related to those of the beam), and 3.3 (further parameters). If most or all experimental conditions are known, it is possible to simulate the experiment and therefore estimate the dose absorbed by the crystal. One piece of software capable of such simulation is RADDOSE.

RADDOSE is being used extensively for research into the nature of radiation damage with a view to improving data collection results (Murray *et al.*, 2005; Owen *et al.*, 2006; Bowler *et al.*, 2010; Holton & Frankel, 2010; Sanishvili *et al.*, 2011; Owen *et al.*, 2012). It is also being used as

a dose estimation tool for protein crystallographers, giving credence to the accuracy of published protein structures (Beitlich *et al.*, 2007; Johansson *et al.*, 2010; Gumiero *et al.*, 2011; Zhai *et al.*, 2011; White *et al.*, 2012). Although RADDOSE was not developed with X-ray free electron lasers (X-FELs) in mind, it has even been used to produce dose estimates in this context (Chapman *et al.*, 2011; Kern *et al.*, 2012). The software package BEST (Bourenkov & Popov, 2010) makes use of RADDOSE to provide radiation-damage aware data collection strategy optimisation. RADDOSE has been integrated at beamlines at the European Synchrotron Radiation Facility, France, at the Diamond Light Source, UK, with a graphical user interface written by Robin Owen, at the Swiss Light Source, Switzerland, and with a graphical user interface written by Masahide Kawamoto, at the SPring8, Japan.

3.1.1 Dose estimation with RADDOSE

There are three major releases of RADDOSE, versions 1 (Murray *et al.*, 2004), 2 (Paithankar *et al.*, 2009), and 3 (Paithankar & Garman, 2010).

At the time of its publication, RADDOSE (v1) was the first software capable of estimating the dose absorbed in an X-ray crystallography experiment and thus advising the experimenter on the available exposure time until the Henderson limit is reached. The complete unit cell composition is modelled using information such as the number of amino acid or nucleotide residues per monomer, number of monomers per unit cell, the unit cell size and angles, and heavy atoms in the protein and the solvent. RADDOSE (v1) only considers the energy absorbed via the photoelectric effect, the dominant effect in the energy ranges usually used in MX (7 keV to 14 keV). While the energy absorption via the Compton effect is neglected, the beam attenuation by the Compton effect is considered, leading to an underestimation of the resulting dose. RADDOSE (v1) internally assumes

that the attenuation and absorption cross-sections are identical ($\mu_{abs}=\mu_{att}$). This treatment results in an overestimate of the absorbed dose: coherent scattering does attenuate the beam, but does not contribute to the absorption.

Alongside information such as the crystal solvent content, the total absorbed dose, absorbed dose per image and temperature rise in the crystal[1], RADDOSE provides the experimenter with information on how much experimental time is available until the crystal reaches the Henderson limit or the Garman limit[2]. This information can then be used in the planning of a data collection strategy.

RADDOSE (v1) also assumes that any absorbed energy stays within the crystal. This further overestimates the dose for crystals involving elements with a high atomic number, as these elements, upon excitation by the incident X-ray, can stabilise via emission of a fluorescent photon instead of an Auger electron (Figure 1.2, page 6). Depending on its origin within the crystal and the crystal size, the fluorescent photon may escape the crystal, and reduce the total absorbed energy and thus the absorbed dose. In combination with higher incident X-ray energies, the error in dose estimate can be substantial: Paithankar *et al.* (2009) report a 37 % overestimation of dose for a selenomethionine protein crystal irradiated at the selenium K-edge. Fluorescent escape is added in RADDOSE (v2), with the minor limitation that the escape probability is always based on a fluorescent photon originating from the centre of the crystal (Paithankar *et al.*, 2009). In RADDOSE (v3) the effects of Compton scattering are included with appropriate angular distribution and self-absorption treatment to the calculations (Paithankar & Garman, 2010). Compton scattering only becomes a relevant process when using high energy incident beams (above 20 keV,

[1] For this the heating model described by Kuzay *et al.* (2001) is used.
[2] Within RADDOSE this limit is called the 'experimental dose limit'.

Section 1.1).

Photoelectrons can travel some distance and even escape from the crystal (Nave & Hill, 2005). For an incident beam energy of 18.5 keV they have an approximate range of 4 µm, which leads to an overestimation of absorbed dose in particular for small crystals (Sanishvili *et al.*, 2011). Currently no version of RADDOSE takes these dose-spreading and dose-reducing effects into account.

In all three RADDOSE versions the crystal is treated as a (strictly cuboid) entity with uniform internal dose distribution. The omission of crystal rotation allows the integration of the dose absorption over the exposed time. Consequently, RADDOSE only reports one dose summary metric, which is the maximum dose. It follows that, when the beam is smaller than the crystal, RADDOSE reports an overestimated dose, as new, non-irradiated crystal volume will have entered the beam during rotation. The proper treatment of the inhomogeneous internal dose distribution was the main motivation for the development of a new version of RADDOSE, RADDOSE-3D. An overview of the different effects and parameters supported in different versions of RADDOSE and RADDOSE-3D is shown in Table 3.4.

3.1.2 Limitations of RADDOSE: the need for RADDOSE-3D

When RADDOSE was initially presented, the assumption of a beam larger than the crystal was generally valid. As X-ray beamline technology advanced, much smaller beams became available to crystallographers. Modern microfocus beamlines, such as ID23-2 at the ESRF, can produce an X-ray beam as small as 6.8 µm × 3.4 µm (Flot *et al.*, 2010). With beams much smaller than the crystals, the assumption of homogeneous dose distribution within a crystal is no longer valid, and the effects of rotation should no longer be neglected (Zeldin *et al.*, 2013c). The fundamental dose limitations still apply in microbeam scenarios, but now they also extend to subsections of the

Inclusion of physical phenomena in different versions of RADDOSE				
	RADDOSE version			
	v1	v2	v3	3D
crystal shape - cuboid	●	●	●	●
crystal shape - spherical	○	○	○	●
photoelectric effect	●	●	●	●
photoelectron path	○	○	○	○
photoelectron escape	○	○	○	○
fluorescent escape	○	●	●	○
Compton scattering (attenuation)	●	●	●	●
Compton scattering (reabsorption)	○	○	●	○
dose distribution within crystal	○	○	○	●
correctly treats $\mu_{att} \neq \mu_{abs}$	○	○	●	●
reports maximum dose	●	●	●	●
advanced dose metrics (Zeldin *et al.*, 2013a,c)	○	○	○	●

Table 3.4: Support of physical phenomena in different RADDOSE versions

crystal. In a simple experiment with a beam smaller than the crystal, and the goniometer rotation axis perpendicular and intersecting with the beam axis, it is intuitively clear that the centre of the crystal suffers a much higher dose than the outer regions, which are only within the beam for a smaller angular range. The experimenter therefore can only resort to a worst-case scenario and set up the data collection strategy in such a way that the total exposure is limited by the maximum dose of the central crystal region, i.e. no portion of the crystal exceeds the specified maximum dose. Alternatively, the researcher can accept that this maximum dose will be exceeded in the centre of the crystal.

The maximum dose can be reduced by efficient use of the available crystal volume (Zeldin *et al.*, 2013c), which is a well known scientific problem (Galton, 1906), and can be achieved in crystallography for example by using a helical scan (Flot *et al.*, 2010). For these data collection strategies, the crystal is continuously translated perpendicularly to the beam axis while it is also translated, as shown in Figure 3.1.

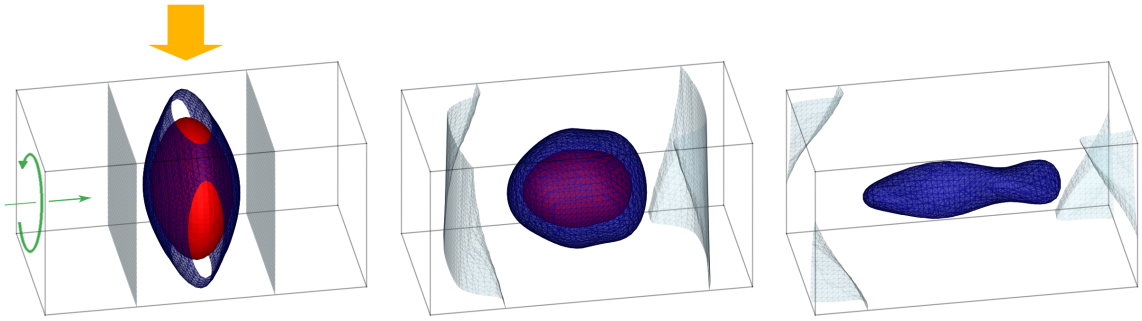


Figure 3.1: Dose spreading through crystal translation
X-ray diffraction experiment involving a $50\times100\times50\,\mu m^3$ crystal exposed for a full 360° rotation in a $20\,\mu m\times20\,\mu m$ FWHM Gaussian beam collimated to $40\times40\,\mu m^2$. The crystal is rotated (green) along its y-axis. The X-ray beam (orange) perpendicularly intersects the rotation axis (green). Resulting dose isosurfaces are shown at 30 MGy (red), 20 MGy (dark blue), and 0.1 MGy (light blue). Figures obtained using the RADDOSE-3D web service (Section 3.5.3). **Left:** Crystal is not translated during exposure. Maximum dose at the crystal centre reaches 88 MGy. **Middle:** Crystal is translated by 0.111 µm/° along the rotation axis during exposure. Maximum dose at the crystal centre reaches 46 MGy. **Right:** Crystal is translated by 0.222 µm/° along the rotation axis during exposure. Maximum dose at the crystal centre reaches 23 MGy.

To establish the actual extent of dose inhomogeneity in realistic scenarios, and ultimately to aid the experimenter in designing a data collection strategy under these circumstances, it became necessary to model the experiment in more detail than a simple cuboid, homogeneous crystal that does not rotate in the beam. A newer version of RADDOSE, RADDOSE-3D was therefore devised, that in addition to the ability of modelling spatial dose distribution, also allowed non-cuboid crystal shapes. It can also support more complex beam profiles than those adequately described by a rectangular top-hat or an ideal Gaussian beam.

3.2 RADDOSE-3D architecture

RADDOSE-3D provides a generic and extensible simulation framework for three-dimensional X-ray crystal diffraction experiments. Its most obvious function is to estimate crystal doses for an ex-

periment after it has taken place. RADDOSE-3D can also be integrated into beamlines and be used for live tracking of doses throughout an experiment, and visualisation of dose states and dose progression. RADDOSE-3D has been successfully used to produce dose estimates for crystallography experiments (Kekilli *et al.*, 2014; Yogavel *et al.*, 2014), to evaluate data collection strategies before the experiment, and to determine the merit of new dose summary metrics (Zeldin *et al.*, 2013a).

The programming language of choice for the development of RADDOSE-3D was Java, an object oriented programming language. Java has a large user base, is widely used in academic settings, and is relatively easy to learn for new programmers. Java code is platform-independent, therefore RADDOSE-3D can be run on any machine with an installed Java Runtime Environment.

In this section how RADDOSE-3D works will be described, how it is designed and what the rationale behind its architecture is.

3.2.1 Object oriented programming (OOP)

Developing RADDOSE-3D in an object oriented language facilitates the creation of reusable and modular code. To understand the RADDOSE-3D architecture some basics of OOP are necessary. There are two fundamental concepts in OOP: classes and objects.

A class is a fixed collection of code, such as functions and methods. Classes can be derived from other classes by overwriting their functions or adding on new functions. Every class has a unique, fixed name and a parent class (called a superclass)[3].

An object is an instance of a class. In addition to its code, which is identical for all the objects

[3] This is necessarily a simplification. There are many more constructs of near-classes such as abstract classes and interfaces. Classes may have multiple parents and no name, etc. Naturally, the interested reader can find out much more about these ideas in a number of relevant computer science books.

derived from the same class, it carries an internal state. This internal state is usually protected from direct outside manipulation, and is only changed via the functions and methods of the object. Thus objects are a key element for modular project design: parts of a large program can run isolated from other modules and only interact with one another via predefined interfaces with specified and documented semantics. As objects are unaware of the internal workings of other objects, objects can be changed and new objects can be introduced without the need to change code in other existing objects. This allows RADDOSE-3D to become an easily extendible platform.

RADDOSE-3D consists of a set of classes, each concerned with a small part of the overall simulation. These classes range from the EXPERIMENT class, which is a central coordinating instance, to the details of how RADDOSE-3D receives and processes experimental descriptions via INPUT classes. Elements of the experiment are internally represented by CRYSTAL, BEAM, and WEDGE classes, while the experimental analysis of simulations happens within OUTPUT and EXPOSEOBSERVER classes. To illustrate how a diffraction experiment is modelled within RADDOSE-3D these most important classes (or superclasses) will be presented in this section.

3.2.2 The RD3D class: Instantiation

If RADDOSE-3D is started from the command line, the main-function in the class RD3D is called. When no command line options are given, the RADDOSE-3D version is printed, along with a short summary of the available command line options (Figure 3.2). Otherwise RD3D will set up helper classes according to the scheme shown in Figure 3.3. By default this will include one INPUT class reading from a configuration file and a predefined set of four OUTPUT classes. These four classes and the files they create are shown in Table 3.5.

Furthermore, RD3D will create an instance of the EXPERIMENT class (Section 3.2.4) which sits

```
raddose   --   command line options  (1.1.1000)
=========================================================
 -?   or --help         show command line help (this)
 -V   or --version      show version information

 -i   or --in filename  read instructions from file
 -i - or --in -         read instructions from STDIN

 -p   or --prefix name  prefix for output files
 -r   or --raddose path path to RaddoseV3 executable
 -t   or --test         test run with no simulation

 -o   or --out <output> for user-defined output
     Default output corresponds to
       -o SummaryCSV:SummaryCSV.csv
       -o Summary:Summary.txt,-
       -o DoseStateCSV:DoseState.csv

=========================================================
Please cite:
 Zeldin, Gerstel, Garman. (2013). J. Appl. Cryst. 46, 1225-1230.
 http://dx.doi.org/10.1107/S0021889813011461
```

Figure 3.2: RADDOSE-3D command line help

RADDOSE-3D version 1.1 relies on a previous version of RADDOSE (either v2 or v3) to calculate the attenuation and absorption coefficients based on the crystal composition.

RADDOSE-3D output files, their content and the associated class (Sect. 3.2.8, page 102)

file name	associated OUTPUT class	content
OUTPUT-DOSESTATE.CSV	OUTPUTFINALDOSESTATECSV	three-dimensional scalar field of crystal dose values at the end of the experiment
OUTPUT-DOSESTATE.R	OUTPUTFINALDOSESTATER	source code that can be loaded directly into R (R Development Core Team, 2011) to visualise the crystal dose state at the end of the experiment.
OUTPUT-SUMMARY.CSV	OUTPUTSUMMARYCSV	comma separated value table containing per-wedge experimental information
OUTPUT-SUMMARY.TXT	OUTPUTSUMMARYTEXT	summary of the experiment, also shown on the command line, Figure 3.9

Table 3.5: Default RADDOSE-3D output files

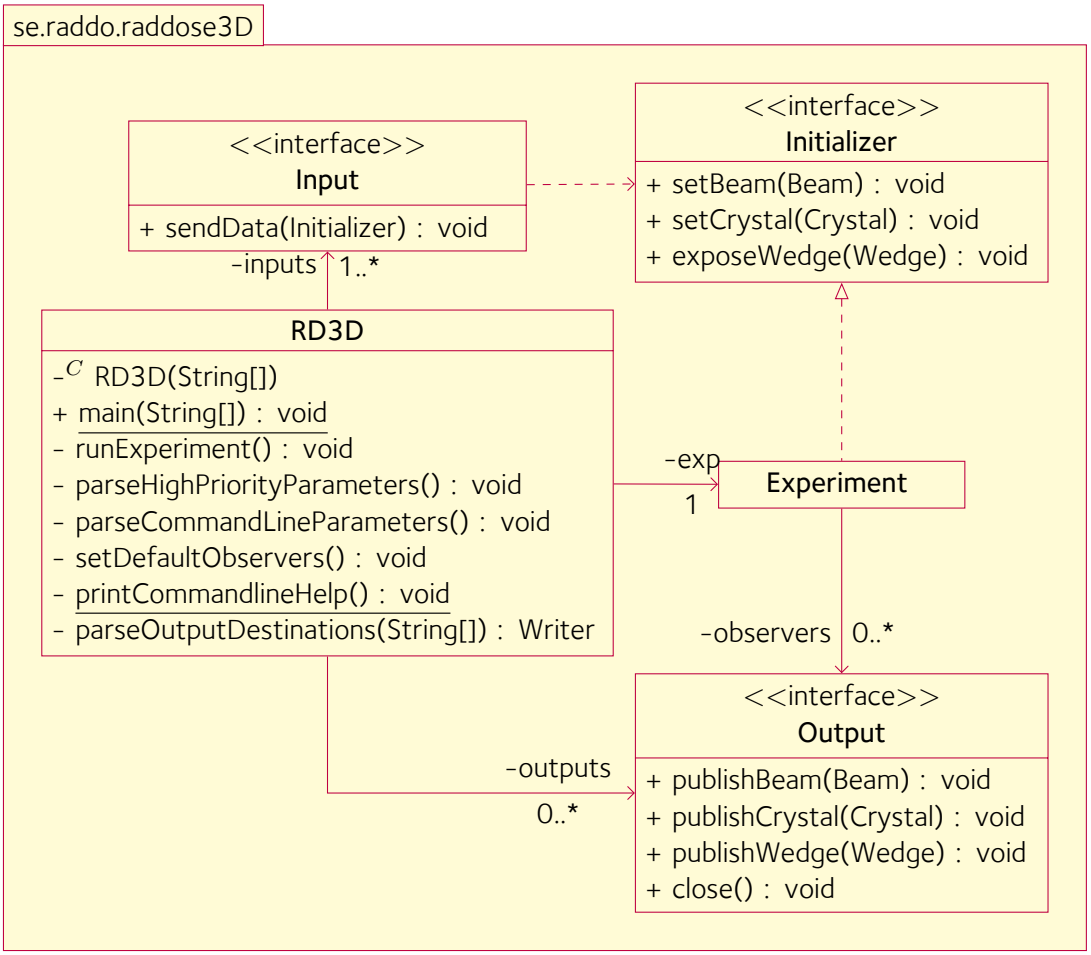


Figure 3.3: Unified Modeling Language (UML) class diagram of the RADDOSE-3D object structure for the experimental setup.

In a UML diagram different classes are represented by individual rectangles. Classes can be annotated with attributes or functions. Private functions (those that are only available within the class, but can not be called from other classes) are prefixed with a '-', public functions with a '+'. For clarity, only a few relevant functions are shown. Relations between classes are represented by different arrows.

When the user starts RADDOSE-3D from the command line an instance of the class RD3D is created (left). This class holds references to one or more classes derived from the INPUT superclass (top left), exactly one EXPERIMENT class (right) and zero or more classes derived from the OUTPUT superclass (bottom right).

The INPUT type classes are responsible for processing the user input. There can be different implementations of INPUT-type classes, for example one for reading from a file, and another one for reading from a database. Once the input is read and processed it will be passed on to an INITIALIZER type class (top right). RADDOSE-3D provides one default implementation of the INITIALIZER class, called EXPERIMENT (right). This class is responsible for the coordination of the simulation, but does not produce any analysis or output by itself. For this purpose a number of OUTPUT classes (bottom right) can be attached to EXPERIMENT, which can observe and analyse the simulation as it happens.

Which INPUT and OUTPUT classes are actually used is decided by RD3D and ultimately by the user. This design makes RADDOSE-3D very flexible and allows it to run in any possible environment with few changes.

between the INPUT and OUTPUT classes and has a coordinating role.

3.2.3 The INPUT class: Instruction

INPUT objects provide a function called sendData to transfer an ordered stream of BEAM, CRYSTAL and WEDGE objects to an INITIALIZER class (Figure 3.4). Currently the class INPUTPARSER is the only existing direct subclass of INPUT. INPUTPARSER runs a parser generated using ANTLR[4] (Parr, 2013), to interpret the descriptions of an X-ray diffraction experiment given by the user in a configuration file. An example configuration file describing a simple X-ray diffraction experiment is shown in Figure 3.5. The RADDOSE-3D configuration file syntax is described in its entirety in the RADDOSE-3D user guide, which can be found in Appendix C (page 214) and online[5].

Future INPUT implementations could obtain information from sources other than text files, such as from graphical user interfaces or even directly from measurements taken at a beamline.

3.2.4 The EXPERIMENT class: Coordination

The EXPERIMENT class contains the controlling logic for the simulation. It receives calls from the INPUT classes providing information about the experiment, such as the details about the crystal, the X-ray beam and data collection strategy.

However, there is no analysis carried out within the EXPERIMENT class itself. Rather other objects, derived from the superclass OUTPUT, can be attached to the EXPERIMENT class (Figure 3.6). EXPERIMENT will then pass on the information it received about the experiment to this list of effectively 'subscribing' OUTPUT objects[6]. Because the EXPERIMENT class only routes information about

[4] http://www.antlr.org

[5] http://raddo.se/rd3d/RD3D-User-Guide.pdf

[6] The arrangement of objects to achieve this relationship is called an 'Observer' or 'Publish-Subscribe' design pattern (Gamma *et al.*, 1994).

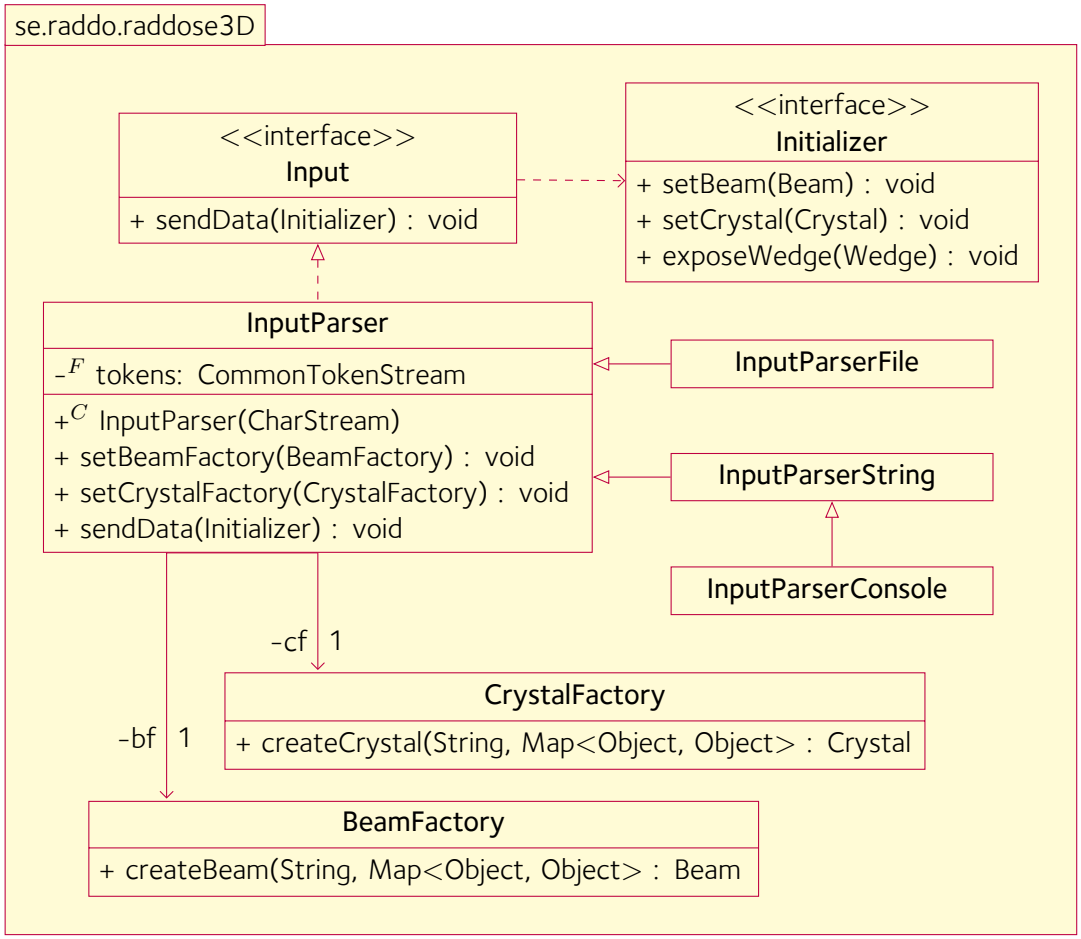


Figure 3.4: UML class diagram of the RADDOSE-3D user input processing architecture.

RADDOSE-3D currently contains three directly usable INPUT-class implementations to process experimental descriptions provided by the user: INPUTPARSERFILE reads and processes a text file. INPUTPARSERSTRING processes commands already contained within a Java STRING object. INPUTPARSERCONSOLE, which is derived from INPUTPARSERSTRING, reads in the experimental description from user input given at the command line interface (STDIN).

All three of these classes are subclasses of INPUTPARSER. INPUTPARSER contains a parser common to all its derivative classes. This parser is generated by ANTLR, and capable of understanding the RADDOSE-3D command syntax described in the RADDOSE-3D user guide.

While the parser is to some extent aware of some aspects of CRYSTAL and BEAM classes, e.g. that a crystal may have a three-dimensional size, it does not know exactly what CRYSTAL and BEAM implementations are available, and how the associated classes are instantiated. This knowledge is contained within the CRYSTALFACTORY and BEAMFACTORY classes. Both are used by the INPUTPARSER class to generate CRYSTAL and BEAM objects, which are then passed on, together with WEDGE classes containing information about crystal exposure events, to an INITIALIZER class for the exposure simulation.

```
Crystal
Type Cuboid
Dimensions 100 100 100

Beam
Type Gaussian
Flux 1e12
FWHM 40 20
Energy 12.1
Collimation Rectangular 80 40

Wedge 0 90
ExposureTime 30
AngularResolution 2
```

Figure 3.5: RADDOSE-3D input file example.
A RADDOSE-3D file describing an experiment involving a $100 \times 100 \times 100$ µm^3 cuboid crystal in a Gaussian beam with 10^{12} photons/s at 12.1 keV with full-width half-maxima of 40 µm (vertical) and 20 µm (horizontal) in a collimated area of 80×40 µm^2 (vertical/horizontal). The crystal is exposed for a total of 30 seconds through an angular range of ϕ from 0 to $90°$. The rotation is simulated in $2°$ steps.

the simulation, it can remain completely unaware of the internal details of the simulation, such as different crystal shapes and beam types.

3.2.5 The CRYSTAL class: Workpiece

CRYSTAL objects contain a representation of the physical crystal volume as a set of evenly spaced voxel coordinates (voxel stands for **vo**lumetric pi**xel**, a unit of volume). For each voxel the associated dose and fluence values are stored.

The size of the voxels and hence the resolution of the simulation can be user-defined[7]. A smaller voxel size results in an increased simulation resolution, but also increases the simulation time and the amount of computer memory required. As a guideline, the crystal voxel element resolution should be chosen depending on the beam profile. For Gaussian beams the size of the voxel elements should be smaller than $1/10$th of the full-width half-maximum (FWHM) of the

[7] By default a voxel corresponds to a volume element of $2 \times 2 \times 2$ µm^3.

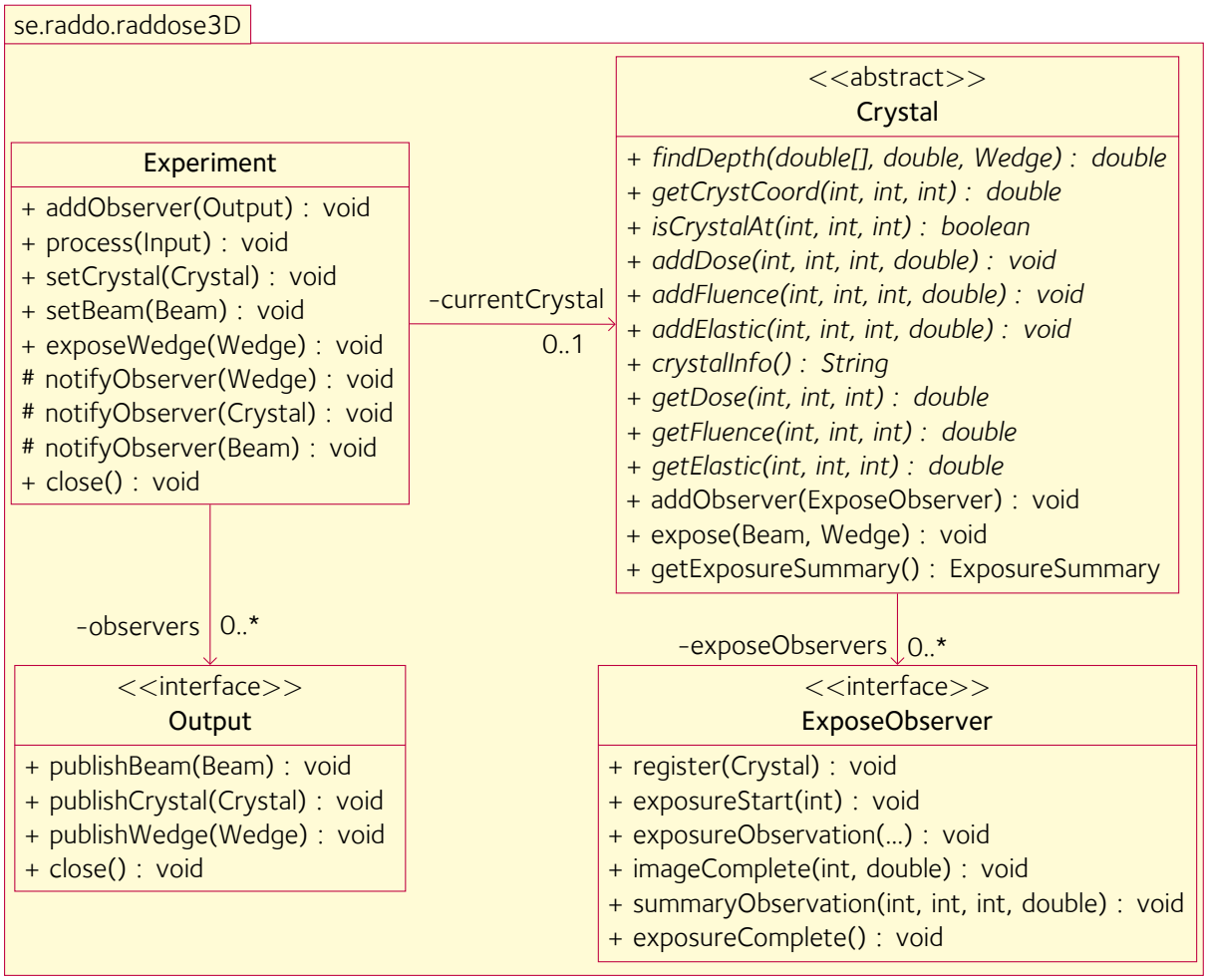


Figure 3.6: UML class diagram of the RADDOSE-3D simulation analysis components.

The EXPERIMENT class (top left) receives a stream of CRYSTAL, BEAM and WEDGE objects from one or more INPUT objects (not shown; Figures 3.3 and 3.4). These are passed on to an arbitrary number of OUTPUT objects (bottom left, which have previously subscribed to EXPERIMENT via its addObserver function.

An OUTPUT object will therefore be informed about any changes in the experimental setup and about any exposures taking place, but it will not receive detailed information throughout the exposure. This level of detail is available to EXPOSEOBSERVER classes, which can be attached to a CRYSTAL object via a its corresponding addObserver function. These classes are then notified for each exposure image (function imageComplete) and even for each individual voxel exposure event (function exposureObservation). Modern Java environments allow optimisation during runtime and will not execute unnecessary function calls, meaning that this architecture is efficient even for EXPOSEOBSERVER implementations that do not require continuous per-voxel information.

beam in its smaller dimension (Zeldin *et al.*, 2013b, §3). A higher resolution may be required to accurately simulate the details of more uneven beam profiles, such as experimental beam profiles.

CRYSTAL objects also contain the necessary functions to translate and rotate the crystal volume, and to calculate the depth of any voxel relative to the beam. The depth calculation is essential for determining the beam intensity incident on the voxel, as the attenuation by the crystal volume between the voxel and the beam source needs to be taken into account. The complexity of the rotation and depth-finding functions depend on the restrictions placed on the crystal geometry. Calculating the depth of a voxel in a cuboid is more easily achieved than in an arbitrarily shaped convex object, which in turn can be calculated faster than the depth of a voxel in an arbitrarily shaped concave object. Subclasses of the CRYSTAL class can make use of these geometry restrictions. Currently there are two subclasses available in RADDOSE-3D: CRYSTALCUBOID, which defines a cuboid crystal shape, and CRYSTALSPHERICAL, which defines a spherical crystal (Figure 3.7). New CRYSTAL classes can be added by implementing the required functions for storing and retrieving dose and fluence values and providing a function to calculate the depth of voxels in a given crystal rotation and translation. A class for arbitrary polyhedron crystals, CRYSTALPOLYHEDRON, is currently being developed by Helen Ginn.

The CRYSTAL classes also hold an internal reference to a COEFCALC object, used to obtain the relevant beam attenuation and absorption coefficients. RADDOSE-3D versions up to 1.1 contain two different COEFCALC implementations. COEFCALCRADDOSE calculates the coefficients from the crystal composition and actual beam energy (Murray *et al.*, 2004; Paithankar *et al.*, 2009) by feeding this information to a RADDOSE version 2 executable[8].

[8]The upcoming RADDOSE-3D version 1.2 will contain a Java implementation by Helen Ginn of the relevant RADDOSE version 2 Fortran code, eliminating the requirement to install a separate (and platform-dependent) copy of RADDOSE.

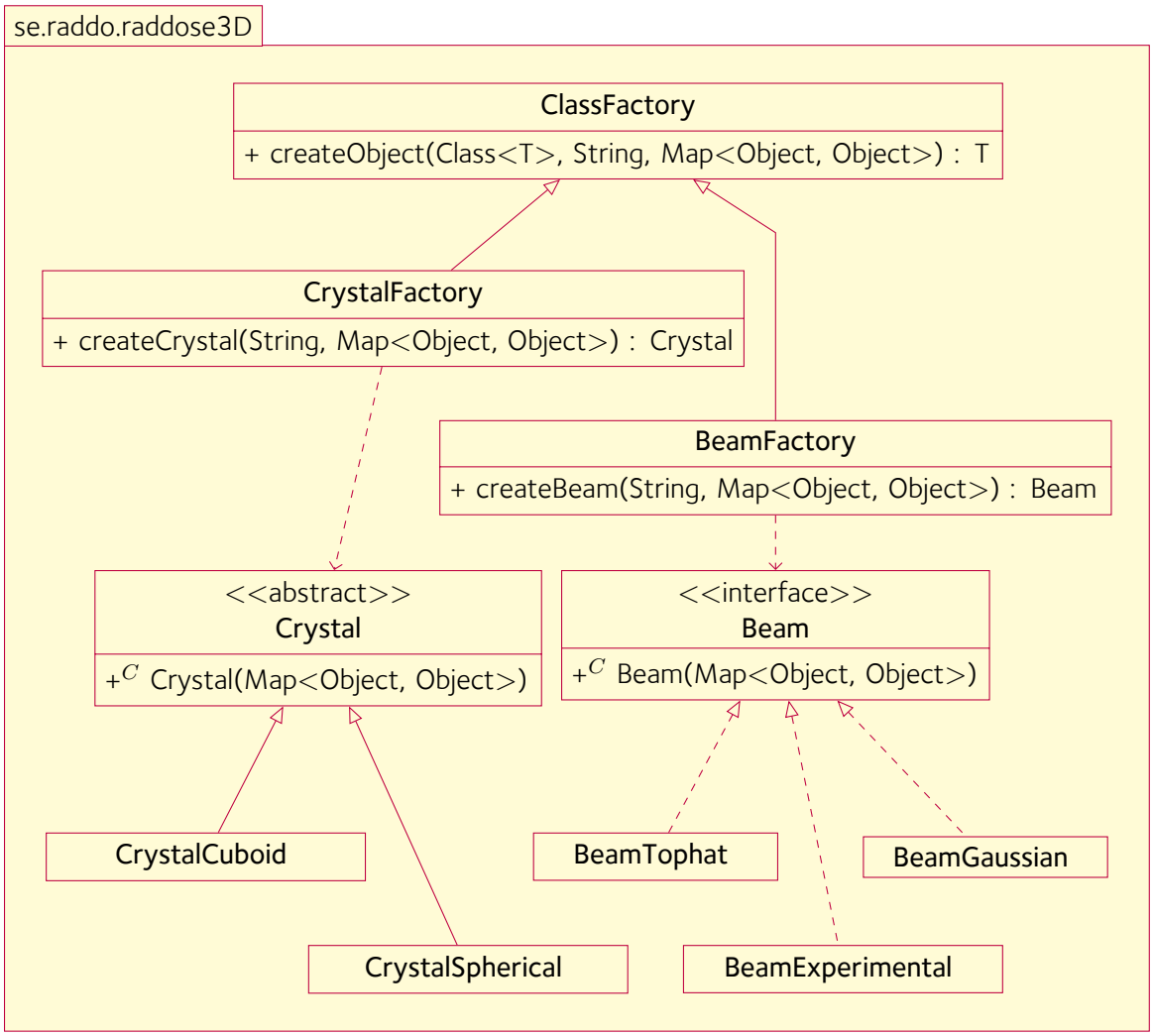


Figure 3.7: UML class diagram of the available RADDOSE-3D beam and crystal classes.

RADDOSE-3D supports different crystal shapes and beam definitions. The class CRYSTAL defines a number of common functions that all subclasses need to implement. RADDOSE-3D currently contains two crystal implementations, CRYSTALSPHERICAL and CRYSTALCUBOID. Similarly, there are three beam implementations, BEAMTOPHAT, BEAMGAUSSIAN, and BEAMEXPERIMENTAL, each providing the functions defined in the BEAM interface.

All CRYSTAL and BEAM classes can work together, and the decision which classes are used is made depending on the user input.

In contrast, COEFCALCAVERAGE returns fixed values: the absorption coefficient is defined as 0.237 mm^{-1} and the attenuation coefficient as 0.281 mm^{-1}. These numbers were calculated (Holton & Frankel, 2010) based on a (hypothetical) average protein with composition $H_{49.8}C_{31.8}N_{8.56}O_{9.54}S_{0.249}$ and assuming a beam energy of 12.4 keV. When using these averages, the crystal density is taken as 1.2 g/ml with an assumed 50% solvent content with solvent mass 1 g/ml, and a protein mass of 1.4 g/ml (Fischer *et al.*, 2004). The absolute dose values calculated by using COEFCALCAVERAGE should never be quoted in publications relating to actual experiments, as there is considerable variance in the coefficients, especially when the crystal contains traces of heavy elements. The average coefficients can however be useful to indicate the approximate spatial dose distribution within the crystal or to compare data collection strategies.

3.2.6 The BEAM class: Tool

The BEAM interface facilitates access to information describing the beam profile, energy, and flux.

There are currently three beam shape implementations (Figure 3.7): BEAMTOPHAT describes a top hat beam, with a uniform intensity at each point within the beam, BEAMGAUSSIAN defines a beam with a continuous 2D Gaussian profile (using the Apache Commons Mathematics Library[9]) with user-defined x and y FWHM, and BEAMEXPERIMENTAL models a beam with an arbitrary profile based on experimental measurements.

The first two beam shapes provide useful approximations for most experimental beams. BEAMTOPHAT can be used when there is no information available about the beam shape, apart from the collimation settings. BEAMGAUSSIAN is a useful approximation for X-ray beams when limited shape information is available, for example from a scintillator or from X-ray sensitive paper.

[9] http://commons.apache.org/math/

BEAMEXPERIMENTAL can be used when the beam has been fully characterised by measuring the beam intensity in a two-dimensional grid. This has so far been implemented for beams at the ESRF (Grenoble, France) and PETRA II (Hamburg, Germany), and work is currently being undertaken to implement this for beamline I02 at the Diamond Light Source (Didcot, UK).

3.2.7 The WEDGE class: Strategy

WEDGE objects contain all the information pertaining to the data collection strategy: rotational range, any offset in either the starting orientation and position of the crystal, or between the rotation axis and the beam axis, and any translation that is to take place during the wedge exposure (e.g. a helical strategy, Flot *et al.*, 2010). They are primarily used as data storage objects and contain little program logic.

3.2.8 The OUTPUT/EXPOSEOBSERVER classes: Analysis

The output of RADDOSE-3D is created from objects that can subscribe to experiments using the Observer design pattern (Gamma *et al.*, 1994). The logic concerned with experimental dose calculations is contained in these classes.

At the moment, RADDOSE-3D offers two different levels of detail for observing objects (Figure 3.6, page 98): where monitoring of the general flow of experimental objects on a high level is required, classes can implement the OUTPUT interface (Figure 3.8) and subscribe to the central EXPERIMENT object. These classes will then be notified of the arrival of new CRYSTAL and BEAM objects as well as of any WEDGE exposure events that have taken place. The observing objects can inspect the CRYSTAL and BEAM objects before and after exposure.

When a more detailed look into the time dependent voxel-exposure events taking place is nec-

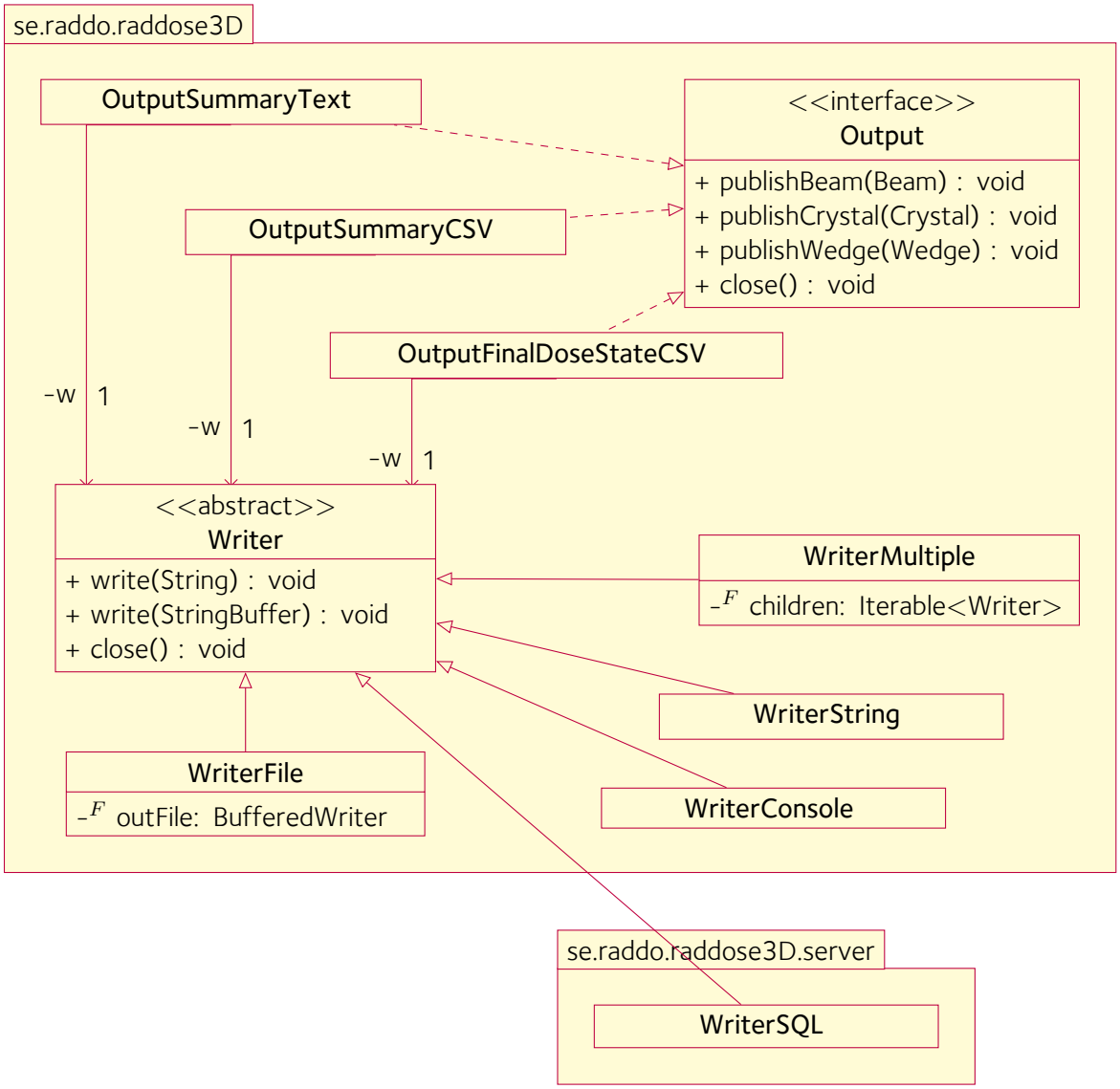


Figure 3.8: UML class diagram of the standard RADDOSE-3D analysis and output components.

Each OUTPUT class implementation (top right) can be attached to an EXPERIMENT class, and receive information about the currently running crystal exposures (Fig. 3.6). There are three RADDOSE-3D OUTPUT implementations shown here: OUTPUTSUMMARYTEXT provides a human-readable experimental summary, an example of which can be seen in Figure 3.9, page 105. The class OUTPUTSUMMARYCSV provides the same information in a machine-readable comma separated value format. Finally, OUTPUTFINALDOSESTATECSV provides the machine-readable three-dimensional scalar dose field of the entire crystal after all exposures have taken place.

Each OUTPUT implementation uses a WRITER class to perform any actual file operations. A WRITER class only receives strings, and is responsible for the storage or further processing of these data. The most obvious WRITER implementation is the WRITERFILE class (bottom left), which writes out received data to a specified file. The WRITERCONSOLE class (bottom right) sends data to the command line interface, where they can be seen during the RADDOSE-3D simulation. Output can be directed to more than one destination using the WRITERMULTIPLE class, which forwards any received data to a list of other WRITER objects. For example the output of OUTPUTSUMMARYTEXT is, by default, shown on the console and written to a file using a mediating WRITERMULTIPLE object.

The OUTPUT classes do not need to be aware of where and how their results are displayed or stored. This allows a clear separation between the experiment analysis logic in the OUTPUT class, and the file, console writing, or – with WRITERSQL used in the RADDOSE-3D server – even database storage logic in WRITER classes.

essary, a class can implement the EXPOSEOBSERVER interface and subscribe to the CRYSTAL object. These classes will then be notified for each individual voxel exposure, between each diffraction image taken, and at the beginning and end of each exposure.

This program structure allows easy creation of new OUTPUT modules that can produce any required statistic in any format.

OUTPUT classes can also provide functionality that is not directly relevant to the simulation results. The class OUTPUTPROGRESSESTIMATE reports the number of crystal voxels and wedge slices, which can be used to produce a simulation runtime estimate. A progress bar can be shown during each wedge exposure using the class OUTPUTPROGRESSINDICATOR, which is particularly helpful for long, high resolution simulations. Finally, for the RADDOSE-3D webservice the class EXPERIMENTYIELDER (despite its name an implementation of OUTPUT) ensures that parallel simulations run alongside fairly, so that no one simulation starves the others of resources.

The main results of an RADDOSE-3D simulation, which is shown on the command line and stored in the file OUTPUT-SUMMARY.TXT, is produced by the OUTPUTSUMMARYTEXT class. This output, an example is shown in Figure 3.9, includes a description of the simulation and relevant dose summary metrics (Zeldin *et al.*, 2013c) including the Diffraction Weighted Dose (Zeldin *et al.*, 2013a).

New output modules can be created and used without the need to touch existing code. The class CLASSFACTORY contains a generic class loader, which can create instances of classes that were unknown at compile time. This allows the independent development of OUTPUT classes as well as INPUT, BEAM and CRYSTAL classes, which do not have to be part of the RADDOSE-3D package, but can still be used in simulations.

Having a large number of observing classes can increase the runtime of a simulation. Therefore

```
Cuboid Crystal of size [100, 100, 100] um [x, y, z] at a resolution of
 2.00 microns per voxel edge.
Simple DDM.
Gaussian beam, 80.0x40.0 um with 40.00 by 20.00 FWHM (x by y) and
 1.0e+12 photons per second at 12.10 keV.

Wedge 1:
Collecting data for a total of 30.0s from phi = 0.0 to 90.0 deg.
Dummy crystal coefficients used.
Absorption coefficient: 2.37e-04 /um.
Attenuation coefficient: 2.81e-04 /um.
Elastic coefficient: 1.80e-05 /um.
Density: 1.20 g/ml.

Average Diffraction Weighted Dose         : 2.94 MGy
Elastic Yield                             : 5.82e+10 photons
Diffraction Efficiency (Elastic Yield/DWD): 1.98e+10 photons/MGy
Average Dose (Whole Crystal)              : 1.17 MGy
Average Dose (Exposed Region)             : 2.89 MGy
Max Dose                                  : 12.865 MGy
Average Dose (95.0 % of total absorbed energy threshold (0.99 MGy)): 4.22 MGy
Dose Contrast (Max/Threshold Av.)         : 3.05
Used Volume                               : 40.4%
Absorbed Energy (this Wedge)              : 1.49e-03 J.
Dose Inefficiency (Max Dose/mJ Absorbed)  : 8.7 1/g
Final Dose Histogram:
Bin  1,  0.0 to  0.1 MGy:  6.8 %
Bin  2,  0.1 to  3.4 MGy: 58.9 %
Bin  3,  3.4 to  6.7 MGy: 23.0 %
Bin  4,  6.7 to 10.1 MGy:  9.0 %
Bin  5, 10.1 to 13.4 MGy:  2.2 %
Bin  6, 13.4 to 16.7 MGy:  0.0 %
Bin  7, 16.7 to 20.0 MGy:  0.0 %
Bin  8, 20.0 to 23.4 MGy:  0.0 %
Bin  9, 23.4 to 26.7 MGy:  0.0 %
Bin 10, 26.7 to 30.0 MGy:  0.0 %
Bin 11, 30.0 MGy upwards:  0.0 %
```

Figure 3.9: RADDOSE-3D command line output corresponding to the input shown in Figure 3.5.

At the top a brief description of the experiment such as the crystal type and size, and the simulation resolution is given. This is then followed by the description of each data collection (wedge) with the associated dose values according to different metrics, which are explained in detail in Zeldin *et al.* (2013c,a). A dose histogram shows how well the dose is distributed throughout the crystal volume.
The output was generated using RADDOSE-3D version 1.1.1000.

RADDOSE-3D allows the selection of relevant output modules via command line parameters and the RD3D class (Section 3.2.2).

3.2.9 The WRITER class: Reporting

With the aim of keeping the analysing OUTPUT classes as focussed and generic as possible, it was decided to separate the code for analysis and for file operations. To this end, each OUTPUT object relies on one instance of a WRITER class to provide reliable processing of its analysis results.

The WRITER class is a subclass of JAVA.IO.OUTPUTSTREAM and provides two basic functions: A write(...) function to process output, and a close() function to signal the end of the output. A number of existing WRITER classes are shown in Figure 3.8. With these, output can be directed to a file, the command line output, or, of particular importance for the RADDOSE-3D web service (Section 3.5.3), to a database.

The versatility of the OUTPUT–WRITER arrangement can be demonstrated using the WRITER-MULTIPLE class: although each existing OUTPUT class only uses one WRITER object, the same output can be written to multiple destinations using a WRITERMULTIPLE object, which copies the output on to multiple WRITER instances.

In a similar fashion the family of WRITER classes could be extended to include, for example classes which receive plain text output, apply compression algorithms and pass the binary data on to another WRITER object. These could be particularly useful for modules producing a number of large files such as dose state progressions.

3.3 Development and quality assurance

Once any non-trivial programming project grows beyond a few files, or multiple developers are involved, software quality should be considered. Software quality can refer to functional quality, how well the software is suited for the intended task, or structural quality, which includes non-functional requirements such as maintainability.

RADDOSE-3D is an academic software project that will probably exist for some time and may see a number of generations of developers. The importance of documentation and maintainability was recognised from the outset and thus software quality was understood to be an essential part of the development process.

This section will describe the approach taken to ensure software quality in a setting with multiple developers working simultaneously on the same project (Section 3.3.1) using continuous integration (Section 3.3.2), testing (Section 3.3.3), static code analysis (Section 3.3.4), and optimisation for speed (Section 3.3.5).

3.3.1 Collaborative development

Whenever multiple developers are working on a single code base there is an issue concerning code synchronisation. Because everybody is working on a local copy of the source code, changes need to propagate between the different development environments in an organised fashion. Revision control and source code management systems, initially Apache Subversion[10] and later git[11] were used to facilitate simultaneous, collaborative software development. Using a revision control system, developers can look up the complete history of each file, track any changes, see who

[10] http://subversion.apache.org/
[11] http://git-scm.com/

last modified a particular piece of code, and download the entire repository not just in its most recent incarnation, but also at any point in time of its history. The git repository containing the RADDOSE-3D source code is publicly available at https://github.com/GarmanGroup/Raddose3D.

3.3.2 Continuous Integration using Jenkins

This RADDOSE-3D code repository is monitored by Jenkins[12], a continuous integration server.

Continuous integration denotes a software development practice where all developers 'continuously' merge their working copies with a central repository. This allows other developers to see the progress, and to advise and assist in the development process. This facilitates team work and also code quality, as code can be reviewed by other developers in small portions while it is written. Compliance to coding standards and the project design philosophy is much easier to ensure with small changes. However, it also means that untested code changes of other developers can interfere with ones own work and with the stability of the common code base.

Jenkins is notified of every source code change and automatically runs a number of tasks defined in an Apache ANT[13] buildscript. It recompiles the entire project and records compiler warnings and errors. All existing functionality tests (Section 3.3.3) are run to ensure RADDOSE-3D works as intended. The source code is then inspected using static code analysis tools, which highlight noncritical problems in the source code such as bad programming habits or inadequately commented code (Section 3.3.4). Finally, if the compilation was successful, a .JAR-archive containing a compiled version of RADDOSE-3D is produced.

For each compilation run, a full report including any compilation issues, unit test failures and code analysis results is generated. This allows quick identification of who caused the build to

[12] http://jenkins-ci.org/
[13] http://ant.apache.org/

become unstable, when it happened, and what the relevant changes in the code were.

3.3.3 Testing

Small changes in the code during development or code maintenance can have application-wide effects. For quality assurance, and to ensure that developers do not inadvertently introduce defects (bugs). RADDOSE-3D uses the unit testing framework TestNG (Beust & Suleiman, 2007).

A unit test is a piece of code that does not directly contribute to the main software project functionality. Each unit test executes some isolated parts of the main software together with predefined appropriate data, and verifies the actual behaviour of the software against the expected behaviour. These code-data 'units' usually concern low-level functions, but may also test overall software behaviour.

At the time of writing, RADDOSE-3D contains 58 unit tests. These unit tests cover internal functionality and ensure the proper handling of objects pertaining to the experiment, and ensure reliable results, for example by verifying the result of crystal rotation operations and depth calculations.

As more code is developed, more unit tests are written. Old unit tests are only removed when the corresponding project code is removed, therefore a large library of unit tests is accumulated over time.

3.3.4 Static Code Analysis

Static code analysis is the process of automatically[14] inspecting a program without executing it (Louridas, 2006; Novak *et al.*, 2010).

[14] If the inspection is performed by a human, the process is usually termed 'code review'.

Program analysis can range from testing coding guideline adherence, identifying overly complex or unreadable sections of code, and detecting common high-level coding error patterns[15]. Basic syntactical inspection, which is covered by the compiler, is not usually understood as code analysis.

RADDOSE-3D uses three different free and open source packages for static code analysis: CheckStyle, PMD, and FindBugs.

CheckStyle

CheckStyle[16] is primarily a coding standard compliance checker. It processes the source code and highlights minor code readability issues, such as missing spaces around operators, overly long lines, or the placement of parentheses on new lines[17]. An example analysis result of CheckStyle on RADDOSE-3D can be seen in Figure 3.10.

Missing comments for classes, class variables, functions and function parameters are reported. CheckStyle also flags comments in code matching certain patterns (e.g. `TODO` or `XXX`), indicating that the developer wanted to revisit this piece of code later.

A major reason for running CheckStyle is to ensure code is understandable for developers who did not write the original code, and thus to avoid possible future errors caused by unnecessary confusion. For example variable name collisions, e.g. local variables having the same name as function parameters, are pointed out. It encourages function parameters to be marked as immutable, which prevents changes in the parameter values within the function, making it easier for

[15] There are fundamental limitations to what code analysis can do: even the question whether a program terminates or continues to run forever is generally, and famously, undecidable (Turing, 1936).

[16] http://checkstyle.sourceforge.net/

[17] These trivial matters can be an unexpected cause of contention, for many developers have idiosyncratic coding styles which are usually not compatible with those of other developers. However, enforcing a single style throughout a project does improve readability.

File: CoefCalcRaddose.java Lines 36 to 55

```
36    * Description of the experiment.
37    */
38   private final String      raddoseInput;
39
40   /**
41    * Identified coefficients and density from last program run.
42    */
43   private double            absCoeff, attCoeff, elasCoeff, density;
44
45   public CoefCalcRaddose(final Double cellA, final Double cellB,
4
4
49       final int numDNA,
50       final List<String> heavyProteinAtomNames,
51       final List<Double> heavyProteinAtomNums,
52       final List<String> heavySolutionConcNames,
53       final List<Double> heavySolutionConcNums,
54       final Double solventFraction) {
55
```

Type	Class	Description
checkstyle	JavadocMethodCheck	Missing a Javadoc comment.
checkstyle	ParameterNumberCheck	More than 10 parameters (found 15).

File: CoefCalcRaddose.java Lines 200 to 219

```
200              + "using the -r command line argument");
201     throw new RuntimeException("Could not run RADDOSE executable.");
202   }
203
204   @Override
205   @SuppressWarnings("PMD.PrematureDeclaration")
206   public void updateCoefficients(final Beam b) {
207     String energy = String.format("ENERGY %g%n", b.getPhotonEnergy());
208     String phoSec = String.format("PHOSEC %g%n", b.getPhotonsPerSec());
209     String exposure = String.format("EXPOSURE %d%n", 10);
210     String debug = "DEBUG\n";
211
212     Process oldRD = runRaddose();
```

Figure 3.10: Source code quality problems identified by CheckStyle

The function CoefCalcRaddose(...) is not properly documented in the code. Additionally, this function is deemed overly complex as it requires 15 parameters.

Later on in the same file the use of a 'magic number' is detected in the function updateCoefficients(...). It is not obvious from the code why the developer decided to use the value 10 (as opposed to 11 or 9). CheckStyle recommends that the developer should use a named constant instead, and then the choice of the particular value 10 should be justified in the accompanying comment.

other programmers to follow the program logic. CheckStyle also warns against 'magic numbers', which are numeric constants used directly in the code, and not explicitly defined and named at the top of a source code file.

The more advanced parts of CheckStyle also concern interface quality. For example it flags functions having a large number of parameters, classes having too many functions and inconsistent naming of functions, variables or constants.

PMD

PMD[18] also processes the source code of a program. In contrast to CheckStyle it focuses less on the source code layout and presentation and more on higher level programming patterns and common pitfalls.

Its extensive ruleset includes patterns examining source code complexity, performance, parallelisation, and internationalisation issues. Its recommendations can lead to faster program execution, lower memory usage, and the prevention of memory leaks by properly closing open resources. The PMD output, when run on RADDOSE-3D, is shown in Figure 3.11.

PMD can detect unused and unnecessary code. Unused code is usually written unintentionally, and may lead the programmer to expect a different program behaviour. PMD also highlights code that violates common object oriented programming principles, which can be very helpful as these violations can cause bugs which are very difficult to trace.

PMD provides examples for and the reasoning behind each warning (Copeland, 2005). Not every warning corresponds to an actual problem. Specific warnings can therefore be disabled for certain lines, functions or the entire project.

[18] http://pmd.sourceforge.net/

PMD Result

Warnings Trend

All Warnings	New Warnings	Fixed Warnings
70	0	0

Summary

Total	High Priority	Normal Priority	Low Priority
70	7	52	11

Details

Packages | Files | Types | **Warnings** | Details | High | Normal | Low

File	Package	Line	Priority	Type	Category
CoefCalcFromPDB.java:614	se.raddo.raddose3D	614	Normal	UseLocaleWithCaseConversions	RADDOSE3D
CoefCalcFromPDB.java:645	se.raddo.raddose3D	645	Normal	PositionLiteralsFirstInComparisons	RADDOSE3D
CoefCalcFromPDB.java:808	se.raddo.raddose3D	808	Normal	UseLocaleWithCaseConversions	RADDOSE3D
CoefCalcFromPDB.java:826	se.raddo.raddose3D	826	Normal	UseLocaleWithCaseConversions	RADDOSE3D
CoefCalcFromPDB.java:834	se.raddo.raddose3D	834	Normal	AvoidCatchingGenericException	RADDOSE3D
CoefCalcFromParams.java:31	se.raddo.raddose3D	31	Normal	ExcessiveParameterList	RADDOSE3D
CoefCalcRaddose.java:20	se.raddo.raddose3D	20	Normal	CyclomaticComplexity	RADDOSE3D
CoefCalcRaddose.java:45	se.raddo.raddose3D	45	Normal	CyclomaticComplexity	RADDOSE3D
CoefCalcRaddose.java:45	se.raddo.raddose3D	45	Normal	ExcessiveParameterList	RADDOSE3D
CoefCalcRaddose.java:132	se.raddo.raddose3D	132	Normal	CyclomaticComplexity	RADDOSE3D
CoefCalcRaddose.java:165	se.raddo.raddose3D	165	Low	AvoidInstantiatingObjectsInLoops	RADDOSE3D
CoefCalcRaddose.java:194	se.raddo.raddose3D	194	High	AvoidThrowingRawExceptionTypes	RADDOSE3D
CoefCalcRaddose.java:201	se.raddo.raddose3D	201	High	AvoidThrowingRawExceptionTypes	RADDOSE3D
CoefCalcRaddose.java:206	se.raddo.raddose3D	206	Normal	CyclomaticComplexity	RADDOSE3D
CoefCalcRaddose.java:206	se.raddo.raddose3D	206	Low	NPathComplexity	RADDOSE3D
CoefCalcRaddose.java:243	se.raddo.raddose3D	243	High	AvoidThrowingRawExceptionTypes	RADDOSE3D

Figure 3.11: PMD inspecting the RADDOSE-3D source code

PMD identified a total of 70 potential problems in the code. They range from too complex interfaces (ExcessiveParameterList in the 'Type'-column) and functions (CyclomaticComplexity and NPathComplexity), unclean Java exception handling (AvoidCatchingGenericException and AvoidThrowingRawExceptionTypes) to potential internationalisation (UseLocaleWithCaseConversions) and performance (AvoidInstantiatingObjectsInLoops) issues.

FindBugs

FindBugs[19] is an analysis software that, unlike CheckStyle and PMD, does not work on source code but rather on compiled Java bytecode. FindBugs is capable of recognising more than 300 common programming mistakes and bug patterns (Ayewah *et al.*, 2008).

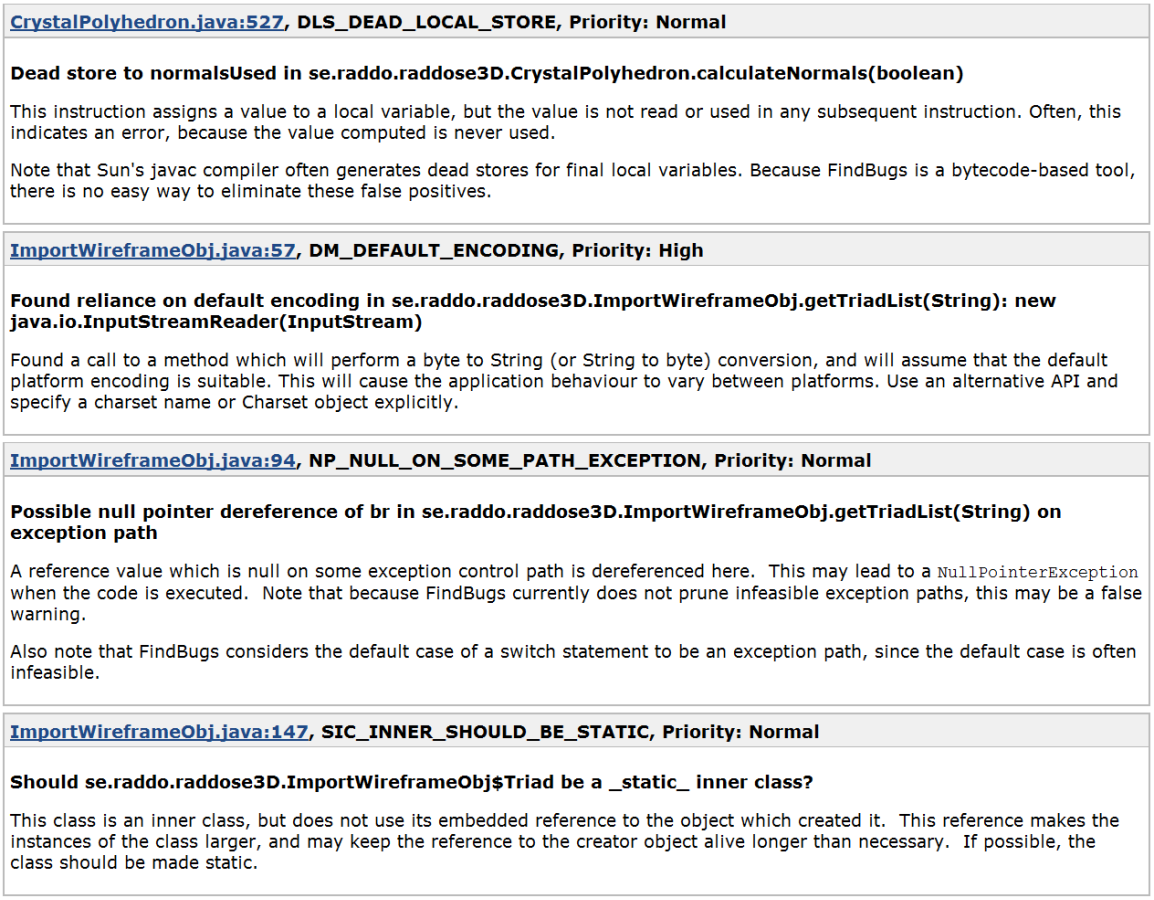
Due to its nature, FindBugs is only concerned with code flow, and not with source code layout. Most bug patterns are more advanced than those used by PMD, so FindBugs may for example warn about potential null pointer dereference bugs (Figure 3.12), which might cause the software to crash only in specific simulations.

Together, CheckStyle, PMD and FindBugs provide immediate, extensive feedback for newly written source code, and encourage the production of well-documented, readable, and robust code. This feedback, including the issues shown in Figures 3.10–3.12, were taken into account during the development of RADDOSE-3D and the program code was fixed accordingly.

3.3.5 Optimisation

During early development of RADDOSE-3D, the issue of simulation speed was deliberately not considered. Optimisation during early development often constitutes a waste of time and resources, as programmers do not yet know and usually have no intuition of which code paths are critical to actual runtime (Knuth, 1974, p. 268). However, for RADDOSE-3D to be useful at beamlines, particularly for the selection of data collection strategies, the calculations need to be quick, and should not take considerably longer than the determination of useful data collection strategies themselves.

[19] http://findbugs.sourceforge.net/



CrystalPolyhedron.java:527, DLS_DEAD_LOCAL_STORE, Priority: Normal

Dead store to normalsUsed in se.raddo.raddose3D.CrystalPolyhedron.calculateNormals(boolean)

This instruction assigns a value to a local variable, but the value is not read or used in any subsequent instruction. Often, this indicates an error, because the value computed is never used.

Note that Sun's javac compiler often generates dead stores for final local variables. Because FindBugs is a bytecode-based tool, there is no easy way to eliminate these false positives.

ImportWireframeObj.java:57, DM_DEFAULT_ENCODING, Priority: High

Found reliance on default encoding in se.raddo.raddose3D.ImportWireframeObj.getTriadList(String): new java.io.InputStreamReader(InputStream)

Found a call to a method which will perform a byte to String (or String to byte) conversion, and will assume that the default platform encoding is suitable. This will cause the application behaviour to vary between platforms. Use an alternative API and specify a charset name or Charset object explicitly.

ImportWireframeObj.java:94, NP_NULL_ON_SOME_PATH_EXCEPTION, Priority: Normal

Possible null pointer dereference of br in se.raddo.raddose3D.ImportWireframeObj.getTriadList(String) on exception path

A reference value which is null on some exception control path is dereferenced here. This may lead to a `NullPointerException` when the code is executed. Note that because FindBugs currently does not prune infeasible exception paths, this may be a false warning.

Also note that FindBugs considers the default case of a switch statement to be an exception path, since the default case is often infeasible.

ImportWireframeObj.java:147, SIC_INNER_SHOULD_BE_STATIC, Priority: Normal

Should se.raddo.raddose3D.ImportWireframeObj$Triad be a _static_ inner class?

This class is an inner class, but does not use its embedded reference to the object which created it. This reference makes the instances of the class larger, and may keep the reference to the creator object alive longer than necessary. If possible, the class should be made static.

Figure 3.12: Representative output of FindBugs identifying four potential issues in RADDOSE-3D.

In the first warning an assignment of a value to an otherwise unused variable is identified, which may indicate unfinished or incorrect code. The second warning highlights a character encoding issue, which may cause problems when the software is run on computers with different language settings. The third warning indicates a piece of code that may cause the program to crash in certain instances. The last message points out a possible performance issue, which may cause the program to run slower and use more memory than necessary.

The RADDOSE-3D execution time was monitored with VisualVM[20] (Figure 3.13). VisualVM runs in Eclipse[21] and provides a hierarchical breakdown of function calls and corresponding execution times. Software like VisualVM aids in identifying potential optimisation targets and may help in comparing the runtime behaviour of different approaches for solving identical problems, but it is limited on a fundamental level. Analogous to the well-known observer effect in particle physics, there is an observer effect in information technology: observing a running software process changes the runtime behaviour of that process.

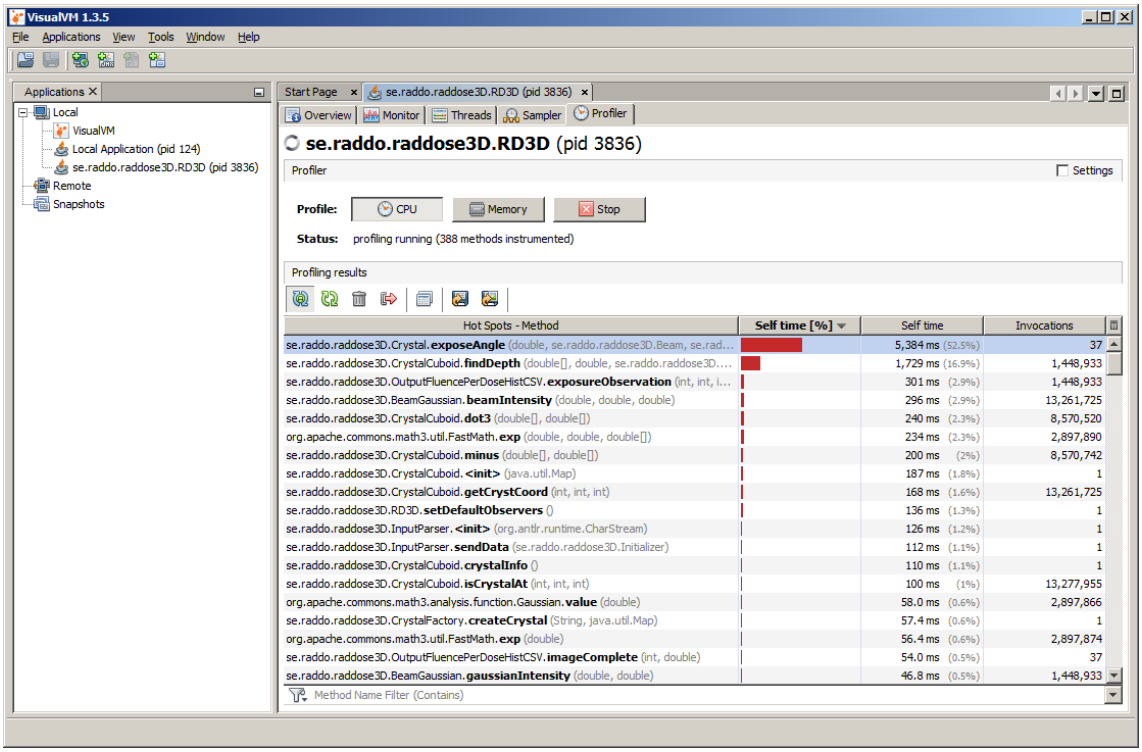


Figure 3.13: VisualVM profiler examining a running RADDOSE-3D simulation
While RADDOSE-3D is running in the background (not shown), VisualVM presents the programmer with a continuously updated list of called functions alongside information about the relative and absolute time spent in these functions and an invocation counter. After the RADDOSE-3D simulation is finished, the programmer can traverse the tree of functions calling other functions (not shown) and identify potential candidates for code optimisation. In subsequent runs the effects of these code optimisations can then be tested.

[20] http://visualvm.java.net/
[21] http://visualvm.java.net/eclipse-launcher.html

In the case of VisualVM, the absolute execution time is increased and the measured time spent in functions tends to be systematically overestimated. To verify the actual impact of any optimisation attempts, a compiled, modified version of RADDOSE-3D was run against an unmodified version in a number of simulations. The simulation results were compared to ensure that the simulations had run completely and correctly. In some cases an apparently huge improvement in runtime reported by VisualVM had little or no effect on actual runtimes. VisualVM is nevertheless a useful tool to highlight sections of the code which are of specific interest as optimisation targets.

The optimisation efforts on RADDOSE-3D are shown in Figure 3.14. The largest effect on execution speed was achieved by version 1.0.630, in which the simulation of dead zones of the crystal outside the active beam was disabled. Moving calculations out of the main simulation loop and caching intermediary or partial results caused measurable performance increases in the versions 1.0.745 (depth finding), 773 (crystal simulation) and 790 (beam parameters). Replacing a computationally expensive fluenceToDose(...) function by a per-image constant resulted in a small speed increase for version 1.0.783. No low-level optimisation of RADDOSE-3D (e.g. inspection of translated byte code) was attempted. It is not clear that this would have resulted in reasonable reductions in runtime considering the sizeable optimisation effort required.

The current speed of RADDOSE-3D appears to be suitable for tracking an experiment on a beamline in realtime.

3.4 Extending RADDOSE-3D

RADDOSE-3D provides a stable, solid, fast and extendible simulation package. A number of useful modules can be built on this platform with minimal or no changes required to the existing code.

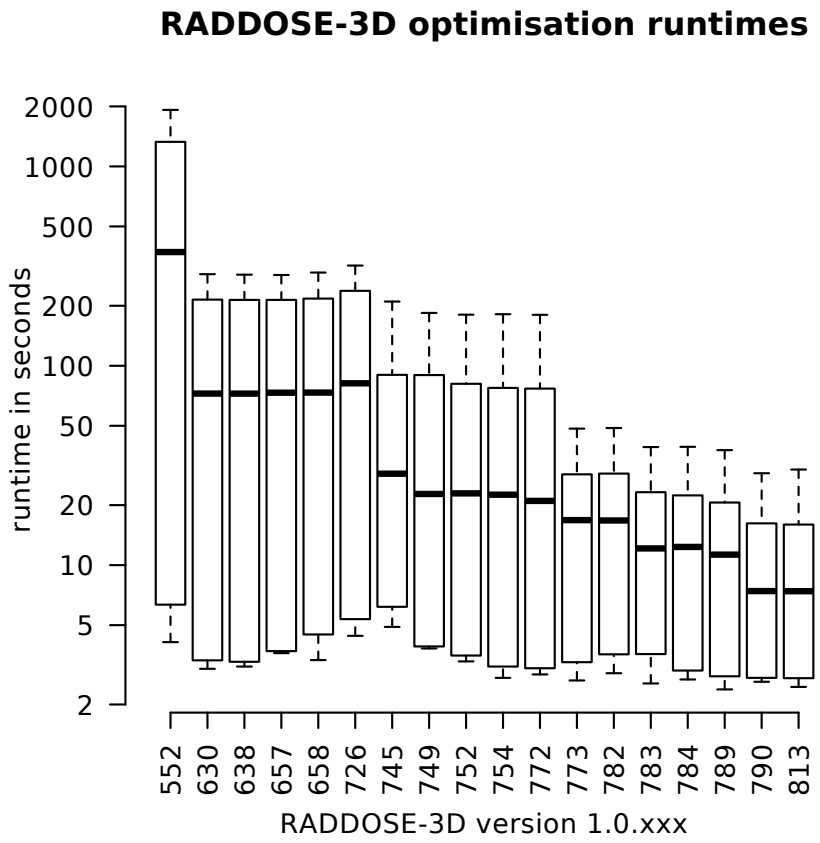


Figure 3.14: Optimisation of RADDOSE-3D during development

Five different RADDOSE-3D simulations were defined. Before the optimisation and after every optimisation run, all five simulations were run three times each to verify that the RADDOSE-3D output is unchanged, and to measure any changes in execution time. Shown here are the test execution times (on a log-axis) using 18 RADDOSE-3D versions, from version 1.0.552 to 1.0.813. The median run time is marked with a bold line, the box marks the interquartile range.

3.4.1 Crystal shapes and composition

New CRYSTAL classes can be developed, offering more complex or even arbitrary, including concave, crystal shapes. Used in conjunction with technology to digitise the three-dimensional crystal topography in an actual beamline experiment, this would greatly increase the modelling power of RADDOSE-3D. The crystal composition could at least be partially read from a PDB entry, allowing users to easily obtain attenuation and absorption figures which are much closer to the actual values than the average values based on the PDB survey. Work on both of these ideas is currently being undertaken by Helen Ginn.

Attenuation and absorption coefficients can currently be calculated easily for protein crystals, but RADDOSE-3D does not offer the option to easily calculate coefficients for other types of crystal, or to use user-specified coefficients. With an appropriate extension, RADDOSE-3D could be used to produce proper dose estimates for small molecule crystallography or small-angle X-ray scattering.

3.4.2 Beamline integration

In the same way, new BEAM classes can be introduced that more accurately describe existing experimental beams. RADDOSE-3D has already been used successfully with real beam profiles (Zeldin *et al.*, 2013a). Future versions of RADDOSE-3D could take beam wavelength spread into account, which would allow proper modelling of Laue crystallography and 'pink beam' experiments.

For beamline integration, the appropriate modelling of attenuators and arbitrary collimators (slits, apertures) could prove useful. Here, RADDOSE-3D could model the effects of attenuators not just on the total flux, but also on the flux distribution.

RADDOSE-3D could be complemented by an interface to take in a feed of goniometer, shutter and sample changer commands, and run in an automated mode online at a beamline. It could then eventually be used to track and display the live crystal damage state and provide on-line advice on data collection strategy decisions.

3.4.3 Improved dose metrics

RADDOSE-3D allows the calculation of new dose metrics, such as Diffraction Weighted Dose (Zeldin *et al.*, 2013a). Brooks-Bartlett *et al.* (2013) are currently investigating if and how these metrics can be improved further by using the dose progression information that RADDOSE-3D provides.

3.4.4 Current limitations

RADDOSE-3D currently offers no direct support for time-dependent, or more generally, beam-independent processes. The canonical use of RADDOSE-3D deals with dose distribution in a crystal caused by a beam-matter interaction, but it would be possible to extend the principle to additionally track the spatial distribution of other variables, such as a laser pumped electron excitation state or the local crystal temperature. Tracking the crystal temperature distribution could be implemented easily with the current code given an appropriate heating model, but heat dissipation and the effects of an external cryostream could only be modelled with some changes to the RADDOSE-3D core. With such changes made, RADDOSE-3D could for example be used to simulate damage processes such as the dark-progression of radiation damage at temperatures above 180 K (Warkentin *et al.*, 2011).

It is well known that even at 100 K, radiation damage is not confined to the exact location of the initial beam–matter interaction. For an incident beam energy of 18.5 keV, produced photo-

electrons have an estimated range of up to 4 µm with a characteristic spatial distribution based on beam polarity (Sanishvili *et al.*, 2011). In coarse grained simulations with large crystals and large beams, with a voxel size of well above $4 \times 4 \times 4\,\mu m^3$, this may not make much of a difference, but for simulations using both very small micro-beams and voxels, the current version of RADDOSE-3D will return a more uneven dose distribution than that which is expected in reality. An advanced dose deposition model could be devised which would take all these factors into account.

The dose distribution suffers from another inaccuracy, in that only the primary beam is modelled. Secondary beams (beams that were already elastically scattered inside the crystal) also suffer absorption, and thus the dose increases in areas of the crystal that are not placed directly in the beam. Similarly the probability of fluorescent escape or the escape of photoelectrons to the side of the crystal is not yet taken into account in RADDOSE-3D.

A more tangible and important limitation is that RADDOSE-3D does not currently allow the use of kappa goniometers. The rotation axis for data collection can only be offset (translated), but there is no option for a secondary rotation axis, or even a rotation axis that is not orthogonal to the beam axis.

3.5 Obtaining RADDOSE-3D

RADDOSE-3D is open source software, available free of charge under the GNU Affero General Public License (AGPL), Version 3[22]. RADDOSE-3D can be run in a number of ways: developers wishing to extend RADDOSE-3D can compile it from the publicly available source code (Section 3.5.1), users who want to run a number of large simulations can download RADDOSE-3D to their computers

[22] http://www.gnu.org/licenses/agpl-3.0.html

(Section 3.5.2), and casual users who want to quickly obtain some dose estimates can use the RADDOSE-3D web service (Section 3.5.3).

3.5.1 Building RADDOSE-3D from source code

The source code of RADDOSE-3D can be downloaded from https://github.com/GarmanGroup/Raddose3D. RADDOSE-3D can be compiled with a Java Development Kit[23], version 6 or newer. An installation of Apache ANT[24] is recommended, but not strictly necessary. With ANT installed, RADDOSE-3D can be compiled into a single .JAR file simply by running the command `ant build-debug jar` in the project directory. This compiles all relevant .JAVA source code files into .CLASS files and then aggregates these together with third-party libraries into a single .JAR archive.

3.5.2 Running RADDOSE-3D

RADDOSE-3D is distributed in a single .JAR file, which can be downloaded from http://raddo.se/. A Java Runtime Environment[25], version 6 or newer is required. RADDOSE-3D can be run from the command line with `java -jar raddose3d.jar`, which will bring up the command line help (Figure 3.2, page 93).

3.5.3 The RADDOSE-3D web service

Aside from installing a local copy of RADDOSE-3D, users can also submit jobs to the RADDOSE-3D web service at http://raddo.se/. This is particularly useful to quickly obtain dose estimates, as it does not require the installation of any software.

[23] http://www.oracle.com/technetwork/java/javase/index.html
[24] http://ant.apache.org/
[25] https://www.java.com/

The web service accepts jobs via two different interfaces. The first is a user-friendly interface (Figure 3.15), built by Jean-Paul Ebejer, that guides the user through the entry of the required parameters of an MX experiment. The user is immediately alerted to any missing parameters and also gets a short description for every parameter. However, it is not possible to utilise the whole spectrum of possible parameters. For example the user is currently limited to an MX scenario that only uses one beam, and exposes only one single wedge of the crystal.

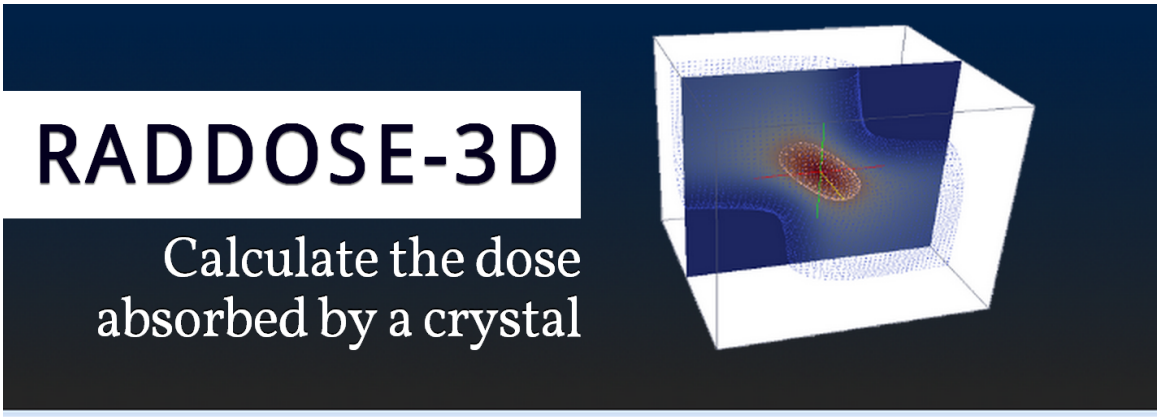
Alternatively the user can click on *'define a job manually'* and use the full range of available RADDOSE-3D features as described in the RADDOSE-3D user guide in Appendix C (page 214) and online[26].

The server version of RADDOSE-3D consists of a regular RADDOSE-3D version encapsulated in the RADDOSESERVER class. It regularly checks for new available jobs, and passes them on to an instance of RADDOSE-3D. For every successful simulation six images are produced showing the crystal from fixed angles and dose isosurfaces at fixed levels. These are currently set at 0.1 MGy, 20 MGy and 30 MGy, representing minimally exposed volume and volume exposed above the Henderson and Garman limits respectively. Example images can be seen in Figure 3.1 (Page 90).

The web service runs on an Apache HTTP Server[27] using PHP[28] as a server side scripting language connected to a MySQL database server[29] to store submitted input files and results for a limited time.

As of the 8th of November 2014, the web service has been used by over 1,800 users from at least 40 different institutions[30] to run more than 3,000 simulations.

[26] http://raddo.se/rd3d/RD3D-User-Guide.pdf
[27] http://httpd.apache.org/
[28] http://php.net/
[29] http://www.mysql.com/
[30] Counted from the domains of mail addresses, when given by the users. E-mail addresses are deleted from the database six months after the last visit of the user, so this number is likely higher.



Email
Get an email notification when your job is ready

Job Title
A brief description of the job

Crystal

Crystal Type
Cuboid or Spherical
Cuboid

Crystal Dimensions*
Dimensions of the crystal in X,Y,Z in µm. Z is the beam axis, Y the rotation axis and X completes the right handed set
Positive real dimensions should be specified!

Pixels per Micron
*The resolution of the grid used to calculate dose in voxels/µm. Defaults to **0.5**.*

Absorption Coefficient
*The method for calculating the absorption and attenuation coefficients. **Average** uses average atomic crystal compositions based on a PDB survey. **RADDOSE** explicitly calculates the coefficients using the crystal parameters.*
Average

Beam

Beam Type*
*Either **Gaussian** or **TopHat**. If it is **Gaussian**, you must*
Gaussian

Figure 3.15: RADDOSE-3D web service interface, programmed by Jean-Paul Ebejer

3.6 Further development

The modular and open architecture of RADDOSE-3D has already enabled the use of RADDOSE-3D for purposes other than basic dose calculations:

Zeldin *et al.* (2013c) demonstrated how RADDOSE-3D can be used to compare different data collection strategies in terms of dose distribution and efficient use of available crystal volume. More recently, RADDOSE-3D has been extended to simulate small-angle X-ray scattering (SAXS) experiments.

Improved support for small molecule chemical crystallography is planned for the future. While RADDOSE (and RADDOSE-3D) can already be used to calculate doses for these experiments (Horton *et al.*, 2014), it makes assumptions regarding solvent content and constituents based on macromolecular crystallography, which easily leads to incorrect dose estimates.

RADDOSE-3D is now used for data collection strategy evaluation at the new Massif-3 microfocus beamline (ESRF, Grenoble), and first steps are being undertaken at the I02 macromolecular crystallography beamline (Diamond Light Source, Didcot) to enable RADDOSE-3D to follow and guide the user through the entire diffraction experiment, from the initial dose-aware evaluation of data collection strategies to the final estimation of experimental doses for collected datasets.

Chapter 4

Quantifying Radiation Damage in Real Space

Quantifying Radiation Damage in Real Space

In this chapter another approach for the quantification of radiation damage is explored. In contrast to the earlier work on B_{Damage} (Chapter 2) this method is based on having multiple datasets from the same protein crystal and measures radiation damage based on the decay of electron density around affected residues[1].

A brief overview of the theory behind electron density maps is given in Section 4.1, and the possibility of quantifying specific radiation damage based on the decay of real space electron density is explored.

To test the validity of this approach, the electron density decay is calculated for radiation damage series previously obtained by Choi (2013) and Weik *et al.* (2000) to analyse the disulphide bond decay in *Gallus gallus* lysozyme and *Torpedo californica* acetylcholinesterase in Section 4.2.

To identify possible stabilising or sensitising effects of the local environments of residues, an extensive radiation damage study is carried out on 21 crystals of three mutants of a *Homo sapiens* signalling protein inhibitor (RhoGDI) protein and presented in in Section 4.3. Obtained by surface-entropy reduction (Longenecker *et al.*, 2001; Derewenda, 2011), these mutants have a high sequence identity, yet crystallise with different crystal contacts and thus provide different local environments for particular residues.

In the final Section, 4.4, the investigation into electron density decay is extended to a time-resolution limit using a new data collection and analysis protocol specifically designed for time-

[1] Parts of the work described in this chapter were presented at the 28th European Crystallography Meeting, Warwick, United Kingdom, August 2013 (Gerstel *et al.*, 2013).

resolved radiation damage research.

4.1 Radiation damage and electron density maps

This section will detail why electron density maps can provide a particularly useful tool in the detection and characterisation of specific radiation damage. For this, the sources of electron density maps and their interrelationship with atomic structural models first need to be explored. This is followed by a discussion of the effects of radiation damage on both electron density maps and atomic structural models.

4.1.1 Obtaining electron density maps

In a regular X-ray diffraction experiment the intensities (I) of a large subset of reflections are measured. These intensities can be used to calculate structure factor (F) amplitudes (proportional to the square root of the intensities; $|F| \sim \sqrt{I}$), which, together with the phases of the reflections (ϕ), can be transformed into a three-dimensional distribution of electron density using an inverse Fourier transform.

Unlike the intensities, the phases of the reflections are not directly *measured*. It is, in principle, possible to obtain phase information directly by experiment (Weckert & Hümmer, 1997). However, this requires a suitably stable protein with a small unit cell, a carefully controlled experiment using a specifically engineered diffractometer, and the alignment of very high quality crystals at room temperature (Weckert *et al.*, 1999), and is therefore, in practice, unfeasible.

To overcome this 'phase problem', other, far more practical, routes are used to *estimate* phase information. These include the use of phases derived from a known protein with a similar primary

sequence (molecular replacement), the collection of datasets from heavy-atom derivatives (single or multiple isomorphous replacement; SIR, MIR) which may include the use of the signal from anomalous scattering (SIR, MIR with anomalous scattering; SIRAS, MIRAS). Datasets can be collected from the same crystal containing a heavy element at different X-ray wavelengths around the absorption edges of that particular element, causing the signal of the heavy element to vary between the datasets (MAD). Further techniques include single-wavelength anomalous dispersion (SAD), and radiation-damage induced phasing (RIP; Ravelli *et al.*, 2003; c.f. Section 2.1.2, page 28), see Taylor (2010) for a review. It is also possible to determine the phases *ab initio* from the intensities alone. This approach, called the 'direct method', has been demonstrated to work for proteins of more than 2,000 non-hydrogen atoms with heavy atoms present (Frazão *et al.*, 1999), and up to 1,914 non-hydrogen atoms without heavy atoms present (Hakim *et al.*, 2013).

The combination of phases ϕ obtained in such ways and the experimentally observed structure factors (F_{obs}) are then used to obtain a first electron density map through the inverse Fourier transformation:

$$\rho_{(x,y,z)} = \frac{1}{V} \sum_h \sum_k \sum_l F_{(h,k,l)} e^{-2\pi i(hx+ky+lz)}$$

A structure model of the protein can be then be fitted automatically or manually by placing atoms at the position of peaks in the electron density function ρ.

This first model of the protein structure forms the basis for an iterative refinement process. Using the atoms placed in the model a new set of calculated structure factors (F_{calc}) and, crucially, phases are obtained. For this step, refinement software such as REFMAC5 (Murshudov *et al.*, 2011) is used. Modern refinement software is based on statistical analysis of the available data

and the use of maximum-likelihood methods to determine the most probable model, given the experimental observations and the latest structure model as a starting point (Bricogne, 1997). The new phases resulting from the refinement process, combined with the original F_{obs}, are used to calculate a new, refined electron density map. The protein model may then be updated to fit this new electron density, and the process is repeated, until F_{calc} converges on F_{obs}.

4.1.2 Electron density maps, atomic B factors and occupancy

At the end of the refinement process stand two separate, yet intertwined, results.

The first of these is the protein structure model. It consists of a set of three-dimensional coordinates, an occupancy, and one or more atomic displacement parameters for each protein atom[2]. This structure model forms the most relevant basis for further research concerning that protein, such as the analysis of biological function, ligand interactions, or the re-use of the structural information in molecular replacement to solve another protein structure. The metric B_{Damage}, presented in Chapter 2, is also based on structure models.

The second result from the structure determination process is the final calculated electron density map. These electron density maps are available to download from the Electron Density Server (http://eds.bmc.uu.se/; Kleywegt *et al.*, 2004) if the structure factors were published alongside the protein structure model in the Protein Data Bank (PDB; Berman *et al.*, 2003). The deposition of structure factors became mandatory for PDB submissions of structure models based on X-ray crystallography on February 1, 2008 (Markley *et al.*, 2008).

Protein structure models provide an abstract, and in some aspects idealised, view of the protein. Each atom is defined as being located around a certain position with an exact uncertainty

[2] Hydrogen atoms are not usually part of the model in macromolecular crystallography, unless very high resolution diffraction data are available.

described by a sphere (isotropic B factor) or an ellipsoid (anisotropic B factor). Electron densities do not follow these idealised shapes or distributions: The electron density map, on which the abstract view is based, is comparatively noisy. It can be understood as the experimental *observation*, while the clean structure model corresponds to the *interpretation*.

The information available from an electron density map and the possible representations in the related protein structure model are compared in Figure 4.1. In this figure, the electron density along a line segment through the three-dimensional space of a protein crystal unit cell is considered. This electron density is modelled, as closely as possible overall, by placing atoms into the unit cell.

In this simplified 1-dimensional example, each atom additively contributes a Gaussian distribution to the modelled density. The parameters of the distribution depend on the number of electrons in the particular atom (a function of the element and its ionisation state), and its distance normal to the line segment, as well as on parameters in the protein structure model. These parameters are the occupancy value and atomic displacement parameters. The occupancy value ranges from 0 to 1, and describes the probability of encountering the atom at this position in any one randomly selected unit cell. In terms of the atom's contribution towards the electron density map, the occupancy is multiplied by the atom's distribution. The atomic displacement parameters describe the uncertainty on the atom's location, which correspond to a change in the variance of the distribution. A more detailed discussion of the occupancy and atomic displacement parameters can be found in Section 2.2.2, page 33.

It is obvious that it is possible to perfectly model the expected electron density by adding a large number of Gaussian distributions (placing a large number of atoms) with small variance

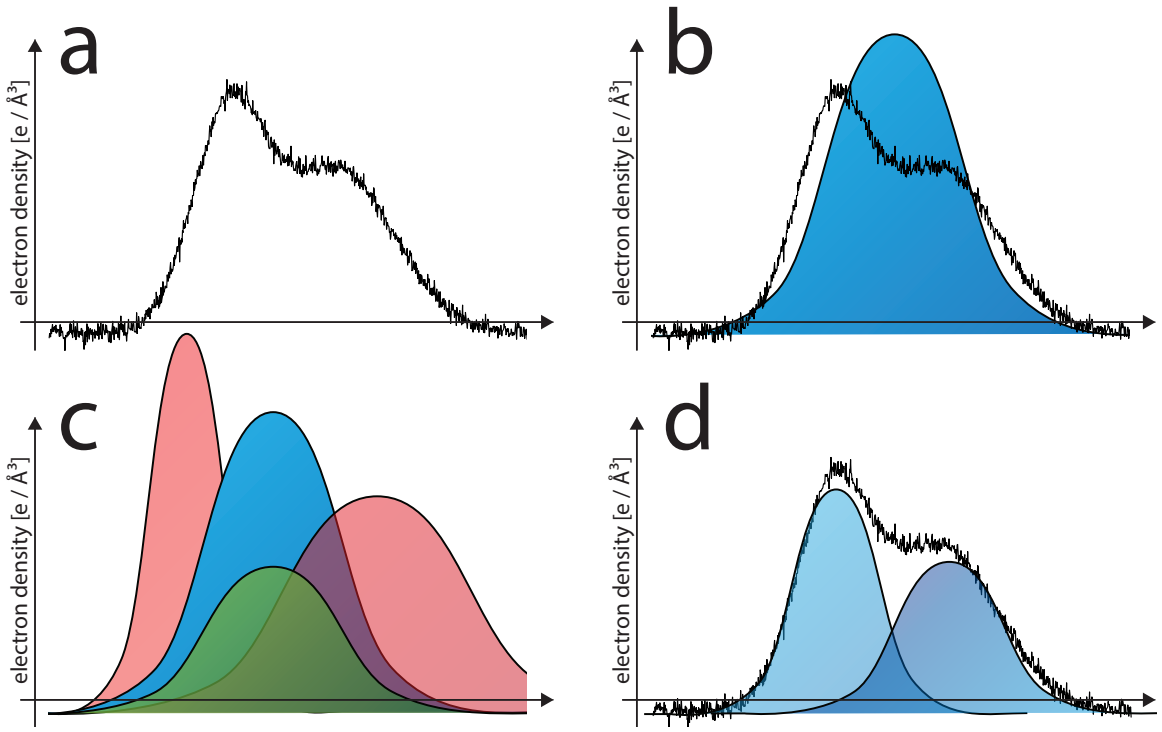


Figure 4.1: (a): A one-dimensional section of a normally three-dimensional electron density map is given.

(b): When building a protein structure model, atoms are placed in such a way that they explain the electron density as well as possible. Each atom contributes a single, essentially Gaussian, electron density distribution (blue distribution). A single atom can only approximate the observed electron density to a certain degree.

(c): The contribution of each atom is parametrised by the position of its peak (x, y and z coordinates in three dimensions, here only: x), its atomic B factor and occupancy. A smaller atomic B factor (red distribution on the left) results in a sharper and stronger electron density contribution. A larger atomic B factor (red distribution on the right) in a wider and flatter contribution. The integrated area (in three dimensions: volume) stays constant. In contrast to this, a reduction in occupancy (green distribution in front) reduces the integrated area without changing the spread of the distribution.

(d): By using all available parameters, and placing more than one atom (alternative conformations) and then adding their contributed electron density distributions, the identified electron density can be approximated much more accurately. An unwarranted increase in parameters eventually leads to an overfitted model: the model complexity increases in an attempt to explain experimental noise.

and appropriate multiplicative factors (small atomic B factors and small occupancies). It is indeed possible to 'improve' the overall quality of a protein structure model by building large numbers of alternative conformations. However, each new alternative atom which is placed increases the number of parameters in the model, eventually leading to an overfitted model even before the number of parameters grows beyond the number of observations (Brünger, 1997; Mayer *et al.*, 2010).

For this reason, the electron density for each atom is usually approximated using one single Gaussian distribution, and alternative conformations are avoided unless strong evidence for these is present. Raw electron density maps contain information about the electron density distribution that is not represented in the structure models, but may be valuable in understanding the process of specific radiation damage.

4.1.3 Electron density and specific radiation damage

In a radiation damage series of protein crystal structures, radiation damage can manifest itself in a number of ways. Throughout the series, ordered water molecules and other small molecules in the solvent disappear, as well as parts of the protein side chains, particularly of those on the protein surface and at the protein termini. In this context 'disappearance' means that parts of the protein are completely removed from the structure model, once the remaining evidence (the electron density) no longer justifies modelling those regions. While the loss will be gradual in the electron density maps, it can be represented both gradually (decreasing occupancy and/or increasing atomic B factor) or suddenly (removal of the atom) in the structure models.

Because all observations in X-ray diffraction experiments are a weighted time- and space-average of all illuminated protein crystal unit cells, disappearing electron density means that there

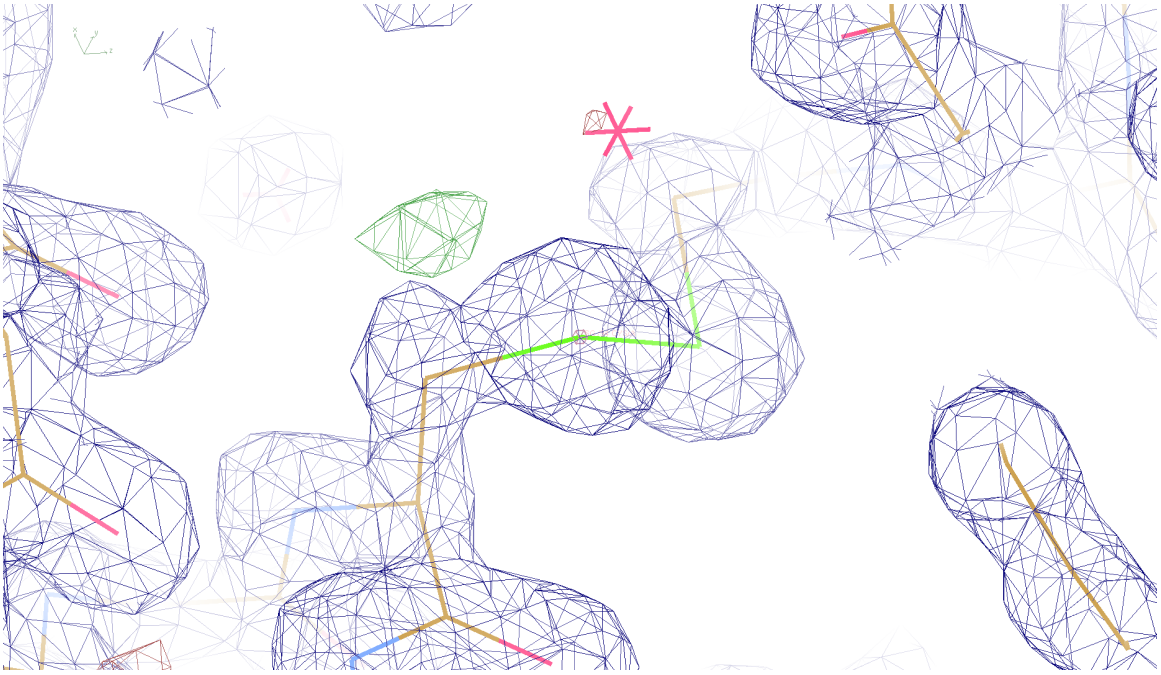


Figure 4.2: In the first dataset of the Choi (2013) lysozyme radiation damage series an alternative rotamer can be observed at CYS 94: additional electron density (green) is visible above the disulphide bond, indicating the alternative position that the CYS 94 sulphur occupies after the disulphide bond is broken in a number of unit cells. Electron density (blue) contoured at 1.3 σ, electron density difference map (green) contoured at 3.0 σ. Dataset was obtained with a Diffraction Weighted Dose of 4.7 MGy.

is enough disorder between the unit cells' contents that the electron density contributions of the relevant atoms are no longer localised. However if the same radiation damage processes were to occur in all the unit cells in unison, then residues and atoms would not seem to disappear, and instead seem to move or to adopt new conformations.

One particular example of this is the S_γ atom in the CYS 94 residue of Chicken-Egg White Lysozyme: It forms a disulphide bond with the CYS 76 residue of the same chain (Canfield & Liu, 1965). When this bond is broken by radiation damage, the CYS 94 side chain adopts a new conformation (Figure 4.2; Weik *et al.*, 2000).

In these cases, where specific radiation damage causes well defined structural changes, the identification and quantification of specific damage based on structural models, using metrics such

as B_{Damage} (Chapter 2), becomes increasingly difficult. When comparing the original state, where the CYS–CYS disulphide bond is in place, to the damaged state, with an alternative CYS conformation, all the atoms are present in both cases. While the atomic B factors may have increased in the latter case, the main information on the structural changes that have taken place is contained in the location of the atoms combined with the knowledge that a disulphide bond should, and used to be, present.

The movement of side chains and substructures is not taken into consideration in the B_{Damage} metric. In a radiation damage series of electron density maps, this kind of specific radiation damage becomes comparatively easy to identify: Both radiation damage symptoms, the increased disorder of an atom between unit cells, and the ordered move of the atom to another location, lead to a decrease of the electron density at its original location. It is this idea, the localised decrease of electron density caused by radiation damage, and, conversely, the quantification of radiation damage by observing the electron density decay, that will be explored throughout this chapter.

4.2 Decay of real space electron density

In this section the method of obtaining relevant electron density information is demonstrated using two radiation damage series. The theoretical underpinnings and the software that is necessary to acquire electron density distributions is presented. A method is then proposed to reduce these distributions to a useful, meaningful, and manageable summary statistic quantifying the cumulative radiation damage at specific protein locations.

4.2.1 Lysozyme as a model protein

Gallus gallus Chicken Egg-White Lysozyme (CEWL) is a common model protein in structural biology. It is readily available from commercial suppliers and crystallises quickly and reliably in an unusually broad range of conditions. Lysozyme accounts for approximately 1.5 % of all structure depositions in the PDB[3], making it one of the most studied proteins.

Lysozyme has four disulphide bonds between cysteines 6–127, 30–115, 64–80 and 76–94 (Canfield & Liu, 1965). These disulphide bonds exhibit different decay rates at 100 K, but so far a universal order has not been agreed upon. According to Weik *et al.* (2000) 6–127 is the most susceptible bond, followed by 76–94, which is partially cleaved at ∼10 MGy. The two remaining bonds, 30–115 and 64–80, are the least susceptible bonds, with no particular defined order between them. Ravelli & McSweeney (2000) arrived at a similar ordering, but identified the 30–115 bond to be more susceptible than the 64–80 bond. In contrast, recent work by Sutton *et al.* (2013) identifies 6–127 and 30–115 as the most susceptible bonds, with evidence for radicalisation already present at 0.14 MGy. Negative electron density begins to appear from 0.84 MGy and radiation damage can be certainly identified at 1.05 MGy. According to Sutton *et al.* (2013) the 64–80 disulphide bond is less affected, and the 76–94 bond is the least susceptible bond to cleavage. A summary of these findings is shown in Table 4.1.

4.2.2 Measuring electron density decay

To investigate the decay of electron density distributions, a radiation-damage series of 10 data-sets obtained from a single *Gallus gallus* Chicken Egg-White Lysozyme (CEWL) protein crystal

[3] A search on http://www.pdb.org returns 1,544 structure hits out of a total of 103,921 published structures (as of October 7, 2014; Berman *et al.*, 2003).

	Weik *et al.* (2000)	Ravelli & McSweeney (2000)	Sutton *et al.* (2013)
susceptible	CYS 6–CYS 127	CYS 6–CYS 127	CYS 6–CYS 127†
↓	CYS 76–CYS 94	CYS 76–CYS 94	CYS 30–CYS 115†
↓	CYS 30–CYS 115†	CYS 30–CYS 115	CYS 64–CYS 80
stable	CYS 64–CYS 80†	CYS 64–CYS 80	CYS 76–CYS 94‡

Table 4.1: Lysozyme disulphide bonds ordered by decreasing radiation damage susceptibility at cryotemperatures (100 K) according to three different sources in the literature. †No ordering between those two bonds is implied. ‡The appearance of an alternative conformation of the CYS 94 side chain is noted, casting doubt on the apparent stability of this disulphide bond.

held at 100 K was provided by Choi (2013). CEWL was obtained from Sigma Aldrich and crystallised by the hanging drop vapour diffusion method. The hanging drop was produced by mixing 1 µl of protein solution (30 $^{mg}/_{ml}$ CEWL in 100 mM sodium acetate) with 1 µl of well solution (8 % NaCl in 100 mM sodium acetate at pH 4.5), following the protocol given by Barker *et al.* (2009). Datasets were collected at 100 K at beamline I03 of the Diamond Light Source. The cumulative absorbed dose per dataset was 4.7 MGy (calculated using RADDOSE version 2; Paithankar *et al.*, 2009).

An electron density map can be created from a structure factor file (.MTZ) containing amplitude and phase information using FFT (Winn *et al.*, 2011). The resulting electron density maps are stored as a three-dimensional scalar field in .MAP files. These maps can then be cropped to the asymmetric unit with MAPMASK (Winn *et al.*, 2011) for further processing.

To obtain the electron density for all residues and atoms, the volume within this scalar field pertaining to each atom needs to be identified. The software SFALL (Winn *et al.*, 2011) can be used to create a .MAP file resembling an electron density map but with atom identifiers instead of electron density values at each field position. The volumes associated with each atom (region of interest) do not overlap, and not every position in the map is associated with an atom. A maximum

outward distance r from the assumed atom position is inferred from a user defined parameter V (default: 2.5) and the atomic B factor B:[4]

$$r = V \frac{\sqrt{B + 25\,\text{Å}^2}}{2\pi}$$

A larger value of V results in the assignment of a larger volume to each atom. Because the volume is still bounded by neighbouring atoms, it will expand out into solvent space for larger values of V. As the emphasis of the present work is on the dissipation or movement of electron density from a particular space, smaller values of V should result in a more sensitive metric. The idea of increasing the sensitivity of the metric will be explored in detail in Section 4.2.5 (page 148). For the investigation presented here, a value of $V = 1.0$ was found suitable. The current version of SFALL does not support anisotropic displacement parameters, and therefore B refers here only to the isotropic atomic B factor.

A sample resultant region of interest around the ILE 58 residue of Chicken Egg-White Lysozyme is shown in Figure 4.3. For the work described in this chapter, a PERL script (atommap.pl) was developed to map both .MAP files, one defining the regions of interest, the other containing the actual electron density values, onto one another and thus extract all relevant scalar electron density values for each atom.

This results in an electron density distribution for each atom in the structure. These distributions can be compared between atoms or residues and tracked for each atom or residue through time. The electron density decay caused by radiation damage can be visualised by obtaining these distributions for a series of electron density maps (Figure 4.4). For such a time-resolved series

[4] This equation can be found in the SFALL manual (Winn *et al.*, 2011), section 'Keyworded Input', keyword 'VDWR'.

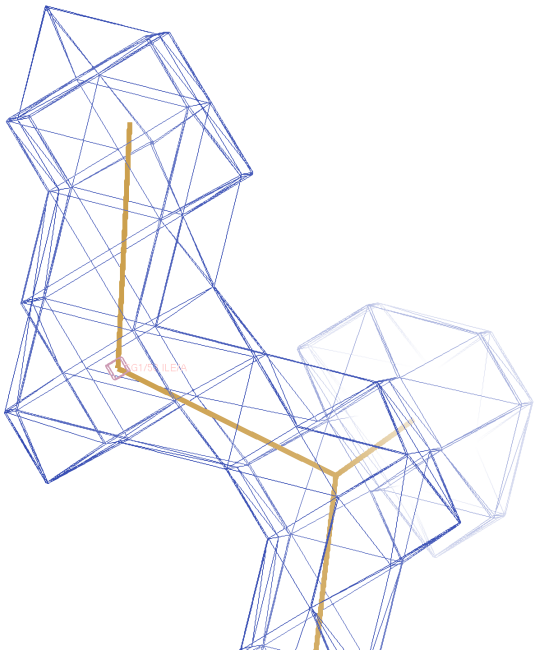


Figure 4.3: Boundary marking the region of interest around the ILE 58 residue side chain in a Chicken Egg-White Lysozyme structure model. Regions of interest were identified as described, and saved into a .MAP file. These could then be loaded and visualised in Coot (Emsley *et al.*, 2010) alongside the structure model. Note that although the boundary is treated and displayed like an electron density map, it does not correspond to the actual electron density in any way.

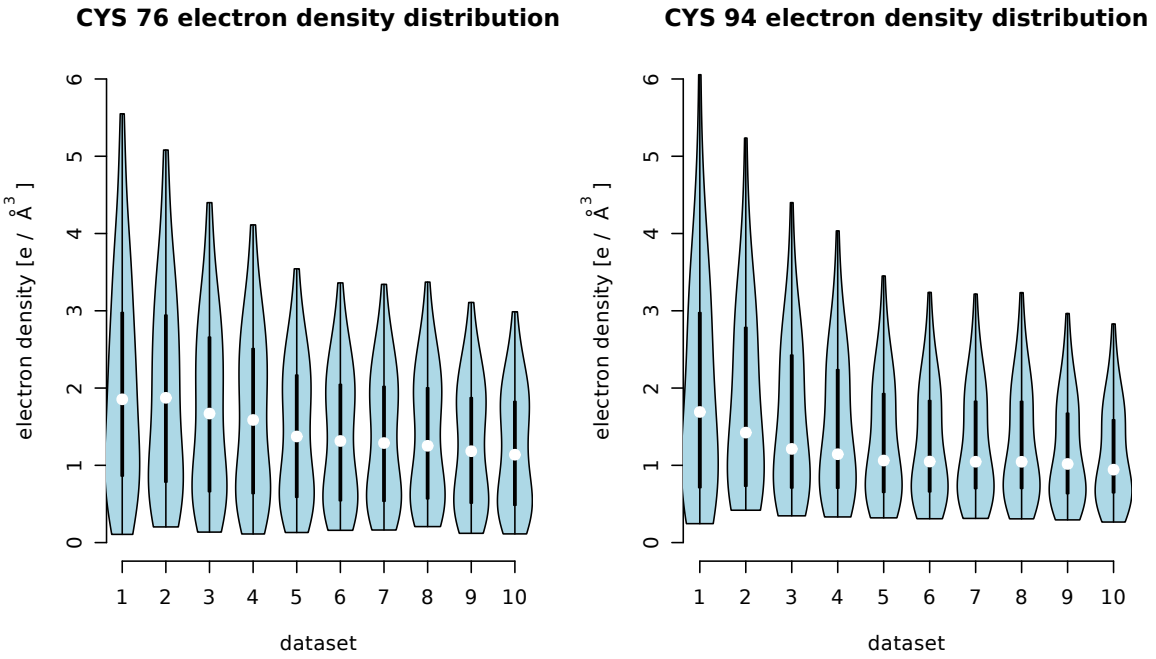


Figure 4.4: Electron density distributions of CEWL CYS 76 (**left**) and CYS 94 (**right**) residue side chains in a violin plot.
Violin plots are extended box plots (Hintze & Nelson, 1998; Adler, 2005), which allow the direct comparison between multiple distributions, by showing their kernel density plots (blue) vertically alongside one another. For each distribution the white dot marks its median value, and the black band in the centre of the density marks the interquartile range. The width of each distribution represents the number of observations, in this case the number of electron density volume elements, at a particular electron density. Both residues show a marked electron density decay through the ten datasets in the radiation damage series. Data collected by Choi (2013) with a corresponding dose of 4.7 MGy per dataset.

the atom locations of the model corresponding to the first, least-damaged dataset are used to identify the regions of interest.

As is apparent from Figure 4.4, suitable summary statistics are required to compare these electron density distributions over time, particularly when the multiple distributions of multiple atoms are involved.

4.2.3 Summary statistics and disulphide susceptibility

A number of summary statistics, including the maximum, mean, and median, for each electron density distribution were evaluated using the lysozyme disulphide bonds of the Choi (2013) damage series.

Intuitively, it may seem that the sum of all electron density values provides useful information about the ordered electron density of each atom. Since the volume assigned to each atom is kept fixed for a radiation damage series, the sum and the mean are related by a constant factor. Both the mean and the maximum of a distribution are affected by outliers. The electron density map is a result of an inverse Fourier transform, which may produce strong peaks, which are evident from the distributions shown in Figure 4.4.

All four disulphide bonds are clearly visible in the reference electron density map of the first dataset in the Choi (2013) series at 4.7 MGy. There are no obvious signs of specific radiation damage for three of the four disulphide bonds. At CYS 94 there is some indication of cleavage with electron density present at the position of an alternative rotamer conformation, indicative of a broken disulphide bond (Figure 4.2, page 136). This is consistent with the observation of the alternative rotamer from 1.05 MGy by Sutton *et al.* (2013) and at 10 MGy by Weik *et al.* (2000).

The summary statistics of each electron density distribution of each cysteine side chain were plotted for each dataset (Figure 4.5, left column). To assess the decay state of entire disulphide bonds, the summary statistic of the two constituting cysteines can be averaged together (Figure 4.5, right column).

The electron density decay process is clearly visible. The magnitudes of the densities for entire disulphide bonds may be correlated with their known susceptibility, but this is less clear. Since the

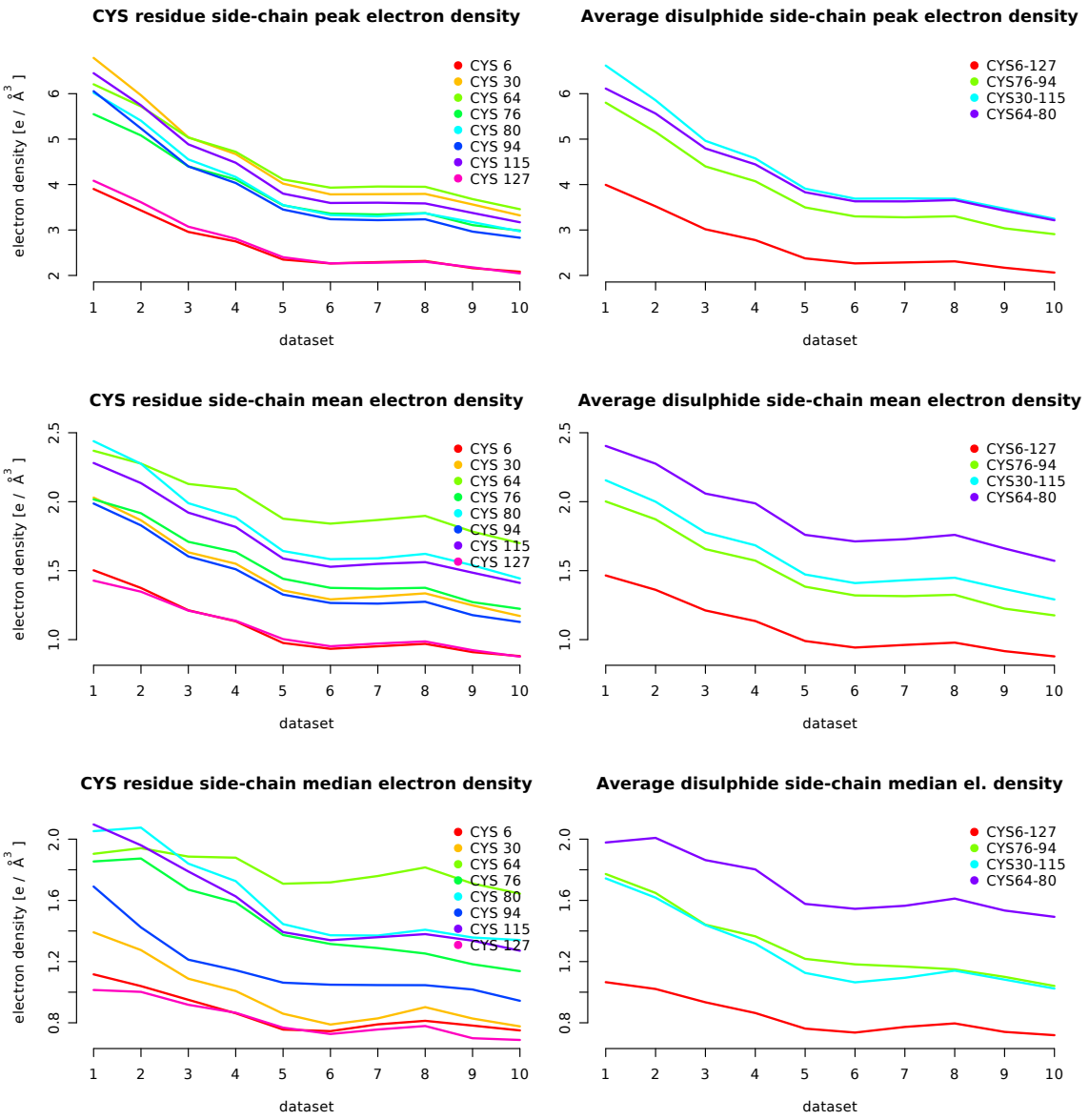


Figure 4.5: Effects of applying different summary statistics to the electron density distributions of disulphide bond side chains in the Choi (2013) CEWL radiation damage series.

All plotted summary statistics (**Top:** maximum, **Middle:** mean, **Bottom:** median) reflect the continuous decay of electron density resulting from radiation damage affecting the disulphide bonds. With the individual summary metrics (**Left**) of the pair of cysteines making up the disulphide bond are averaged together (**Right**) the decay of the disulphide bonds can be assessed. With all three metrics the 6–127 disulphide bond shows the lowest electron density, consistent with it being the most susceptible disulphide bond (Table 4.1). The order of the remaining three disulphides is less clear. Overall, these results are more consistent with the observations by Weik *et al.* (2000) and Ravelli & McSweeney (2000) than with those by Sutton *et al.* (2013).

region of interest is defined using the atomic coordinates of the structure model from the first dataset, specific damage in the first dataset can affect all subsequently obtained electron density distributions. In this series, radiation damage is already visible in the first dataset (Figure 4.2, page 136). To further investigate the real space electron density distribution progression, another radiation damage series is required.

Further work on the Choi (2013) damage series included the calculation of correlation coefficients between the electron density maps of the first and n^{th} data set using OVERLAPMAP (Winn *et al.*, 2011), and the calculation of real-space R-values (Brändén & Jones, 1990) for each residue throughout the series, again with OVERLAPMAP. Neither of these two approaches appeared to distinguish clearly between the disulphide bonds, showing larger variation between the datasets than between the residues (data not shown).

4.2.4 Reinterpreting specific radiation damage to *Tc*AchE

The *Torpedo californica* acetylcholinesterase radiation damage series published by Weik *et al.* (2000) and revisited by Weik *et al.* (2002) consists of 9 datasets (Table 2.1) obtained from a single *Tc*AchE protein crystal. The dose per dataset was calculated by Weik *et al.* (2000) to be on the order of magnitude of 1×10^7 Gy. *Tc*AchE contains seven cysteine residues, of which six form three disulphide bonds (CYS 67–94, 254–265, 402–521). The seventh cysteine (CYS 231) is unbound. The investigation by Weik *et al.* (2000) was focused specifically on disulphide bonds, and included careful refinement of these cysteines as alanines to prevent model bias. Compared to the earlier CEWL series, this series provides a much better coverage of the entire specific damage process. Some early signs of radiation damage were identified on the most susceptible disulphide bond (254–265) in the first dataset (Weik *et al.*, 2000), and this disulphide is broken early in

the series (Weik *et al.*, 2002). The other two disulphide bonds do not exhibit signs of radiation damage in the first dataset, and elongate by 0.7 Å (402–521) and 0.3 Å (67–94) respectively throughout the series (Weik *et al.*, 2002) with an accompanying loss of electron density.

To analyse the electron density, the sulphur atoms were restored into the electron density of the first dataset using Coot (Emsley *et al.*, 2010). Instead of considering entire residues or complete side chains, it is sensible only to inspect the electron density on their most susceptible sites. In the case of cysteine residues, the sulphur atoms are most affected by specific radiation damage. Thus, the electron density was calculated for the region of interest around the cysteine sulphur atoms, using the same software as described above.

Figure 4.6 shows the median electron density of each cysteine sulphur normalised to its median electron density in the first dataset. The median electron density was selected as the measure of location, as it is a more stable statistic with regards to outliers than, for example, the mean.

The clear decay pattern allows the introduction of the concept of a residue half-life, which can be defined as the number of datasets[5] collected before the median electron density is reduced to half the median electron density in the first dataset. In this case the calculated half-lives (Table 4.2) are, again, based only on the sulphur atoms.

Electron densities obtained by X-ray diffraction are time-averaged across multiple exposures from one dataset and space-averaged across a large number of exposed unit-cells. The absolute electron density therefore includes a component relating to the intrinsic order of the disulphide bond. In contrast, the half-life relates mostly to the change in electron density, and thus the sensitivity of the disulphide bond to specific radiation damage as well as the overall loss of detail

[5] Ideally this would be defined in terms of dose rather than datasets. However, reliable dose figures are not available for this series. As the datasets were obtained with identical data collection strategies and without additional burn periods, using the number of the dataset ($1..n$) is an internally consistent metric.

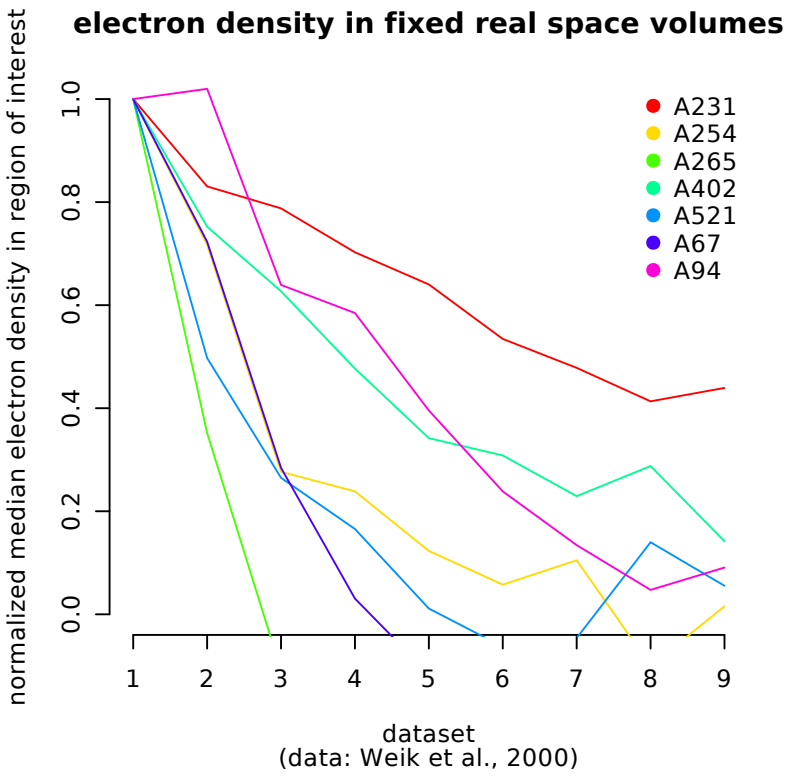


Figure 4.6: Median electron density decay of cysteine sulphurs in a *Tc*AchE radiation damage series (Weik *et al.*, 2000). Electron densities are normalised against those in the first dataset.

CYS residue	half-life (# datasets)	disulphide average
231	5.6	—
254	1.5	1.1
265	0.8	
402	2.8	1.9
521	1.0	
67	1.5	2.5
94	3.4	

Table 4.2: Half-life of *Tc*AchE cysteine residues based on the decay of the median-electron density in the region of interest based on the location of their sulphur atoms in the first dataset (Figure 4.6). Half-life values are given as the number of datasets corresponding to that level of electron density decay. One dataset corresponds to an absorbed radiation dose of the order of magnitude of 10^7 MGy (Weik *et al.*, 2000).
The determined half-lives of the disulphide bonds, calculated by averaging the half-lives of the two cysteines, agree with the experimentally observed susceptibility to radiation damage. Cysteine 231, which does not form a disulphide bond, is remarkably stable in comparison.

due to global radiation damage. The effects of global radiation damage would affect all copies of chemically identical structures, such as all disulphide bonds, at the same rate. A comparison of half-lives of identical groups therefore equates to comparing their susceptibility to specific radiation damage.

4.2.5 Further restriction of the real space region of interest

The analysis of the *Tc*AchE series focused specifically on cysteine sulphurs, and did not consider the electron density of the remaining side chain. This is sensible, as the side chain is obviously less affected by specific radiation damage than the disulphide bond itself. Naturally, this raises the question of whether further restriction of the region of interest would result in an improved separation of the disulphide bond half-lives. To test this idea, the region of interest as determined for the first dataset was reduced further by removing the volume elements by quantiles: i.e. removing the largest possible number of volume elements containing a total of 10%, 20% of the relevant electron density. Figure 4.7 shows the calculated half-lives of each cysteine sulphur as well as the half-life for each disulphide bond (x-axis) against the cut-off quantile. A cut-off quantile of 0 corresponds to using the entire region of interest as defined previously, while a cut-off quantile of 1 results in only tracking the electron density decay of a single value at one point in the electron density map corresponding to the position of the largest electron density peak in the original electron density map.

The reduction of the region of interest does not in fact improve the separation of the disulphide bond half-lives. Rather, the half-life metric appears to become unstable. This is to be expected, because with a reduced volume, the metric is based on fewer measurements and thus it is more strongly affected by random noise. It is therefore recommended not to restrict the region of

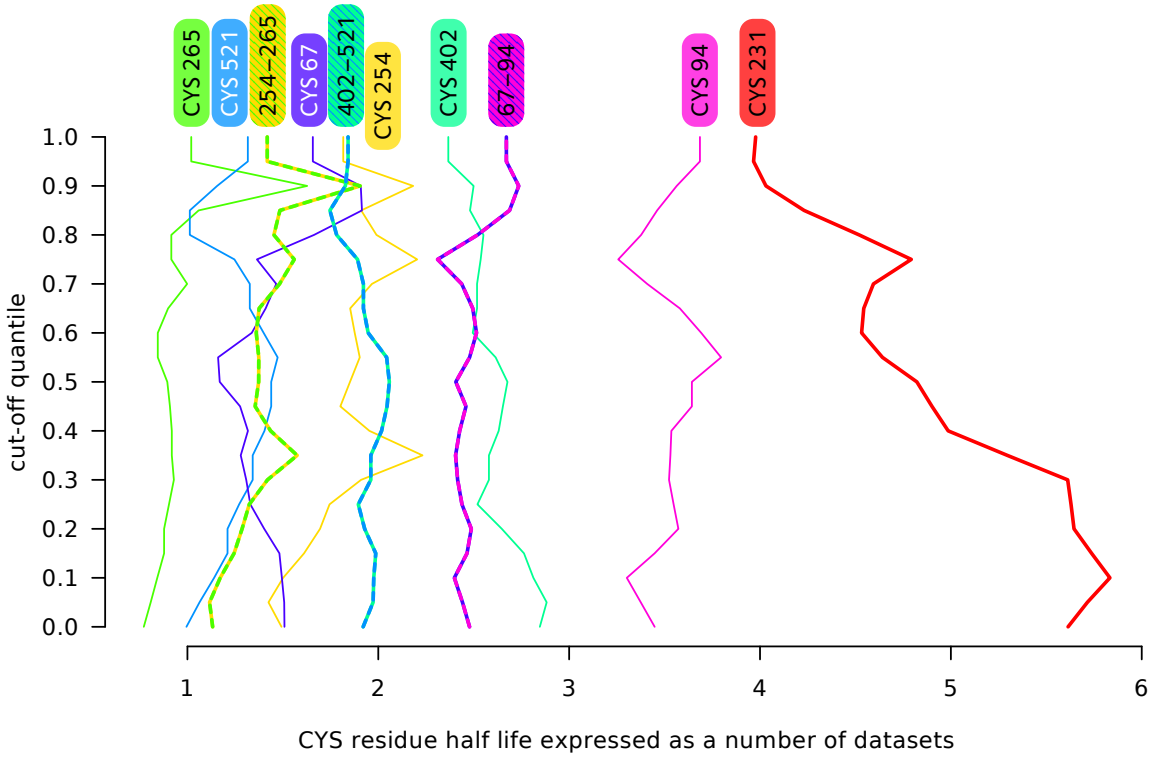


Figure 4.7: Effects of the restriction of the regions of interest by quantile on the calculated half-lives of cysteine residues in a *Tc*AchE radiation damage series (Weik *et al.*, 2000). A low cut-off quantile corresponds to a larger region of interest. With a high cut-off quantile only the electron density in those regions with the strongest peaks in the initial dataset is taken into consideration. Bold lines of alternating colour show the average of two cysteines making up a disulphide bond.

interest based on the initial electron density distribution.

4.3 Experimental investigation

It has been demonstrated in the previous sections that the reduction in electron density can be used to not only qualitatively identify specific radiation damage, but also to quantify the rate at which specific damage develops at certain locations.

The analysis of the *Tc*AchE radiation damage series is interesting in particular, because this radiation damage series cannot be used in a B_{Damage} analysis. This is due to it having been refined with only two atomic B factors per residue, one for the main chain atoms and one for the side chain atoms. A meaningful calculation of B_{Damage} requires per-atom atomic B factors (Section 2.3, page 37; Section 2.7.1, page 77).

Proper quantification of specific radiation damage is a prerequisite to an investigation into the causes and processes of preferential specific radiation damage. Once reliable specific damage rates can be assigned to protein locations, the effects of alterations, such as changes in nearby residues, can be measured. One possible protein candidate for such studies is *Homo sapiens* RhoA GDP/GTP nucleotide dissociation inhibitor (RhoGDI; see PDB ID 1RHO for the original 2.5 Å resolution protein structure model; Keep *et al.*, 1997).

4.3.1 RhoGDI protein

The *Homo sapiens* RhoGDI protein is, in its wild type, difficult to crystallise. It has been used as a model protein in extensive studies on Surface-Entropy Reduction (SER; Longenecker *et al.*, 2001; Derewenda, 2011). By applying site-directed mutagenesis to replace high conformational-

entropy residues on the protein surface, such as lysine, glutamic acid or glutamine, with low-entropy residues such as alanine, threonine, or tyrosine, the tendency of a protein to crystallise can be greatly improved. RhoGDI has a large number of lysine and glutamic acid residues, making it particularly suitable for these studies. The highly flexible N-terminal region of RhoGDI, which can interfere with the crystallisation process, is usually truncated for SER studies by 23 or 66 residues (Longenecker *et al.*, 2001; Cooper *et al.*, 2007).

For the mutations used in this study, 66 residues on the N-terminal were truncated and replaced with an ALA-MET marker, followed by the residues 67–204 of the original protein. This chain of 140 residues still includes 13 lysines and 12 glutamic acids and therefore allows a large number of possible single or multiple mutation variants to be made. With SER, the mutation sites can be systematically chosen to allow these sites to serve as potential crystal contacts (Longenecker *et al.*, 2001; Mateja *et al.*, 2002). In a majority of cases the mutated sites either form crystal contacts themselves, or favour the formation of crystal contacts in their immediate proximity (Cooper *et al.*, 2007).

RhoGDI does not contain disulphide bonds, but it does contain a large number of glutamic and aspartic acid residues, which are also known to be affected by radiation damage. The primary sequences of the RhoGDI mutants are largely identical, which presents an opportunity for differential radiation damage research: observing the changes in the rates of specific radiation damage of the same residue between different mutants could make the effects of the local residue environment visible.

In a single recent study more than 40 different RhoGDI mutants were used (Cooper *et al.*, 2007). Out of these, quantities of purified protein of three mutant forms (CH, EA and DY;

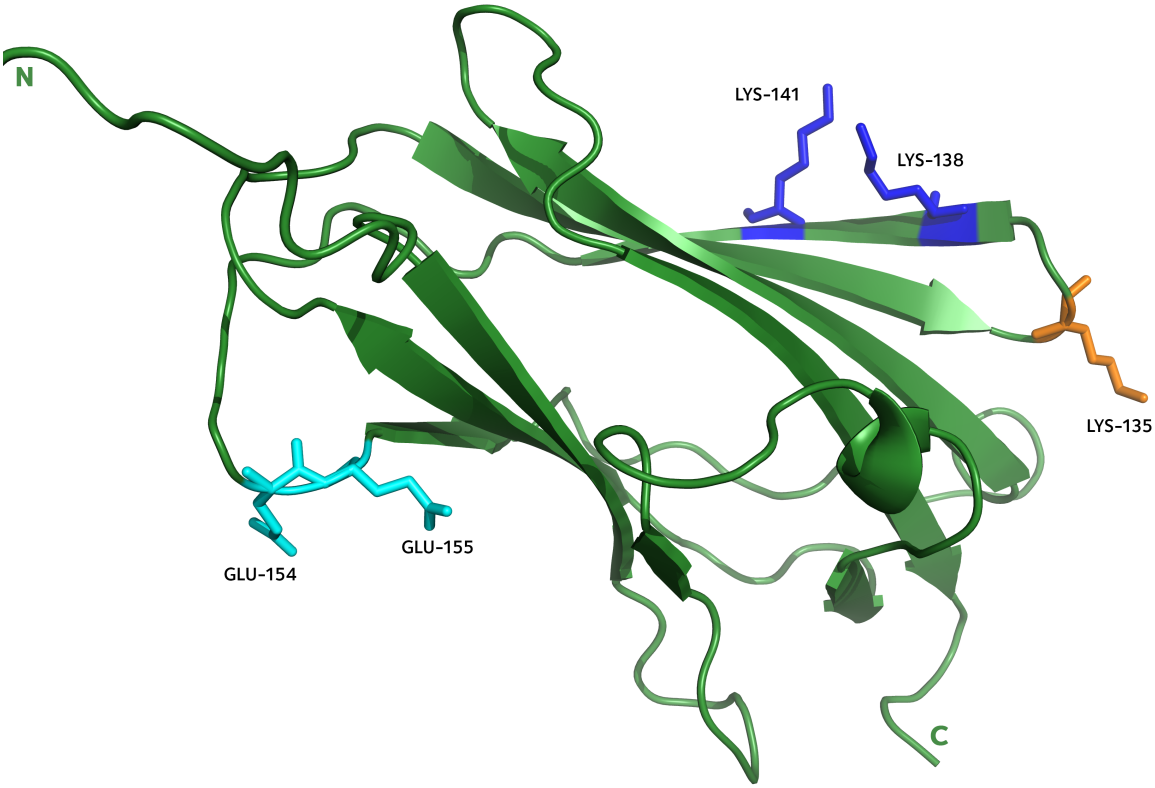


Figure 4.8: Overview of the RhoGDI protein mutants used in this study. RhoGDI mutations have an ALA-MET marker at the N-terminus followed by 138 residues of the original protein, starting from residue number 67. The structure model of the unmodified RhoGDI protein 1RHO (Keep *et al.*, 1997) is shown above. The mutation sites (Table 4.3) are marked. Graphic prepared using PyMOL (Schrödinger, LLC, 2010).

Figure 4.8, Table 4.3) were very kindly provided by Z. S. Derwenda, University of Virginia, USA for the purpose of our radiation damage research.

4.3.2 Materials and methods

The proteins were provided in a 150 mM NaCl, 50 mM Tris, 5 mM 2-mercaptoethanol solution at pH 7.5. The crystallisation solution constituents (Table 4.4) were selected to search around known, successful crystallisation conditions from the literature (Cooper *et al.*, 2007). The crystallisation drop consisted of a mixture of equal parts of centrifuged protein solution and filtered reservoir solution (generally 1 µl+1 µl, see Table 4.4). Solutions were placed and sealed in 24-

RhoGDI mutation		PDB	resolution
CH	K135H/K138H/K141H (Lysine → Histidine)	2JHS	1.95 Å
EA	E154A/E155A (Glutamic acid → Alanine)	2JHU	1.65 Å
DY	K138Y/K141Y (Lysine → Tyrosine)	2JIO	2.10 Å

Table 4.3: RhoGDI mutations kindly provided by Z. S. Derewenda, University of Virginia, USA, with the PDB entries of their structures (Cooper *et al.*, 2007).

well sitting-drop plates and left at room temperature for the vapour diffusion process. Crystals formed quickly within a few days.

Crystals were harvested and soaked in a cryoprotectant solution for approximately one minute. The cryoprotectant solutions (Table 4.4) were based on the relevant reservoir solution with a percentage of water replaced by glycerol. After soaking, the crystals were flash-cooled to 77 K by plunging them into liquid nitrogen, where they were stored until data collection.

In total, 21 radiation damage series were successfully collected from 20 well-diffracting crystals (Table 4.5). 12 series were from crystals of the RhoGDI CH protein, 3 from RhoGDI DY, and 6 from RhoGDI EA. Two independent radiation damage series were obtained from different ends of RhoGDI EA crystal p18 (Figure 1.4, page 16).

Crystal	RhoGDI	Crystallisation condition and Reservoir solution	Cryoprotectant
CH1A1L1	CH	1.4 M sodium citrate, 0.09 M HEPES, pH 7.4	10% glycerol[§]
CH1A5H1	CH	1.3 M sodium citrate, 0.10 M HEPES, pH 7.5	10% glycerol; 90 sec
CH1A6H1	CH	1.3 M sodium citrate, 0.11 M HEPES, pH 7.5	10% glycerol
CH1B4H1	CH	1.3 M sodium citrate, 0.11 M HEPES, pH 7.5	10% glycerol
CH1B6H1	CH	1.3 M sodium citrate, 0.09 M HEPES, pH 7.5	10% glycerol; 113 sec
CH1B6H2	CH	1.3 M sodium citrate, 0.09 M HEPES, pH 7.5	10% glycerol; 150 sec
CH1B6H3	CH	1.3 M sodium citrate, 0.09 M HEPES, pH 7.5	10% glycerol
CH1B6H4	CH	1.3 M sodium citrate, 0.09 M HEPES, pH 7.5	10% glycerol
CH1D3H1	CH	1.3 M sodium citrate, 0.09 M HEPES, pH 7.5	10% glycerol
CH1D3L1	CH	1.3 M sodium citrate, 0.09 M HEPES, pH 7.5	10% glycerol[§]
CH2A1H1	CH	1.4 M sodium citrate, 0.09 M HEPES, pH 7.4	10% glycerol; 130 sec
CH2A2H2	CH	1.4 M sodium citrate, 0.10 M HEPES, pH 7.4	10% glycerol; 130 sec
DY2A4	DY	18% PEG 6000, 95 mM Bicine, pH 7.5*	20% glycerol
DY3C1	DY	17% PEG 6000, 100 mM Bicine, pH 7.5[†]	15% glycerol
DY3C5	DY	19% PEG 6000, 100 mM Bicine, pH 7.5[†]	15% glycerol
M1C1	EA	23.5% PEG 4000, 100 mM Tris, 190 mM $LiSO_4$, 2.5% MPD[‡]	15% glycerol
M1C2	EA	23.5% PEG 4000, 100 mM Tris, 190 mM $LiSO_4$, 2.5% MPD[‡]	15% glycerol
M1C4	EA	24.0% PEG 4000, 100 mM Tris, 210 mM $LiSO_4$, 2.5% MPD[‡]	15% glycerol
M2C4	EA	2.05 M AmmSul, 0.2 M NaCl, 0.1 M NaCac	20% glycerol
p18	EA	2.05 M AmmSul, 0.2 M NaCl, 0.1 M NaCac	20% glycerol

Table 4.4: RhoGDI crystallisation summary. pH adjustment, when stated, using hydrochloric acid or sodium hydroxide. Crystallisation drops consisted of 1 µl protein solution + 1 µl reservoir solution unless stated otherwise. Reservoir solution 500 µl. Cryoprotectant solutions are based on the reservoir solution recipe, in which water is replaced by the cryoprotectant. Crystals were soaked for approximately 60 seconds.

AmmSul: ammonium sulphate; NaCac: sodium cacodylate

*crystallisation drop consisted of 0.6 µl protein solution + 0.6 µl reservoir solution; [†]0.7 µl + 0.7 µl; [‡]0.5 µl + 0.5 µl.

[§]cryoprotectant solution was produced by mixing equal amounts of a 20% glycerol solution and the reservoir solution.

Series	Beamline	Waveln.	Flux (ph/s)	# imgs.	ϕ/t per img.	Degrees collected (continuous collection in degrees @ transmission)
CH1A1L1	DLS I02	0.979 Å	3.5×10^{12}^	5,400	0.2°/ 0.2 s	1080° (360°@20%, 720°@20%)
CH1A5H1	DLS I02	0.979 Å	3.5×10^{12}^	12,600	0.2°/ 0.2 s	2520° (360°@20%, 1440°@20%, 720°@40%)
CH1A6H1	DLS I02	0.979 Å	3.5×10^{12}^	5,400	0.2°/ 0.2 s	1080° (180°@20%, 900°@20%)
CH1B4H1	DLS I02	0.979 Å	3.5×10^{12}^	14,400	0.2°/ 0.2 s[$]	2880° (360°@10%, 720°@20%, 720°@40%, 720°@100%, 360°@100%)
CH1B6H1	DLS I02	0.979 Å	3.5×10^{12}^	1,800	0.2°/ 0.2 s	360° (180°@20%, 180°@20%)
CH1B6H2	DLS I02	0.979 Å	3.5×10^{12}^	7,810	0.2°/ 0.2 s[‡]	1890° (360°@5%, 90°@20%, 1440°@20%)
CH1B6H3	DLS I02	0.979 Å	3.5×10^{12}^	16,200	0.2°/ 0.1 s[§]	3240° (360°@20%, 720°@20%, 720°@40%, 720°@100%, 720°@100%)
CH1B6H4	DLS I02	0.979 Å	3.5×10^{12}^	16,200	0.2°/ 0.1 s[§]	3240° (360°@20%, 720°@20%, 720°@40%, 720°@100%, 720°@100%)
CH1D3H1	DLS I02	0.979 Å	3.5×10^{12}^	12,600	0.2°/ 0.1 s[¶]	2520° (360°@20%, 720°@20%, 720°@40%, 720°@100%)
CH1D3L1	DLS I02	0.979 Å	3.5×10^{12}^	5,400	0.2°/ 0.2 s	1080° (360°@20%, 720°@20%)
CH2A1H1	DLS I02	0.979 Å	3.5×10^{12}^	20,700	0.2°/ 0.2 s	4140° (1800°@20%, 180°@20%, 41°@20%, 139°@20%, 900°@20%, 720°@40%, 360°@100%)
CH2A2H2	DLS I02	0.979 Å	3.5×10^{12}^	5,400	0.2°/ 0.2 s	1080° (360°@20%, 720°@20%)
DY2A4	DLS I03	0.975 Å	5.7×10^{11}*	64,800	0.1°/ 15 ms	6480° (9×720°@100%)
DY3C1	DLS I03	0.975 Å	5.7×10^{11}*	64,800	0.1°/ 25 ms	6480° (9×720°@100%)
DY3C5	DLS I03	0.975 Å	5.7×10^{11}*	50,400	0.1°/ 22 ms	5040° (7×720°@100%)
M1C1	DLS I03	0.976 Å	1.6×10^{12}•	14,402	0.2°/ 0.2 s	2880° (1440°@10%, 360°@10%, 720°@20%, 360°@60%)
M1C2	DLS I03	0.976 Å	1.6×10^{12}•	12,599	0.2°/ 0.2 s	2520° (48°@10%, 1392°@10%, 720°@10%, 360°@100%)
M1C4	DLS I03	0.976 Å	1.6×10^{12}•	11,700	0.2°/ 0.2 s	2340° (360°@20%, 1800°@20%, 180°@100%)
M2C4	DLS I03	0.976 Å	1.6×10^{12}•	16,964	0.2°/ 0.2 s	3393° (720°@10%, 2×720°@20%, 720°@40%, 513°@70%)
p18-T[†]	DLS I04	0.980 Å	2.0×10^{11}*	1,080	0.5°/ 1.0 s	1080° (3×360°@100%)
p18-B[†]	DLS I04	0.980 Å	2.0×10^{11}*	900	1.0°/ 0.7 s	900° (10×90°@100%)

Table 4.5: RhoGDI data collection summary. Beam energy 12.7 keV for all datasets.

[†]damage series obtained from different ends (top & bottom) of same crystal with no overlap and a buffer zone (Sanishvili *et al.*, 2011) in-between.
*flux estimated from measurements taken by beamline staff at the start of beamline run.
^further measurements: 1.4×10^{12}@40%, 6.6×10^{11}@20%, 3.1×10^{11}@10%, 1.5×10^{11}@5% transmission.
•further measurements: 1.1×10^{12}@70%, 9.7×10^{11}@60%, 6.1×10^{11}@40%, 3.0×10^{11}@20%, 1.4×10^{11}@10% transmission.
[‡]the first 160 images (2.25°/ 0.23s) are a low resolution pass (3.0 Å at detector side edges).
[§]0.2 s for the final 3600 images. [¶]0.2 s after the first 1800 images. [$]0.4 s for the final 1800 images.

Series	Primary dataset size	Unit cell [Å, °] a	b	c	α	β	γ	Resolution (Å) low	high shell	Complete-ness (%)	Multi-plicity	$I/\sigma(I)$	R_{merge}	R_{meas}
CH1A1L1	72°	75.5	75.9	91.1	90	90	120	53.14	2.35–2.29	98.8 (99.4)	4.0 (4.1)	15.8 (2.1)	0.05 (0.55)	0.05 (0.68)
CH1A5H1	72°	75.6	75.6	91.5	90	90	120	37.81	2.34–2.28	76.4 (79.3)	5.1 (4.8)	22.8 (2.1)	0.05 (0.74)	0.06 (0.91)
CH1A6H1	72°	75.3	75.3	90.7	90	90	120	52.95	2.31–2.25	99.7 (100)	3.9 (4.1)	22.1 (2.2)	0.03 (0.59)	0.04 (0.78)
CH1B4H1	72°	75.5	75.5	90.9	90	90	120	37.73	2.73–2.66	98.8 (99.3)	3.9 (4.1)	9.1 (1.5)	0.10 (0.79)	0.13 (1.01)
CH1B6H1	72°	75.2	75.2	90.9	90	90	120	30.29	2.76–2.69	86.2 (89.4)	4.4 (4.0)	19.8 (2.0)	0.06 (0.66)	0.07 (0.78)
CH1B6H2	432°	75.5	75.5	90.8	90	90	120	65.34	2.03–1.98	99.8 (99.8)	10.8 (4.1)	23.6 (2.4)	0.11 (0.51)	0.11 (0.68)
CH1B6H3	72°	75.7	75.7	91.0	90	90	120	37.38	2.25–2.19	99.1 (99.2)	3.9 (3.8)	22.8 (2.2)	0.03 (0.52)	0.04 (0.66)
CH1B6H4	72°	75.6	75.6	91.1	90	90	120	37.40	2.54–2.48	99.0 (99.5)	3.9 (4.0)	16.8 (2.0)	0.06 (0.64)	0.07 (0.81)
CH1D3H1	72°	75.5	75.5	91.2	90	90	120	37.75	2.97–2.89	94.1 (95.2)	4.2 (4.3)	11.6 (2.0)	0.10 (0.69)	0.13 (0.87)
CH1D3L1	72°	75.4	75.4	90.4	90	90	120	52.93	2.36–2.30	72.8 (76.2)	5.4 (5.3)	19.5 (2.3)	0.05 (0.64)	0.06 (0.78)
CH2A1H1	72°	75.5	75.5	91.0	90	90	120	34.87	2.43–2.37	89.9 (91.0)	4.4 (4.5)	19.3 (2.0)	0.06 (0.80)	0.07 (0.92)
CH2A2H2	72°	75.2	75.2	91.2	90	90	120	53.02	2.45–2.39	77.7 (81.5)	5.1 (4.6)	22.9 (2.1)	0.05 (0.69)	0.06 (0.84)
DY2A4	180°	32.2	55.9	38.8	90	108.2	90	36.88	2.04–1.99	99.6 (99.1)	3.4 (3.4)	13.6 (2.5)	0.05 (0.40)	0.07 (0.56)
DY3C1	180°	32.1	55.4	38.8	90	108.7	90	36.76	1.80–1.75	95.8 (72.0)	3.1 (2.0)	18.4 (2.2)	0.03 (0.34)	0.04 (0.44)
DY3C5	180°	32.2	55.5	39.1	90	108.3	90	37.11	2.24–2.17	99.4 (99.1)	3.3 (3.4)	10.1 (2.5)	0.06 (0.43)	0.09 (0.57)
M1C1	360°	57.8	63.8	85.4	90	109.1	90	63.84	2.09–2.04	99.7 (99.4)	6.6 (6.7)	13.1 (3.0)	0.09 (0.62)	0.10 (0.72)
M1C2	360°	57.4	63.8	85.4	90	109.1	90	41.20	2.28–2.22	95.7 (64.4)	6.3 (1.4)	11.6 (1.9)	0.13 (0.48)	0.14 (0.68)
M1C4	360°	57.9	63.7	85.4	90	109.0	90	80.74	2.04–1.99	99.9 (99.9)	6.7 (6.8)	12.1 (3.0)	0.09 (0.64)	0.10 (0.74)
M2C4	90°	130.4	130.4	162.9	90	90	120	66.06	2.08–2.03	99.8 (100)	4.9 (4.9)	12.1 (2.4)	0.07 (0.56)	0.08 (0.70)
p18-T	90°	131.0	131.0	164.5	90	90	120	93.40	2.55–2.49	97.4 (98.7)	5.4 (5.5)	13.0 (2.3)	0.07 (0.61)	0.09 (0.74)
p18-B	90°	131.0	131.0	164.7	90	90	120	93.44	2.63–2.56	100 (100)	5.4 (5.5)	12.5 (2.1)	0.08 (0.75)	0.10 (0.93)

Table 4.6: RhoGDI data integration summary for the first dataset of each series. Numbers in brackets are for the highest resolution shell. CH-mutants of the RhoGDI protein (series CH..., see Table 4.5) all crystallised in space group P $3_2\,2\,1$. DY-mutants (series DY...) and EA-mutants in PEG solution (series M1...) crystallised in space group P $1\,2_1\,1$. EA-mutants in ammonium sulphate solution (series M2C4 and p18-...) crystallised in space group R $3\,2$. Their unit cell constants are specified in the setting R $3\,2$:h.

From each radiation damage series a primary dataset was selected of size as detailed in Table 4.6. The size of the primary dataset was chosen so that a multiplicity between 2 and 4 can be expected in the reduced dataset. The first dataset of crystal CH1B6H2 additionally includes a 360° low-resolution pass.

The primary datasets were indexed and integrated using XDS and XSCALE (Kabsch, 2010) via the xia2 3DA II R pipeline (Winter, 2010; Winn *et al.*, 2011) using all images for indexing. Data from the M1C2 crystal proved impossible to process with this software pipeline and required DISTL (Zhang *et al.*, 2006) and LABELIT (Sauter *et al.*, 2004) (via the xia2 3DR pipeline) for successful indexing and integration. The p18 datasets were unsuitable for fine-sliced processing and therefore were indexed via the xia2 2DA pipeline, which uses MOSFLM (Leslie & Powell, 2007) for indexing and integration.

Integrated data were processed using POINTLESS (Evans, 2006) and subsequently scaled with AIMLESS (Evans & Murshudov, 2013). Initial structure models were obtained by molecular replacement with PHASER (McCoy *et al.*, 2007) using the known structure of the relevant RhoGDI mutant (Table 4.3).

For the crystals M1C1, M1C2, and M1C4 the initial molecular replacement resulted in an unsuitable, in the case of M1C4 unrefineable, structure model. The structures could be solved by integrating these datasets in P 1 and then carrying out molecular replacement with the CH mutation reference structure 2JHS, resulting in 8 protein copies in the unit cell, confirming the number expected by calculating the Matthews coefficient (Matthews, 1968) with the CCP4 program matthews_coef (Winn *et al.*, 2011). The correct space group, P 2_1, (with 4-fold non-crystallographic symmetry) was confirmed with ZANUDA (Lebedev & Isupov, 2014), and the da-

Series	R_{work}	R_{free}	r.m.s. dev. bond lengths	r.m.s. dev. bond angles	Ramachandran plot Allowed	Ramachandran plot Additional	Ramachandran plot Disallowed	Rotamer outliers
CH1A1L1	0.200	0.229	0.011	1.508	135 (97.83 %)	3	0 (0.00 %)	0 (0.00 %)
CH1A5H1	0.202	0.258	0.009	1.373	133 (96.38 %)	4	1 (0.72 %)	0 (0.00 %)
CH1A6H1	0.196	0.234	0.010	1.416	135 (97.83 %)	3	0 (0.00 %)	2 (1.61 %)
CH1B4H1	0.191	0.235	0.013	1.745	135 (97.83 %)	2	1 (0.72 %)	2 (1.61 %)
CH1B6H1	0.196	0.241	0.016	1.917	134 (97.81 %)	2	1 (0.73 %)	3 (2.42 %)
CH1B6H2	0.200	0.234	0.008	1.233	133 (96.38 %)	4	1 (0.72 %)	0 (0.00 %)
CH1B6H3	0.208	0.254	0.023	2.298	133 (97.08 %)	4	0 (0.00 %)	2 (1.61 %)
CH1B6H4	0.196	0.243	0.019	2.178	135 (98.54 %)	2	0 (0.00 %)	2 (1.61 %)
CH1D3H1	0.174	0.239	0.016	2.010	133 (96.38 %)	4	1 (0.72 %)	4 (3.20 %)
CH1D3L1	0.205	0.259	0.018	2.054	134 (97.81 %)	3	0 (0.00 %)	1 (0.81 %)
CH2A1H1	0.196	0.241	0.019	2.239	134 (97.81 %)	3	0 (0.00 %)	2 (1.61 %)
CH2A2H2	0.190	0.253	0.018	2.011	132 (97.06 %)	4	0 (0.00 %)	3 (2.42 %)
DY2A4	0.174	0.226	0.018	2.074	132 (95.65 %)	6	0 (0.00 %)	4 (3.25 %)
DY3C1	0.214	0.262	0.015	1.840	133 (96.38 %)	5	0 (0.00 %)	4 (3.25 %)
DY3C5	0.206	0.247	0.013	1.676	133 (97.08 %)	4	0 (0.00 %)	4 (3.25 %)
M1C1	0.224	0.277	0.017	2.087	508 (93.90 %)	18	15 (2.77 %)	21 (4.36 %)
M1C2	0.250	0.300	0.016	1.933	513 (95.00 %)	21	6 (1.11 %)	32 (6.64 %)
M1C4	0.237	0.277	0.020	2.175	520 (96.30 %)	12	8 (1.48 %)	31 (6.43 %)
M2C4	0.249	0.281	0.011	1.548	268 (98.17 %)	4	1 (0.37 %)	10 (4.13 %)
p18-B	0.213	0.251	0.013	1.764	266 (97.08 %)	8	0 (0.00 %)	13 (5.37 %)
p18-T	0.213	0.268	0.013	1.702	267 (97.45 %)	7	0 (0.00 %)	10 (4.13 %)

Table 4.7: Refinement statistics for RhoGDI structures from primary datasets.

tasets were reprocessed accordingly. The failure of the initial molecular replacement was probably caused by incorrectly specifying the dimeric domains of the reference structure model 2JHU in PHASER, but was not investigated further.

At this point all models obtained by molecular replacement reasonably matched the diffraction data (R_{work} values between 20 % and 30 %). Structure models were subsequently refined manually for a number of iterations using COOT (Emsley *et al.*, 2010) and REFMAC5 (Murshudov *et al.*, 2011) using restrained refinement with TLS refinement cycles, isotropic atomic B factors, and without specifying hydrogen atom locations. The final refinement statistics for the structures derived from the primary datasets are shown in Table 4.7.

To ensure unbiased R_{free} figures (Brünger, 1997), R_{free} sets were reused: for each RhoGDI

mutant the structure factors of the relevant PDB structure used in molecular replacement (Table 4.3; 2JHS for M1C1, M1C2, and M1C4) were obtained from the PDB. The R_{free} set was then extended to the maximum observed resolution of any radiation damage series involving that mutant, using the CCP4 programs CAD and uniqueify (Winn *et al.*, 2011). The resulting R_{free} set was then used for all datasets of that particular RhoGDI mutant.

Residues involved in crystal contacts were identified using the CCP4 program NCONT (Winn *et al.*, 2011). A crystal contact residue is defined as a residue having at least one atom within 4 Å of any non-hydrogen atom of a residue of another protein molecule. Results show that the mutated residues are involved in crystal contacts for the RhoGDI CH mutation (Figure 4.9, top), but not for the DY (see Figure 4.9, bottom) or EA mutations (see Figure 4.10). The EA mutant crystallises in two different space groups depending on the buffer composition ($P\,1\,2_1\,1$ for the PEG buffer and $R\,3\,2$ for the ammonium sulphate buffer conditions, see Table 4.6), leading to a different set of crystal contact residues. The full list of identified crystal contacts for all RhoGDI structures is given in Appendix D (page 223).

For each series, the subsequent datasets based on the identical data collection wedge (and thus: crystal volume) as the first dataset[6], were indexed and integrated using xia2 as before, but with the refined first dataset used as reference to speed up processing. Some datasets late in the radiation damage series of crystals DY3C5, M1C2, and M2C4 could not be processed automatically due to excessive global radiation damage (see Table 4.8 for an overview). In the interest of retaining optimum comparability between datasets it was decided not to pursue manual indexing and integration of these datasets.

[6] For crystal CH1B6H2 the subsequent datasets were only based on the 72° wedge used in the high-resolution pass of the primary dataset.

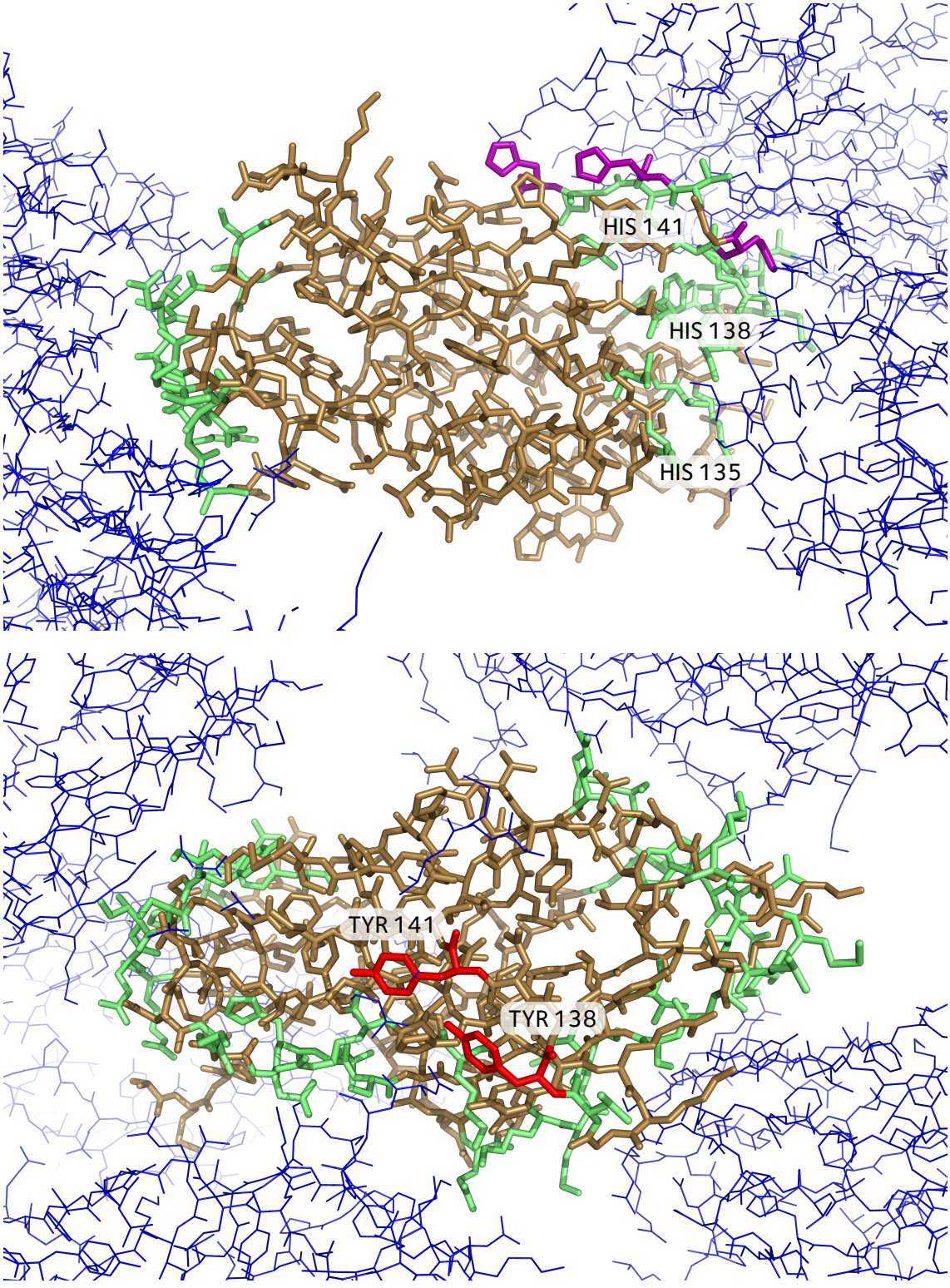


Figure 4.9: The structure models (brown) for the primary dataset of the radiation damage series CH1A1L1 (**top**) and DY2A4 (**bottom**) are shown alongside their neighbouring symmetry-related protein copies (blue). Residues within 4 Å of a symmetry related non-hydrogen atom are identified as crystal contacts, and marked green. The mutated residues HIS 135, 138, and 141 (Table 4.3) are all forming crystal contacts in CH1A1L1 and are shown in purple. In the DY2A4 crystal the mutated residues TYR 138 and 141 (both red) lie on the protein surface, but do not form crystal contacts.

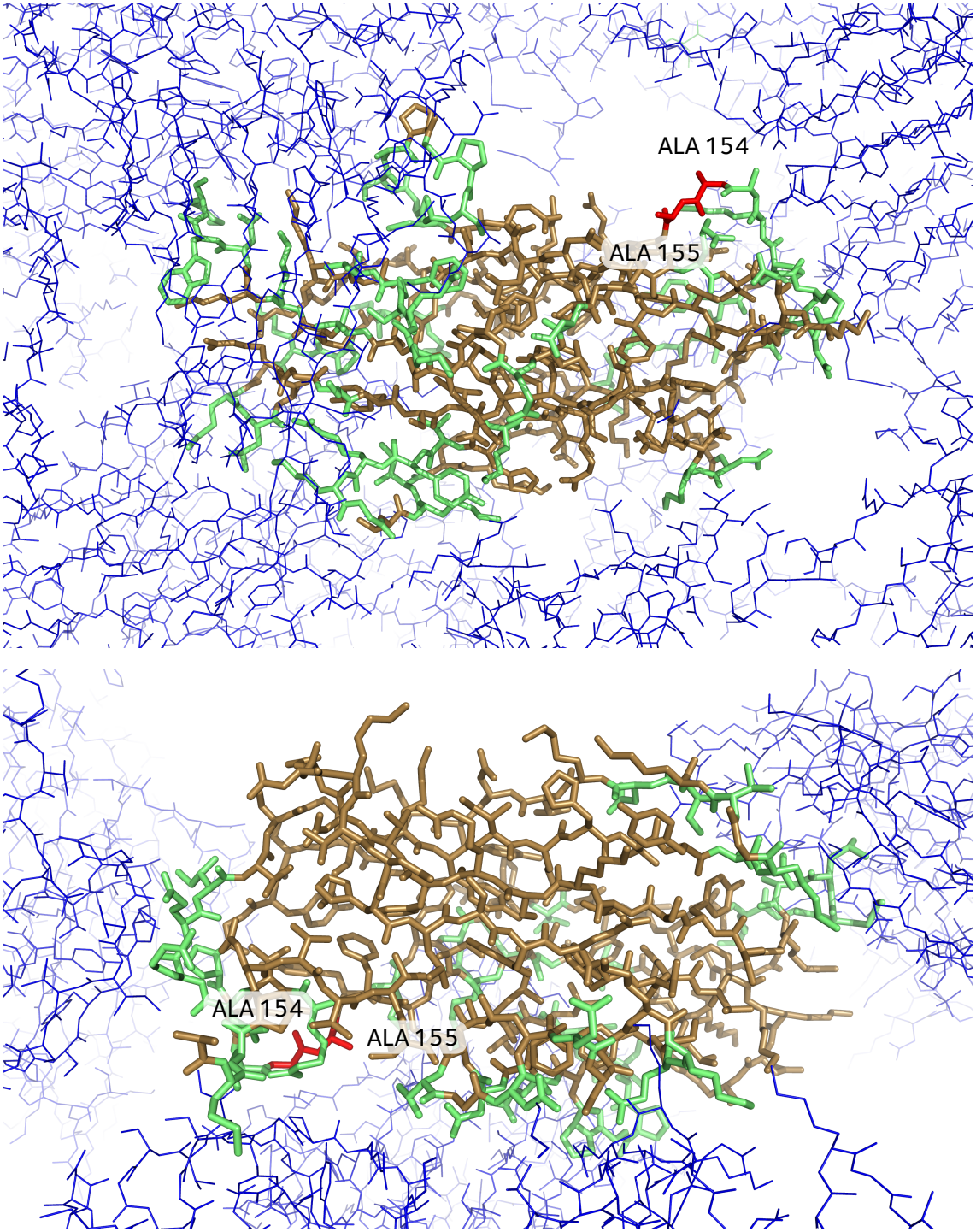


Figure 4.10: The chain A structure models (brown) for the primary dataset of the radiation damage series M1C1 (**top**) and M2C4 (**bottom**) are shown alongside the remaining chains and neighbouring symmetry-related protein copies (blue). Residues within 4 Å of a symmetry related non-hydrogen atom are identified as crystal contacts, and marked green. The protein in both crystals is the RhoGDI EA mutant, yet the crystal contacts are different due to the crystallisation in different space groups ($P\,1\,2_1\,1$ for M1C1 and $R\,3\,2$ for M2C4, Table 4.6). The mutated residues ALA 154 and 155 (shown in red; Table 4.3) do not form crystal contacts in either of the structures.

Series	Datasets									
CH1A1L1	2.6 MGy	18.9 MGy	35.1 MGy							
CH1A5H1	2.4 MGy	17.2 MGy	32.1 MGy	46.9 MGy	61.8 MGy	79.3 MGy	110.8 MGy			
CH1A6H1	3.0 MGy	20.7 MGy	38.5 MGy							
CH1B4H1	1.0 MGy	8.2 MGy	21.2 MGy	36.5 MGy	64.0 MGy	98.1 MGy	166.9 MGy	246.8 MGy		
CH1B6H1	2.1 MGy									
CH1B6H2	2.6 MGy†	7.6 MGy	21.7 MGy	35.7 MGy	49.8 MGy					
CH1B6H3	1.1 MGy	8.1 MGy	15.2 MGy	23.5 MGy	38.5 MGy	57.0 MGy	94.4 MGy	137.7 MGy	212.5 MGy	
CH1B6H4	1.3 MGy	9.9 MGy	18.5 MGy	28.5 MGy	46.7 MGy	69.0 MGy	114.6 MGy	167.0 MGy	258.2 MGy	
CH1D3H1	1.4 MGy	13.2 MGy	33.8 MGy	57.7 MGy	101.5 MGy	154.4 MGy	263.9 MGy			
CH1D3L1	2.5 MGy	18.2 MGy	33.9 MGy							
CH2A1H1	2.9 MGy	21.1 MGy	39.3 MGy	57.5 MGy	75.7 MGy	93.9 MGy	112.1 MGy	130.3 MGy	148.5 MGy	166.7 MGy
	184.9 MGy	254.7 MGy								
CH2A2H2	2.4 MGy	18.5 MGy	34.6 MGy							
DY2A4	1.6 MGy	7.9 MGy	14.1 MGy	20.4 MGy	26.7 MGy	32.9 MGy	39.2 MGy	45.5 MGy	51.7 MGy	58.0 MGy
	64.3 MGy	70.5 MGy	76.8 MGy	83.1 MGy	89.3 MGy	95.6 MGy	101.9 MGy	108.1 MGy		
DY3C1	2.3 MGy	11.3 MGy	20.4 MGy	29.5 MGy	38.5 MGy	47.6 MGy	56.6 MGy	65.7 MGy	74.7 MGy	83.8 MGy
	92.8 MGy	101.9 MGy	111.0 MGy	120.0 MGy	129.1 MGy	138.1 MGy	147.2 MGy	156.2 MGy		
DY3C5	1.8 MGy	9.0 MGy	16.2 MGy	23.3 MGy	30.5 MGy	37.7 MGy	44.9 MGy	52.0 MGy‡	59.2 MGy	66.4 MGy‡
	73.6 MGy‡	80.7 MGy‡	87.9 MGy‡	95.1 MGy‡						
M1C1	2.8 MGy	8.2 MGy	13.7 MGy	19.1 MGy	24.6 MGy	33.2 MGy	44.9 MGy	69.9 MGy		
M1C2	3.2 MGy	9.4 MGy	15.6 MGy	21.8 MGy	28.1 MGy	34.3 MGy	73.4 MGy‡			
M1C4	7.9 MGy	23.6 MGy	39.3 MGy	55.0 MGy	70.7 MGy	86.4 MGy				
M2C4	1.0 MGy	7.0 MGy	14.2 MGy	27.0 MGy	39.9 MGy	52.8 MGy	67.9 MGy	94.1 MGy‡	123.8 MGy‡	171.0 MGy‡
p18-T	20.8 MGy	143.3 MGy	265.8 MGy							
p18-B	7.3 MGy	20.3 MGy	33.2 MGy	46.2 MGy	59.2 MGy	72.2 MGy	85.1 MGy	98.1 MGy	111.1 MGy	124.1 MGy

Table 4.8: Diffraction Weighted Doses of the RhoGDI radiation damage datasets used in this survey. Subsequent datasets of each crystal use the same crystal volume and are of the same angular size as the primary dataset of the crystal, Table 4.6.

†Dose estimated from the second, high-resolution sweep. RADDOSE-3D does not currently support the calculation of a combined diffraction weighted dose of two sweeps with different beam settings. This is an overestimate, the real dose should be below this value.
‡Dataset excluded from study as it could not be successfully indexed/integrated with xia2.

4.3.3 Isomorphism

Xia2 (version 0.3.6.3, build 4411) determines the appropriate resolution cut-off by analysing merged and unmerged $I/\sigma(I)$ statistics (Winter *et al.*, 2013). Here, xia2 was run with its default settings, resulting in a high resolution cut-off at the point where either the merged $I/\sigma(I)$ falls below 2 or the unmerged $I/\sigma(I)$ falls below 1.

The derived high resolution limits for the initial and all subsequent datasets are shown alongside the R_{merge} value in Table 4.9. These show that for some crystals, such as CH1A1L1 and CH1B6H4, the resolution limit is essentially kept constant throughout the series. For other crystals, a steady decrease in the apparent resolution (increase in the resolution limit in Ångströms: i.e. a steady reduction in the number of used reflections) can be observed with a concomitant increase in R_{merge} values.

4.3.4 Estimating dose

The beam fluxes were measured at the beamline by placing a calibrated silicon pin diode (500 µm thickness, no aluminium filter; S3204-09; Hamamatsu Photonics, Hamamatsu City, Japan) at the crystal position and recording the X-ray induced current using a Keithley picoammeter (Table 4.5) as described in Owen *et al.* (2009). For the flux measurements, apertures and collimation settings were set to those later used in the diffraction experiment. Measurements were taken at different transmission settings to accurately determine the effect of different attenuators. Fluxes could not be measured for all experiments due to time constraints at the beamline and the necessary complications involved with the need to rearrange and interfere with beamline equipment optimised for high crystal throughput. In these cases, which are noted in Table 4.5, the flux measurements

Series	High resolution limit [Å] / R_{merge} [%] of datasets									
CH1A1L1	2.29/4.5	2.28/4.5	2.29/4.3							
CH1A5H1	2.28/4.6	2.33/4.8	2.41/4.9	2.49/5.1	2.52/5.3	2.47/5.0	2.59/5.3			
CH1A6H1	2.25/3.3	2.30/3.1	2.35/3.2							
CH1B4H1	2.66/9.9	2.60/7.8	2.64/7.7	2.53/6.0	2.52/6.3	2.38/4.6	2.41/4.6	2.33/3.7		
CH1B6H1	2.69/5.7									
CH1B6H2	1.98/11.0	2.05/3.2	2.12/2.9	2.17/3.1	2.23/3.2					
CH1B6H3	2.19/3.2	2.22/3.3	2.24/3.4	2.17/2.9	2.22/3.0	2.15/2.7	2.24/3.0	2.27/3.0	2.41/3.5	
CH1B6H4	2.48/5.5	2.49/5.4	2.49/5.7	2.41/5.0	2.41/5.1	2.29/4.1	2.36/4.1	2.33/3.7	2.44/3.8	
CH1D3H1	2.89/10.3	2.74/8.7	2.75/8.5	2.64/6.9	2.65/6.7	2.49/5.3	2.53/5.3			
CH1D3L1	2.30/4.7	2.29/4.7	2.31/4.7							
CH2A1H1	2.37/5.7	2.62/6.7	2.67/7.1	2.80/7.6	2.85/8.4	2.54/8.0	2.60/7.7	2.60/7.8	2.60/7.9	2.55/7.1
	2.56/7.0	2.42/5.7								
CH2A2H2	2.39/4.9	2.40/4.8	2.38/4.9							
DY2A4	1.99/5.0	2.02/4.8	2.05/4.6	2.10/4.7	2.17/4.9	2.27/4.8	2.34/5.1	2.47/5.3	2.57/5.6	2.68/6.3
	2.78/6.9	2.90/7.1	2.99/8.3	3.04/8.9	3.13/10.0	3.21/11.4	3.26/11.5	3.31/11.5		
DY3C1	1.75/2.7	1.81/2.9	1.85/3.0	1.90/3.2	1.95/3.3	2.02/3.6	2.07/3.8	2.15/4.0	2.19/4.1	2.29/4.1
	2.36/4.4	2.42/4.5	2.54/4.5	2.61/4.8	2.69/4.9	2.77/5.3	2.88/5.4	2.94/5.8		
DY3C5	2.18/6.4	2.25/6.4	2.36/7.0	2.45/7.3	2.47/7.8	2.56/8.2	2.65/8.5	—†	2.78/8.3	—†
	—†	—†	—†	—†						
M1C1	2.04/8.5	2.11/8.7	2.19/8.7	2.28/8.7	2.26/10.6	2.39/10.2	2.48/12.3	2.68/11.7		
M1C2	2.22/12.5	2.35/14.0	2.49/14.0	2.47/14.4	2.65/13.9	2.79/15.2	—†			
M1C4	1.99/8.7	2.30/9.6	2.58/12.2	2.80/15.1	3.00/16.8	3.19/19.5				
M2C4	2.03/6.9	2.17/7.4	2.33/7.2	2.46/8.2	2.66/8.5	2.80/8.8	2.85/8.4	—†	—†	—†
p18-T	2.49/7.2	3.26/12.2	4.88/12.4							
p18-B	2.56/8.4	2.71/9.3	2.84/9.2	2.99/9.8	3.15/10.5	3.18/15.2	3.47/13.0	3.53/16.3	3.38/20.3	3.41/24.1

Table 4.9: High resolution cut-off and R_{merge} progression for the RhoGDI radiation damage datasets used in this survey.

†Dataset excluded from study as it could not be successfully indexed/integrated with xia2.

provided by the beam line staff were used. For the damage series of crystals DY2A4, DY3C1, and DY3C5 flux estimates were obtained at the beamline via the Diamond GDA software during the experiment. These numbers later proved to be incorrect as they were based on (correct) in-line ion chamber readings but then were converted to flux readings using a conversion table from another beamline. Attempts to obtain correct figures after the fact, either through corrected conversion tables or through a flux measurement, proved unsuccessful. An approximate flux reading taken by beamline staff at the begin of the run was used instead. The error associated with these readings is unknown.

Diffraction-weighted doses (DWD; Zeldin *et al.*, 2013a) were calculated using RADDOSE-3D (Chapter 3, page 81; Zeldin *et al.*, 2013b). The calculated doses for each dataset used in this analysis are shown in Table 4.8. The spatial resolution used for the RADDOSE-3D crystal simulation was chosen in accordance with Zeldin *et al.* (2013b, §3) so that voxel elements are smaller than $^1/_{10}$th of the full-width half-maximum of the beam in its smaller dimension (see also Section 3.2.5, page 97).

The dose for the first dataset of crystal CH1B6H2 is an overestimate: The dataset is composed of two sweeps, one for low-resolution reflections and one for high-resolution reflections. RADDOSE-3D does not currently allow the calculation of DWD dose values for diffraction datasets composed of more than one sweep. For CH1B6H2 therefore the dose of its second, high-resolution sweep is taken as an estimate for the total DWD of the composite dataset.

Apart from the error on the flux readings the largest remaining uncertainty in the dose values is estimated to be due to the unknown beam profile. The stated dose values for the CH crystals assume a flat top-hat beam. Assuming a sensible Gaussian beam profile ($80 \times 20\,\mu m^2$ full-width

half-maximum) for e.g. crystal CH1B6H4 results in an increase of the DWD of 12.5 %.

As all damage series of CH- crystals were collected on the same day at the same beamline it is expected that any effect of the beam profile on dose will affect all CH-datasets to approximately the same degree. The same holds within all DY-, M- and p18-series. Comparisons across these series are expected to be affected by this error.

The obvious remedy for this unfortunate situation is to characterise beam profiles routinely at the beamlines, preferably for the particular data collection settings of the particular experiment. Given that, as described above, even the determination of beam fluxes still poses a problem at modern beamlines today, although technology such as ion chambers or even active beam stops with integrated silicon pin diodes (Ellis *et al.*, 2003) are widely available, another medium-term solution needs to be found. In hindsight, a statistical block design could have been the appropriate strategy for radiation damage series: by growing a number of crystals of each type and randomly assigning them to beam line experiments, such that each crystal type is used at least once at each beam line. The systematic effects of different beam profiles and errors on the beam characterisations could then be made visible.

4.3.5 Calculating the electron density

To calculate the electron density as described in Section 4.2.2 a refined coordinate and structure factor file is needed for each damaged dataset. A diagrammatic overview of the entire data processing pipeline is shown in Figure 4.11.

Theoretically it would suffice to combine the observed structure factor amplitudes from the damaged dataset with the phases of the reference dataset and pass these blended structure factors through an inverse Fourier transformation to obtain electron density maps. While this will

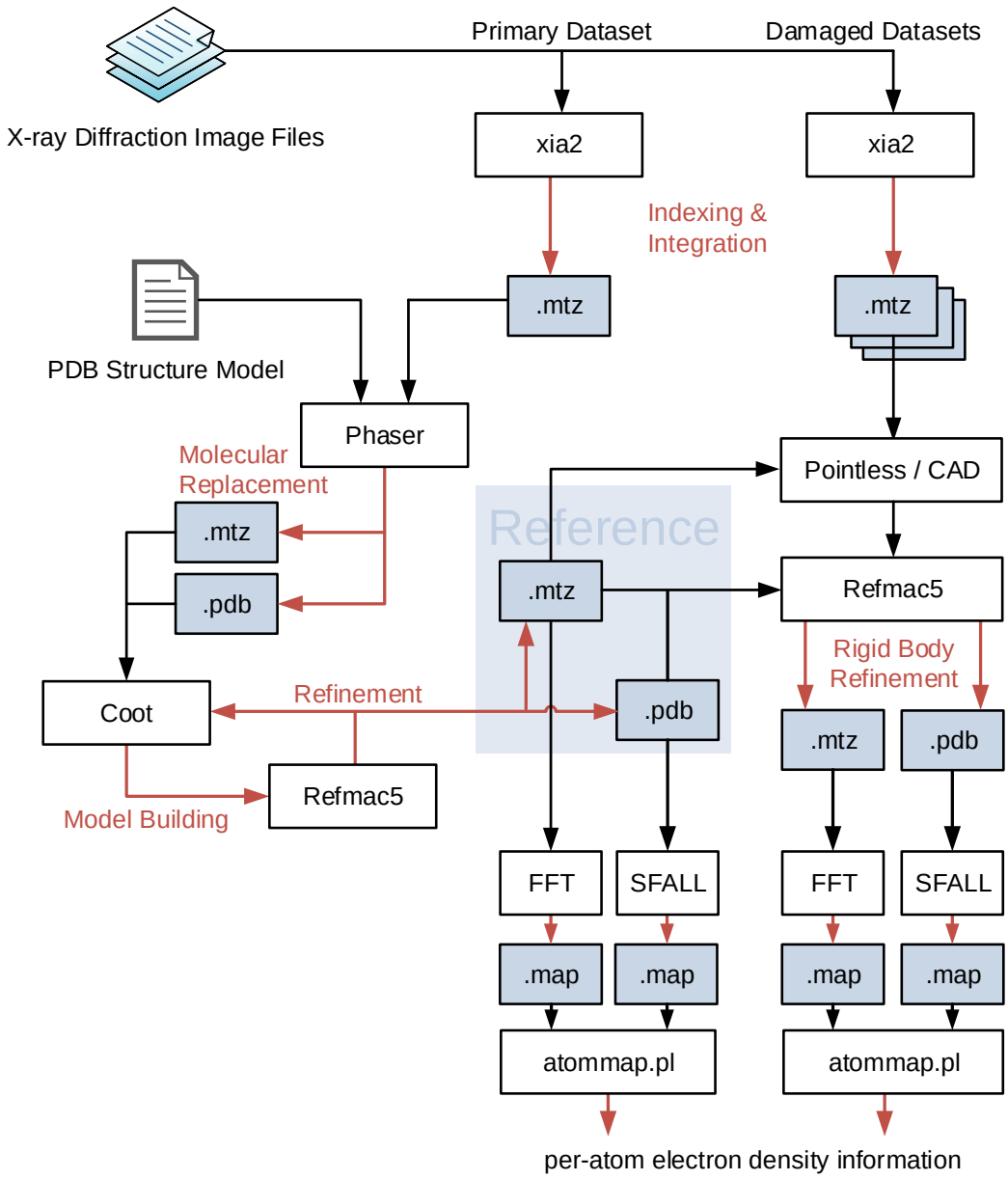


Figure 4.11: Overview of the data processing pipeline. Red arrows indicate processing steps using different programs (white boxes) resulting in either a structure factor file (.MTZ) or coordinates file (.PDB, both in blue boxes).
For each radiation damage series (Table 4.6) X-ray diffraction images (top left) are processed separately for each dataset. The first acquired dataset is solved with molecular replacement and then iteratively manually refined to obtain a low-dose reference structure factor and coordinates file (.MTZ and .PDB-boxes in the centre). Damaged datasets are prepared with POINTLESS and CAD to ensure they use the same indexing scheme and R_{free} set as the reference dataset. Rigid body refinement handles any protein translation or rotation due to radiation damage. Finally, the per-atom electron density is calculated for each atom in each dataset by combining the electron density map (via FFT) with an atom location map (via SFALL) as described in Section 4.2.2.

result in apparently useful electron density maps it fails to account for two sources of error: firstly the automatic data processing pipeline of xia2 does not in all cases[7] guarantee that the indexing of the damaged dataset is consistent with the indexing of the reference dataset. For example the space group P 3_2 2 1 has two alternative indexing schemes, which means that the (identical) unit cell can be defined in two different ways. If the indexing of the datasets is not consistent, then the amplitudes and phases will not match up. When phases and amplitudes from two radically different sources are combined, the result of the inverse Fourier transform will be dominated by the source of the phases (Taylor, 2010, Figure 3). In this case the result will always resemble an electron density similar to the reference structure model. This error can not be identified by visual inspection of this electron density model, but will result in R values above 50 % when the combined structure factors are used in any refinement process. A consistent indexing scheme can be forced by using the software POINTLESS (Evans, 2011) with the `HKLREF` parameter. At this stage the R_{free} assignment of the reference dataset can also be copied with CAD.

The simple combination of structure factor components also does not account for the second source of error: the movement of the protein due to genuine translation, rotation, or the apparent expansion of the unit cell. This may result in a systematic cumulative error on the results. By processing the structure factors of the damaged dataset together with the reference protein structure model in a rigid body refinement (using REFMAC5; Murshudov *et al.*, 2011) any movement is handled properly. In rigid body refinement the entire protein (including defined solvent atoms) is treated as a single unit and placed at the optimal position in the unit cell. Side chains are not moved, which is ideal for the electron density distribution analysis for the reasons outlined in

[7] Xia2 does offer a command line option to set a reference dataset. This setting may however be ignored if xia2 (version 0.3.6.3) decides to reindex the dataset at a later stage in the processing pipeline.

Series	Rotation	Translation						
CH1A1L1	0.00°	0.00 Å	0.00 Å	0.00 Å				
CH1A5H1	0.05°	0.01 Å	-0.00 Å	-0.01 Å				
CH1A6H1	0.00°	0.00 Å	-0.01 Å	-0.00 Å				
CH1B4H1	-0.01°	-0.00 Å	-0.01 Å	0.00 Å				
CH1B6H2	0.00°	0.00 Å	-0.02 Å	-0.01 Å				
CH1B6H3	-0.02°	0.02 Å	-0.04 Å	-0.02 Å				
CH1B6H4	0.02°	0.01 Å	0.01 Å	-0.01 Å				
CH1D3H1	0.00°	0.01 Å	0.01 Å	-0.00 Å				
CH1D3L1	0.00°	0.00 Å	0.00 Å	0.00 Å				
CH2A1H1	0.05°	-0.03 Å	-0.04 Å	0.01 Å				
CH2A2H2	0.00°	0.00 Å	-0.00 Å	-0.00 Å				
DY2A4	-0.16°	0.07 Å	-0.01 Å	-0.01 Å				
DY3C1	-0.12°	0.05 Å	0.01 Å	-0.03 Å				
DY3C5	-0.27°	0.03 Å	-0.00 Å	-0.04 Å				
M1C1	-0.01°	0.05 Å	-0.00 Å	-0.03 Å	0.00°	-0.01 Å	-0.01 Å	0.02 Å
	0.04°	0.01 Å	0.02 Å	0.02 Å	-0.01°	0.01 Å	-0.02 Å	-0.02 Å
M1C2	-0.07°	0.01 Å	0.00 Å	-0.02 Å	-0.02°	0.01 Å	0.01 Å	-0.02 Å
	-0.05°	0.00 Å	-0.02 Å	0.00 Å	0.13°	-0.05 Å	-0.01 Å	-0.04 Å
M1C4	0.06°	-0.02 Å	0.01 Å	-0.01 Å	-0.00°	0.07 Å	-0.00 Å	-0.04 Å
	0.52°	-0.05 Å	0.00 Å	-0.03 Å	0.06°	-0.01 Å	-0.02 Å	0.00 Å
M2C4	0.09°	0.09 Å	-0.09 Å	-0.00 Å	0.13°	-0.01 Å	-0.04 Å	0.04 Å
p18-T	0.40°	-0.06 Å	-0.08 Å	0.02 Å	0.40°	0.18 Å	-0.19 Å	-0.02 Å
p18-B	0.43°	-0.02 Å	0.02 Å	0.06 Å	0.03°	-0.11 Å	-0.04 Å	-0.01 Å

Table 4.10: Effective rotation angle (χ) and translation of the individual protein domains between the first and the last successfully processed dataset of each radiation damage series. Rigid body refinement statistics given by REFMAC5 (Murshudov *et al.*, 2011).

Section 4.1.3.

After the rigid body refinement the electron density maps, and thus the electron density distributions, can be obtained as described previously. The rigid body refinement statistics of the later datasets give a further indication of the degree non-isomorphism developing during the experiment, as shown in Table 4.10.

4.3.6 Observed electron density decay: a case study on DY2A4

Up to 11 GLU and 9 ASP residues were observable in the first structure of each radiation damage series. All residues were visually inspected in each primary dataset to ascertain how well defined their initial electron densities were. An overview of these results is shown in Table 4.11.

The remainder of this section will concentrate on one radiation damage series, DY2A4, and investigate the change in the electron density distributions specifically at the GLU and ASP residues. For this, the electron density distributions were calculated for each of those residues in each dataset, using the methods described earlier.

Additionally, an expected electron density distribution, T*, can be calculated from the reference structure models. It is based on the calculated structure factors F_{calc} of the structure model, and corresponds to a perfect protein crystal made up of model molecules. T* is based on the atom positions, occupancies, and atomic B factors in the model. It is therefore not the distribution of a perfect model of the protein, and not a zero-dose extrapolation of an undamaged structure model, but rather represents the expected ideal electron density distribution under the assumption that the structure model accurately describes the protein. As radiation damage processes cause a smearing out of the electron density or a loss of electron density at the location of interest due to side chain movement, T* is expected to generally contain higher electron density values compared to the observed electron density distributions.

The differences between T* and the observed electron density distributions can be attributed to three factors: an incorrect protein structure model (which would also be reflected in the model statistics, R_{work} and R_{free}), the fundamental inability of the protein structure model to accurately represent the true electron density distribution (Section 4.1.2), and radiation damage.

Series	GLU 87	95	106	109	121	154	155	157	163	164	193	ASP 89	93	140	143	183	184	185	188	201
CH1A1L1	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	◕	◕	●	○
CH1A5H1	◕	●	●	●	●	●	●	●	●	●	●	●	◕	●	●	●	◕	◕	●	◑
CH1A6H1	◕	◔	●	●	●	●	●	●	●	◕	●	●	◕	●	●	●	●	●	●	○
CH1B4H1	◑	◔	●	●	●	●	●	●	●	●	◕	●	●	●	◕	●	◕	◕	●	◔
CH1B6H2	◕	◑	●	●	●	●	●	●	●	●	●	●	◕	●	●	●	◕	●	●	○
CH1B6H3	◑	◑	●	●	●	●	●	●	●	●	●	●	◕	●	●	●	◕	●	●	◔
CH1B6H4	◑	◑	●	●	●	●	●	●	●	●	◕	●	◕	●	●	●	●	●	●	◔
CH1D3H1	◔	◕	●	●	●	●	◕	●	●	●	◕	●	◕	●	◕	●	◑	◑	●	◑
CH1D3L1	◑	●	●	●	●	●	●	●	●	●	●	◕	◕	●	●	●	●	●	●	◕
CH2A1H1	◑	◑	●	●	●	●	●	●	●	●	●	●	●	●	●	●	◕	◑	●	◔
CH2A2H2	◑	◑	●	●	●	●	●	●	●	●	●	●	◕	●	●	●	◕	◕	●	◔
DY2A4	◕	◑	●	◔	◔	◑	●	◕	◔	○	◕	◕	●	●	●	●	◕	◕	●	◑
DY3C1	◔	◕	◕	◕	◑	◔	●	◔	◑	◑	●	●	●	◕	●	●	●	◕	●	◕
DY3C5	◔	◔	◕	◔	◑	○	◕	◔	○	◔	◑	◕	◕	◕	●	●	◕	◔	●	◑
M1C1	◑◑○◕	◕◑◔●	●◕◔●	◕●◔○	●●◕◔	—	—	◕◕◕●	●◔◕◕	◕◕◑◔	●●●●	●●●●	●●◕●	●●●◕	●●●●	●●●●	●●●●	●●●●	●●●●	●×◕◔
M1C2	◔◔◕○	◔◔●●	◕●○◑	●◕◔◑	●●◔◔	—	—	◑◕◑●	○●◑◔	◔●○○	●●●◕	●●●◕	●●◕◕	●◔◑◔	●◕◕◑	●●●◕	◕◕◔◕	●●◕◑	●●●●	◔◕○◔
M1C4	○◔◔◔	○◕◕◕	●●●◑	◕◕◔◔	◕●○◔	—	—	◔◕◔●	○●◔◔	◔◕◑◑	●●●●	●●◔◕	●●◔◕	●●◔◑	◑◕◔●	●●●●	●◔◔●	●●◔●	●●●●	◑●××
M2C4	◕◑	◑◑	●●	◕●	●●	—	—	●●	◑◕	●●	●◑	●●	◕●	◔●	●●	●●	●●	●◕	●●	◕◑
p18-T	○◔	○◔	◔◕	◔◕	●◕	—	—	●◕	○○	◔◑	●○	●●	◕◔	◕●	◑◕	●●	◕◕	◔◑	●●	○○
p18-B	○○	○◔	◕◕	◕●	●●	—	—	●●	◔●	◕◕	●◑	●◕	◕○	●●	◔●	●●	●●	◑◕	●●	◔◔

Table 4.11: RhoGDI residue quality overview. This table indicates the electron density quality for all GLU and ASP residues in the first dataset (Table 4.6) of each radiation damage series except CH1B6H1.

Each circle represents the electron density at the centre of the carboxyl oxygen atom locations. The circle is filled (●) if the electron density exceeds 1.5 σ at both oxygens. The circle is filled three quarters (◕) when one of the two oxygens is between 1.0 σ and 1.5 σ. Further statuses indicate: both oxygens between 1.0 σ and 1.5 σ (◑), one oxygen atom above 1.0 σ and one oxygen atom below 1.0 σ (◔), and both oxygen atoms below 1.0 σ (○). When the residue is missing, or the electron density of its main chain falls below 1.0 σ, the residue is marked with an '×' instead. Residues replaced by alanine in the RhoGDI EA mutations are marked with '—'. Red markings (●◕◑◔○) indicate the presence of negative difference electron density above 3 σ at the carboxyl group.

The first two factors are dose-independent, the latter is not. A gap between T* and the electron density distribution of the first dataset is caused by the inability of the structure model to describe the dataset, on which it is modelled. Any subsequent change throughout the experiment can be expected to be caused by radiation damage.

Figure 4.12 shows the observed and T* electron density distributions for the carboxyl oxygens of the ASP 188 and GLU 155 residues in the radiation damage series DY2A4. Both residues have excellent electron densities in the reference dataset (Table 4.11), yet GLU 155 has a much smaller gap between T* and the first dataset than ASP 188. The median and the maximum of the electron density distributions, relative to either T* or the first dataset, may hold promise for a metric describing the susceptibility of the residue's carboxyl group.

The gap between the first electron density distribution and the T* distribution is closely related to F_{obs}-F_{calc} electron density difference maps: In these maps the differences between the observed and calculated electron densities are calculated. The maps are shown above a certain threshold (usually 3 σ based on the distribution of the entire density difference map) alongside the (absolute) electron density maps as an aid in model building. Rather than comparing the entire region of interest at once, the electron density is compared at each location.

The electron density and difference maps of ASP 188 and GLU 155 throughout the radiation damage series are shown in Figures 4.13 and 4.14.

In the case of ASP 188 negative electron density difference appears at the carboxyl carbon at 7.9 MGy and the electron density begins to visibly shrink from 20.4 MGy onwards. At GLU 155 negative electron density difference is only observed from 20.4 MGy, at which point the electron density of the carboxyl oxygens rapidly disappears. Between 70.5 MGy and 89.3 MGy the

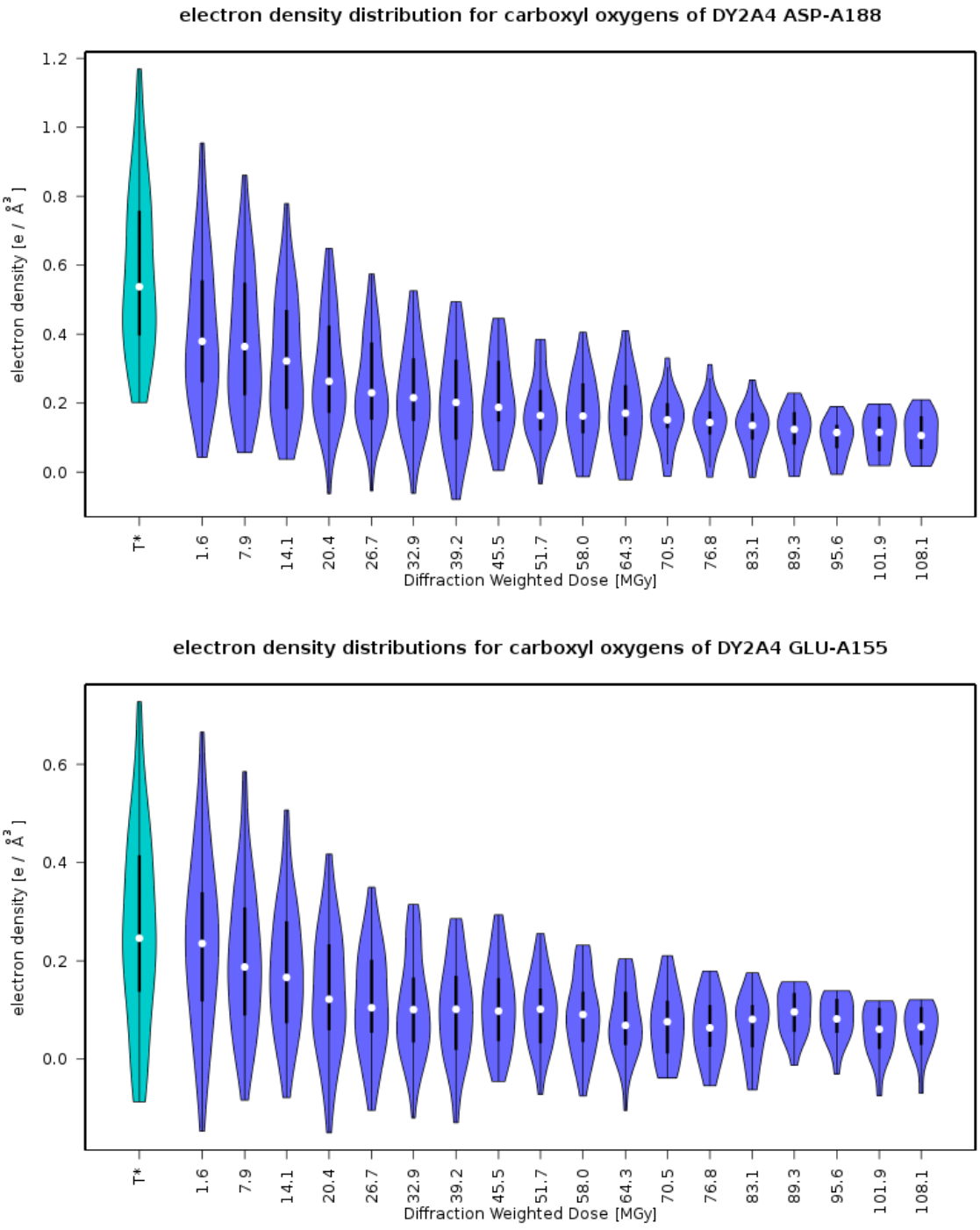


Figure 4.12: Violin plots (explained in Figure 4.4, page 142) of the electron density (blue) at the carboxyl oxygen positions of ASP 188 (**top**) and GLU 155 (**bottom**) of the DY2A4 radiation damage series against dose. A theoretical electron density distribution T* (cyan) was calculated from the protein structure model of the first dataset and the calculated structure factors (F_{calc}). The gap between T* and the distribution of the first dataset at 1.6 MGy is considerably larger for the ASP 188 residue than for the GLU 155 residue, indicating that the structure model does not describe the initial electron density as accurately.
The electron density decay over time is clearly visible in both cases.

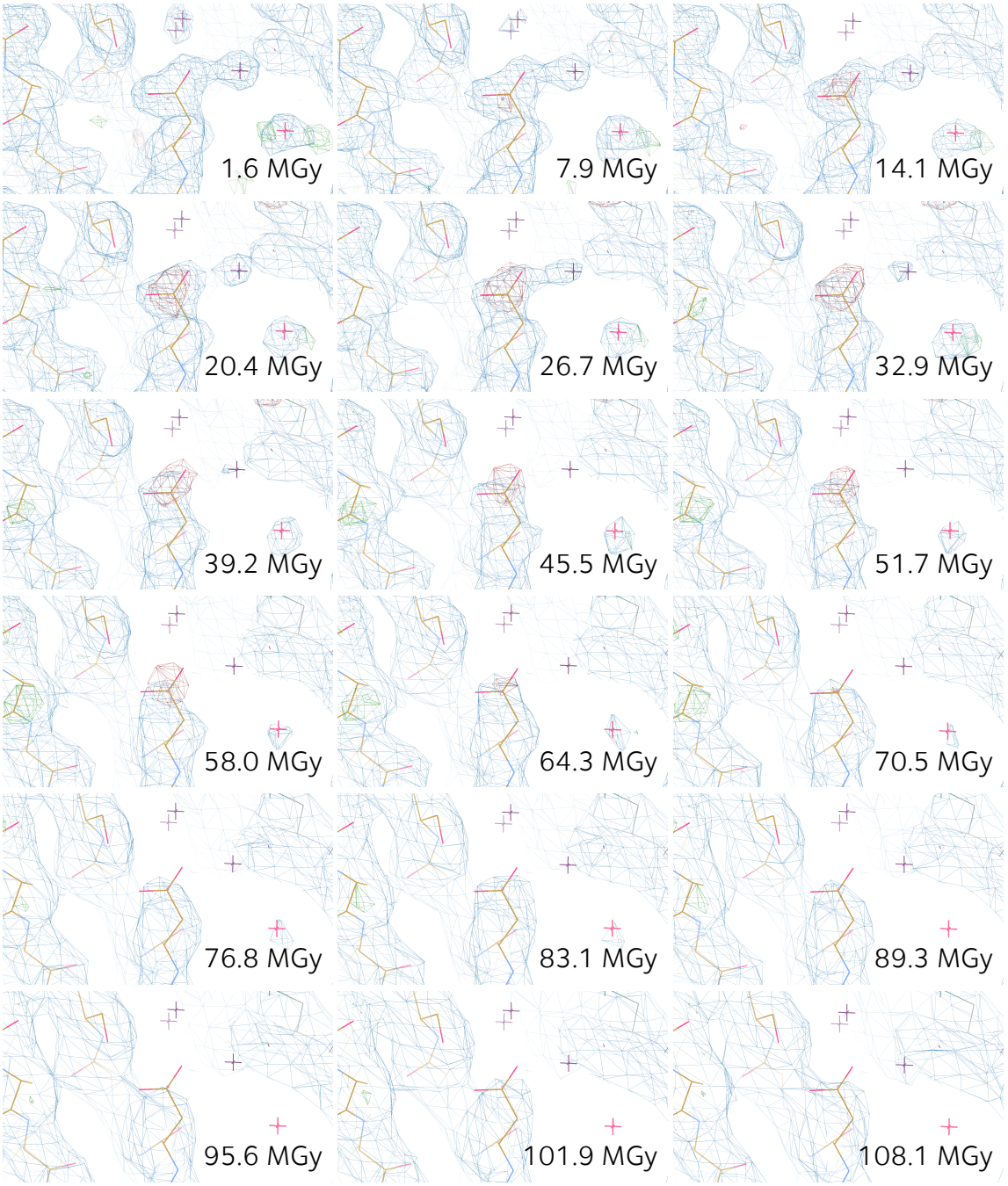


Figure 4.13: Electron density maps (blue) of the ASP 188 residue in the DY2A4 radiation damage series plotted at 1.0 σ level. F_{obs}-F_{calc} difference electron density (red negative, green positive difference) plotted at 3.0 σ.

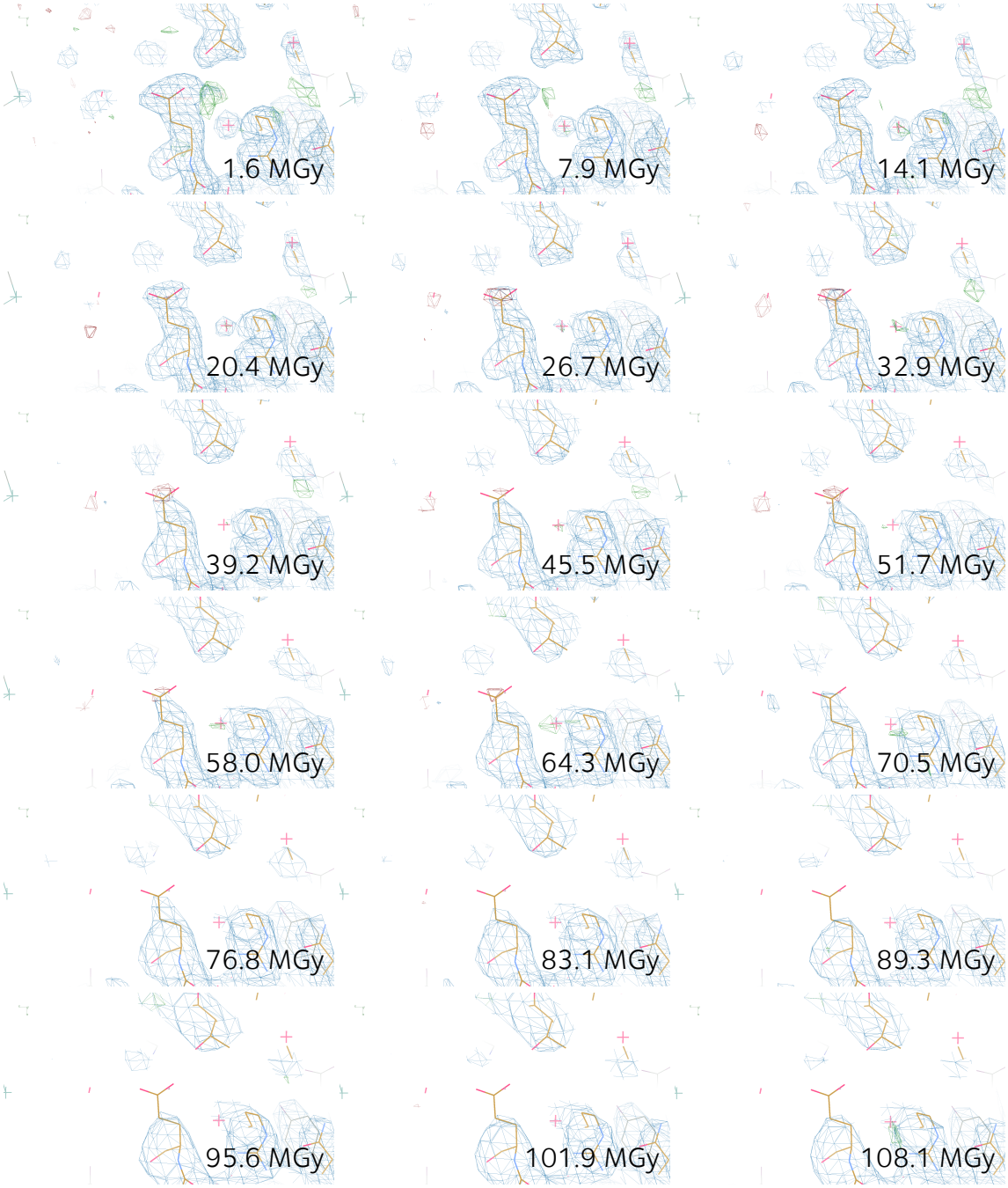


Figure 4.14: Electron density maps (blue) of the GLU 155 residue in the DY2A4 radiation damage series plotted at 1.0 σ level. F_{obs}-F_{calc} difference electron density (red negative, green positive difference) plotted at 3.0 σ.

electron densities of the C_δ and C_γ atoms are lost.

Interestingly, in both cases the negative electron density difference precedes the loss of observed electron density, but the negative difference itself disappears shortly thereafter. As difference densities are only shown above a $3\,\sigma$ cutoff, this disappearance could be caused by an overall increase in electron density difference throughout the model, leading to an increase of the absolute threshold. The difference density distributions within the regions of interest were obtained for both residues (Figure 4.15). At both residues an initial decline phase, a maximum negative difference (at 20.4 MGy for ASP 188 and 32.9 MGy for GLU 155) and a subsequent recovery of the absolute difference density can be observed. In the case of GLU 155 the median difference density even becomes positive in very late datasets.

The reasons for this difference density recovery are unclear. It is not caused by an increase in actual electron density, which would be visible in the electron density distribution plot (Figure 4.12). The atomic B factors of the protein model do not change in the rigid body refinement. By the process of elimination, this leaves only a final model-wide scaling step as an explanation: in order to obtain absolute electron densities in $^{\text{electrons}}/_{\text{Å}^3}$, the electron density map calculated by the Fourier transform is related to the expected electron density map from the structure model. A (global) scaling change should affect all residues simultaneously.

This therefore presents an opportunity for more hypothetical site specific radiation damage quantification metrics. If the electron density in a region of interest decays at a faster rate than the global decay of the whole protein, which is apparently negated in the global scaling step, the electron density difference minimum will be reached earlier. The dose associated with the smallest electron density difference median, and the magnitude of the minimum electron density difference

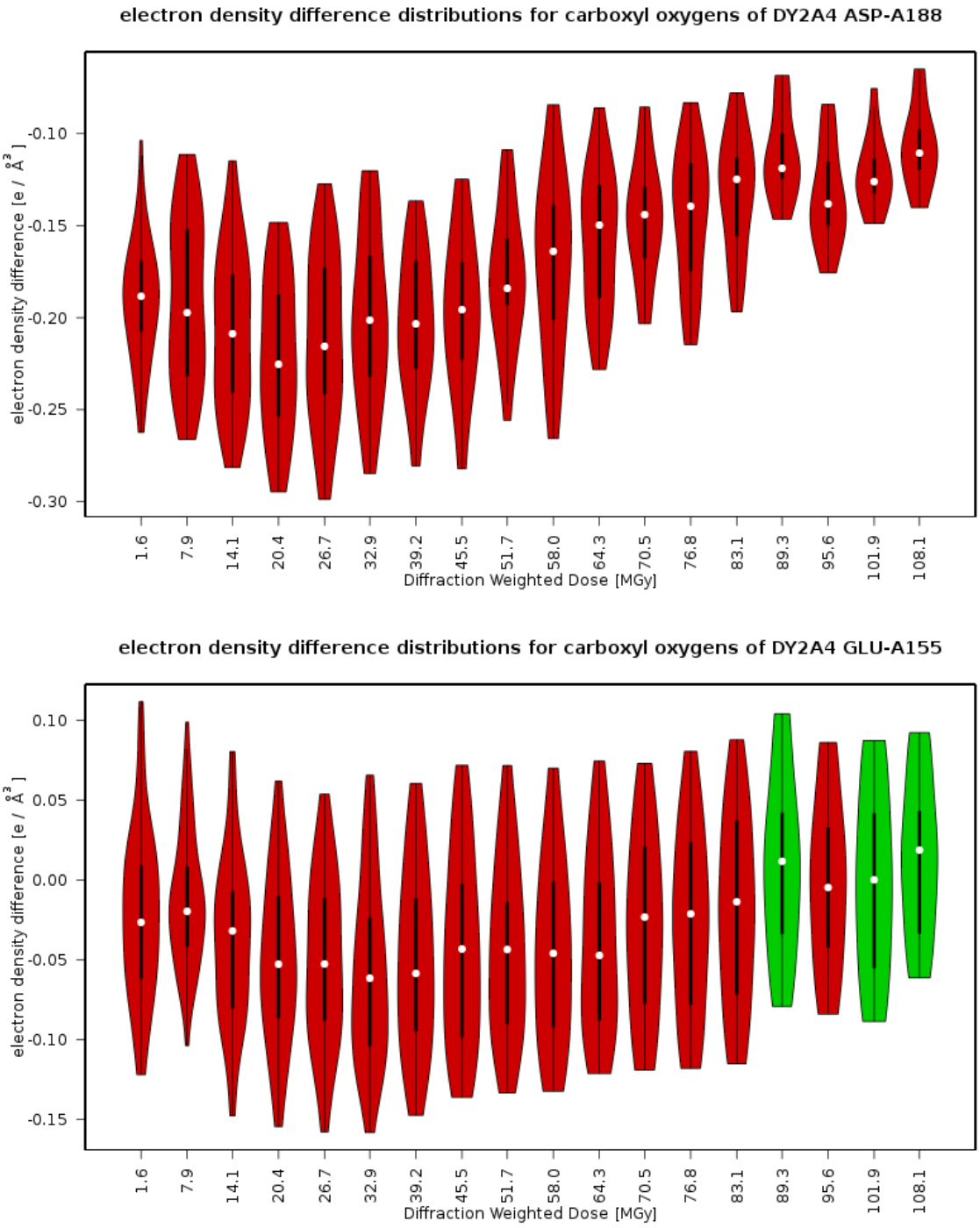


Figure 4.15: Violin plots (explained in Figure 4.4, page 142) of the F_{obs}-F_{calc} electron density difference (red for mainly negative, green for mainly positive) at the carboxyl oxygen positions of ASP 188 (**top**) and GLU 155 (**bottom**) of the DY2A4 radiation damage series against dose.

At both residues a negative electron density difference can be observed. In later datasets the trend reverses, and the negative electron density difference disappears.

median (in $^{electrons}/Å^3$) could both be suitable indicators to describe a residue's susceptibility to radiation damage.

4.3.7 Electron density and B_{Damage}

To further characterize the developing specific damage on the GLU 155 and ASP 188 carboxyl groups, the metric B_{Damage} (previously described in Section 2.3, page 37) was calculated. To this end, the structure model of the refined, first, low-dose dataset of each series was combined with the observed structure factor amplitudes in a restrained refinement cycle using REFMAC5, resulting in a structure model for each dataset in the radiation damage series. B_{Damage} was then calculated using atomic contact numbers with a 14 Å radius, grouped by fixed bins of size 10 (Section 2.3).

Again, only considering the carboxyl oxygens, B_{Damage} indicates that throughout the DY2A4 radiation damage series the ASP 188 carboxyl group is more damaged than the carboxyl group of GLU 155, Figure 4.16. For doses up to 32.9 MGy, this is in good agreement with the earlier observations of the specific decay of these two residues based on electron density (Figures 4.13 and 4.14).

At higher doses the observations apparently contradict each other. Up to 83.1 MGy B_{Damage} identifies GLU 155 as the more stable residue (Figure 4.16), yet there is clearly more residual electron density visible at ASP 188 compared to the completely decarboxylated GLU 155 (Figures 4.13 and 4.14). This can be attributed to the underlying protein structure model: B_{Damage} is based on a restrained refinement model, which allows some side chain movement, and the electron density maps used in the B_{Damage} calculation (Figure 4.17) do show a higher level of specific radiation damage on the ASP 188 carboxyl group at 64.3 MGy. This could indicate that the dam-

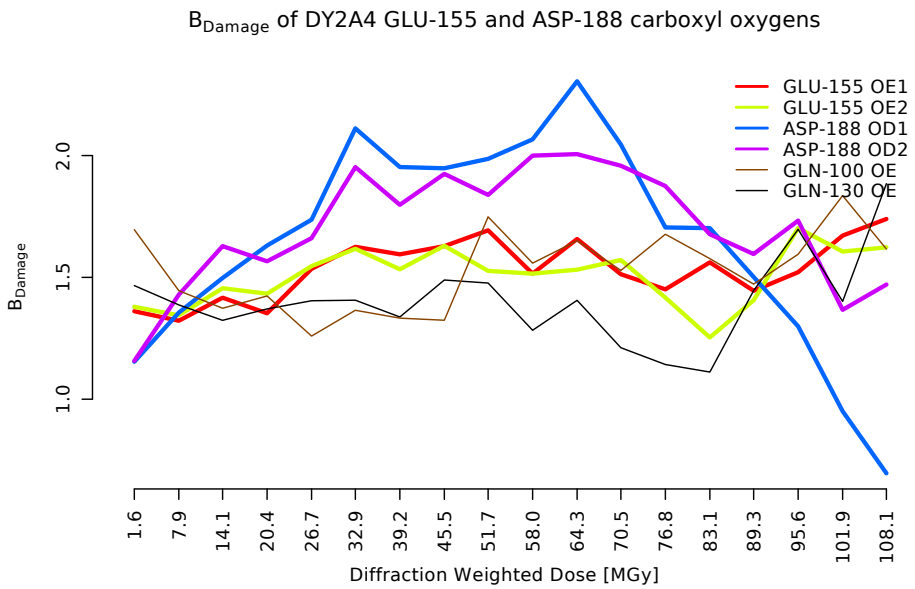


Figure 4.16: B_{Damage} on DY2A4 GLU 155 and ASP 188 carboxyl oxygens, indicating that, throughout most of the radiation damage series, the ASP 188 carboxyl group is more damaged than the carboxyl group of GLU 155. The B_{Damage} values of the oxygens of both glutamine residues in the protein, GLN 100 and GLN 130, only change for very high datasets.

age identified by the electron density decay method can be attributed, at least to some degree, to radiation damage induced movement of the residue side chains.

4.3.8 Quantifying site-specific radiation damage based on electron density

Previously, six possible candidates for metrics quantifying the susceptibility of a carboxyl group were identified: The median and the maximum of the electron density distributions, relative to either T* or the first dataset, and the lowest observed electron density difference median and its associated dose.

Figure 4.18 shows the electron density medians in the DY2A4 radiation damage series relative to the T* median and the median of the first dataset. The lines for the GLU 155 and ASP 188 residues are marked in bold.

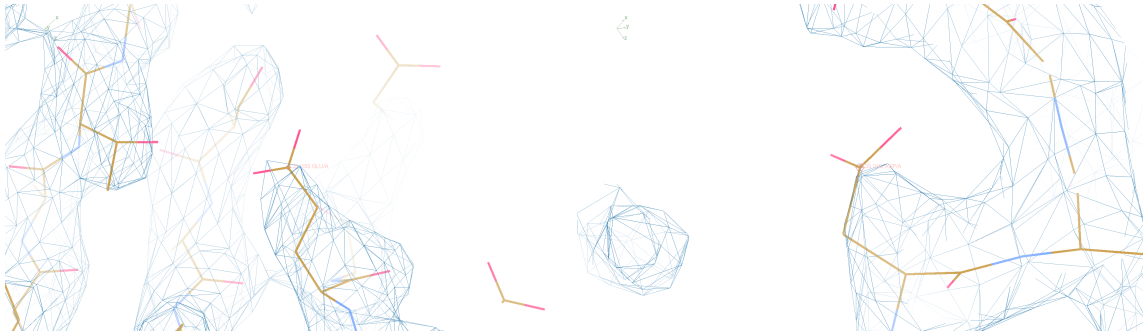


Figure 4.17: Electron density maps (blue) of the GLU 155 (**left**) and ASP 188 (**right**) residue in the DY2A4 radiation damage series at 64.3 MGy. The calculation of B_{Damage} is based on these maps, which were obtained by restrained refinement. The previously shown electron density maps of the same residues (Figures 4.13 and 4.14) are based on a rigid body refinement, which prevents the movement of side chains in the model. Electron density plotted at 1.0 σ level, F_{obs}-F_{calc} difference electron density (red negative, green positive difference) plotted at 3.0 σ.

Describing the residue susceptibility by a reduction in the median of the associated electron density distributions has a number of associated problems. The median contains a considerable noise component, making the identification of any patterns difficult. The median does not decrease monotonically over time, and, for some residues, does not decrease particularly strongly at all. Specifically, the relative median will not necessarily ever reach 50 % for GLU and ASP residues, unlike the strong peaks associated with cysteine sulphurs, as shown previously in Section 4.2 and Figure 4.6 (page 147). This is caused by the electron density of carboxyl oxygens being much closer to the average crystal background than the strong sulphur peaks.

The maximum appears to be the more useful basis for a metric, Figure 4.19. In theory the median should be more stable, as it can not be dominated by single outliers. Due to the nature of the electron density distribution (strong skewness, with a few high values describing the electron density peaks) in practice the maximum is more stable.

Both relative maximum density plots appear to mainly follow an exponential decay, decaying quickly over the first few datasets, and then slowly throughout the remaining series.

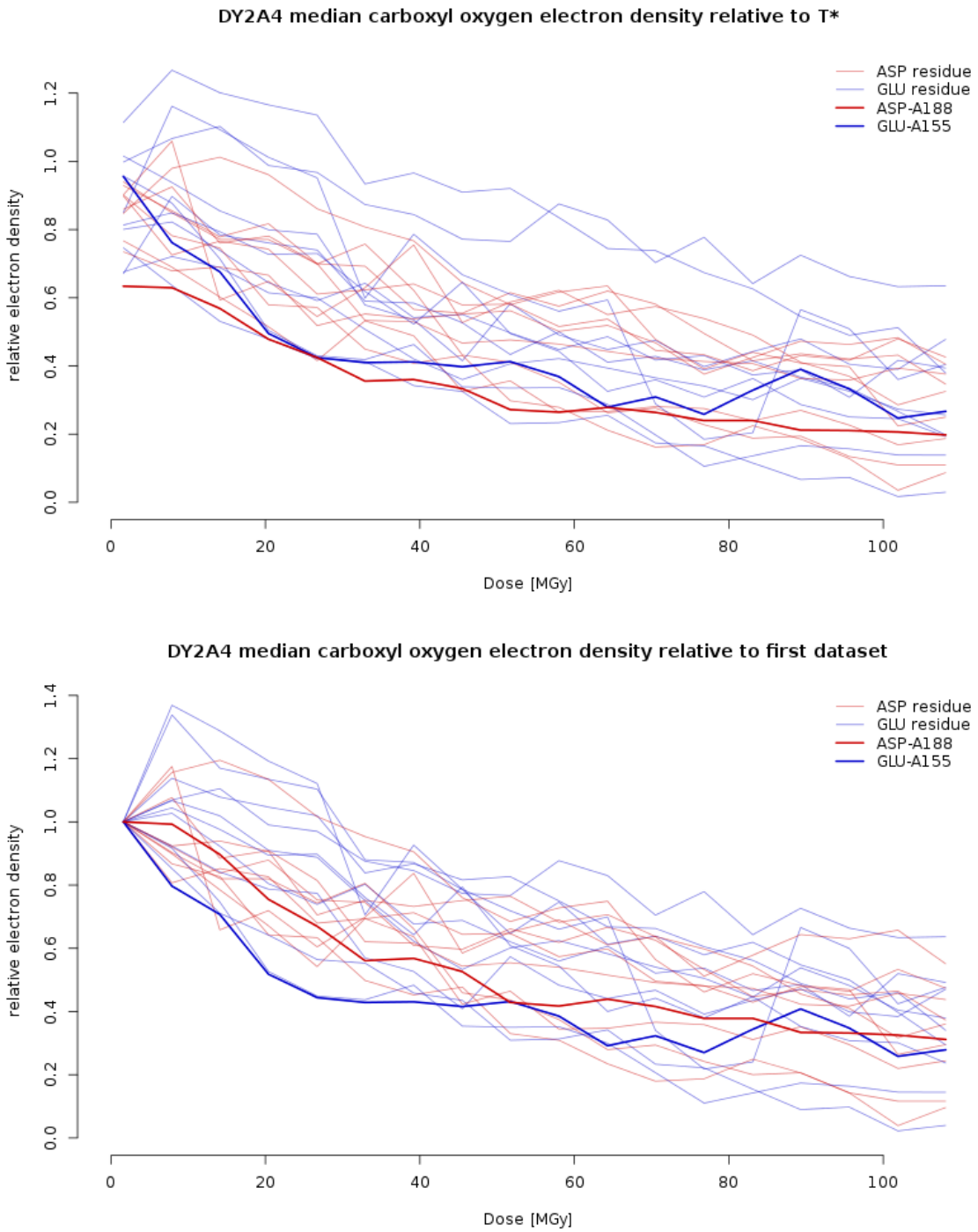


Figure 4.18: Median electron density in the region of interest of carboxyl oxygens of GLU and ASP residues in the DY2A4 radiation damage series, relative to the theoretical electron density median T* (**top**) and the electron density median observed in the first dataset (**bottom**). The lines representing the GLU 155 and ASP 188 residues are marked in bold. The median electron density has a strong noise component for some residues, and does not decay clearly for all residues.

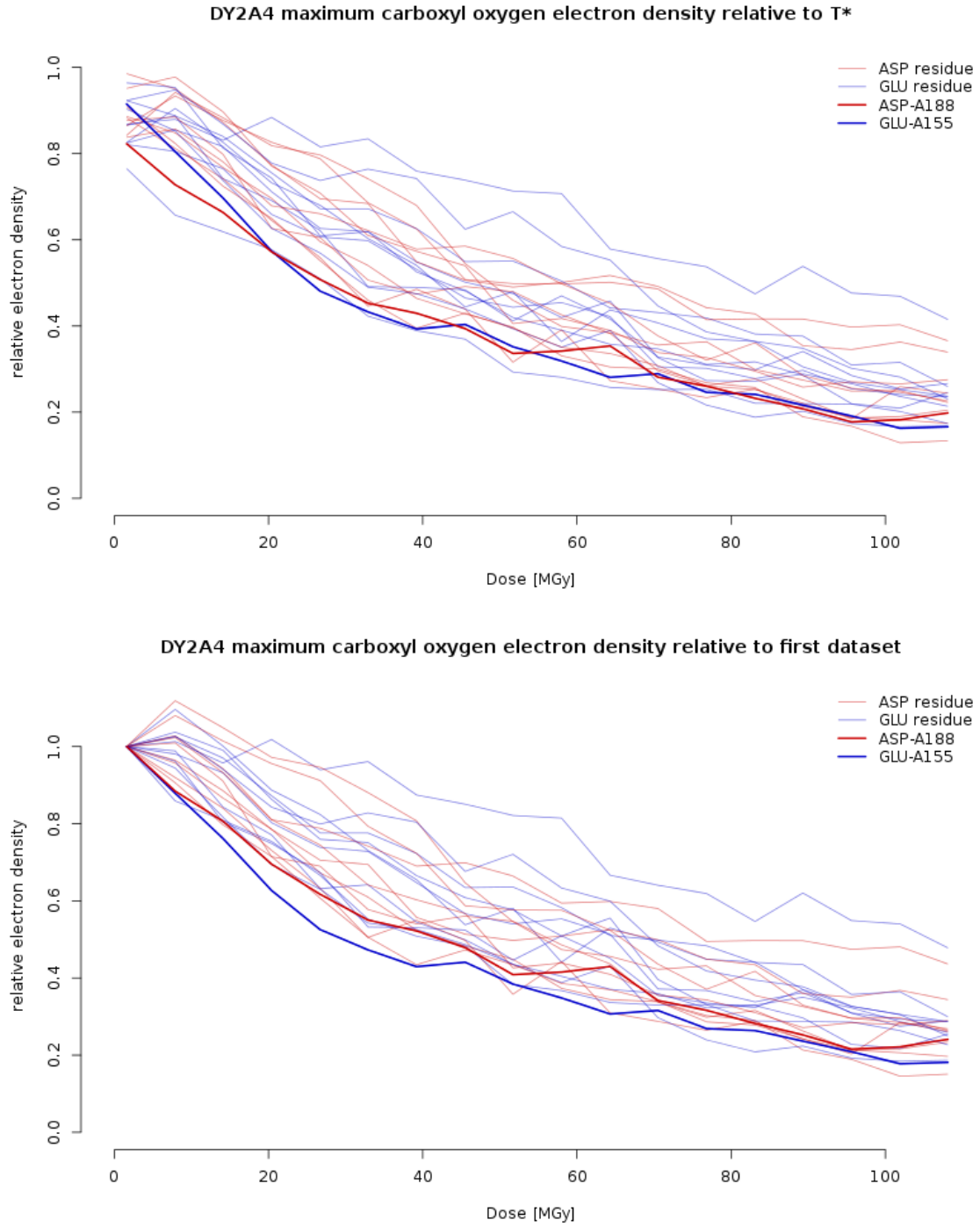


Figure 4.19: Maximum electron density in the region of interest of carboxyl oxygens of GLU and ASP residues in the DY2A4 radiation damage series, relative to the theoretical electron density maximum T* (**top**) and the electron density maximum observed in the first dataset (**bottom**). The lines representing the GLU 155 and ASP 188 residues are marked in bold. The maximum electron density has a weaker noise component and does show a more monotonic behaviour compared to the median (Figure 4.18). It may thus be more suitable to describe the residue susceptibility.

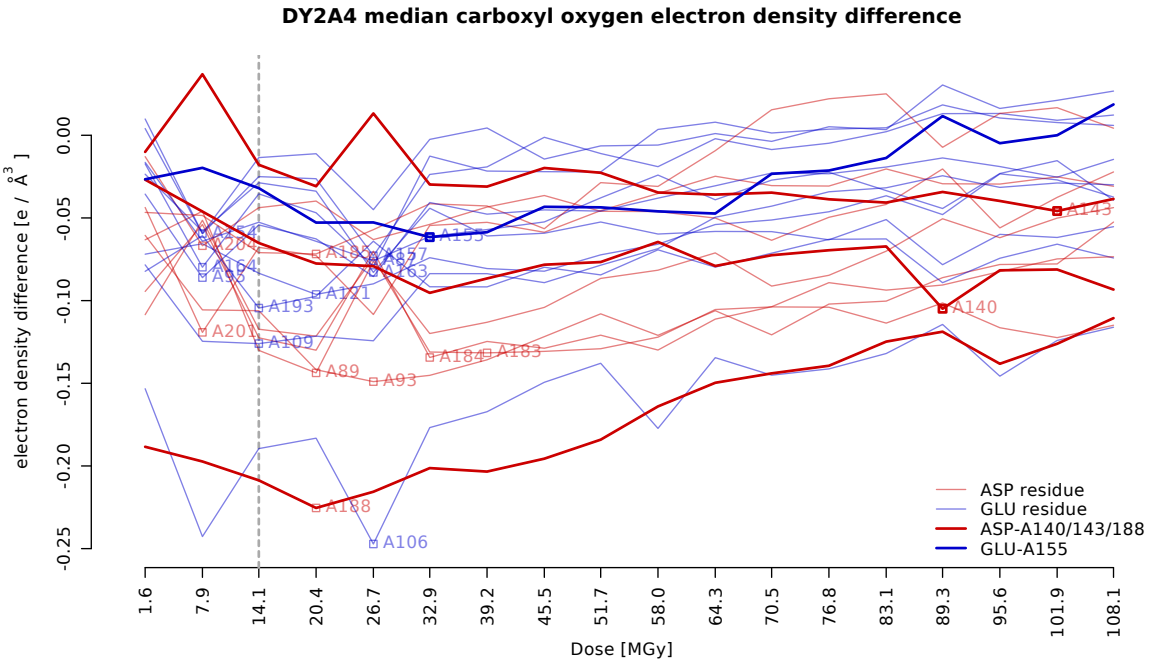


Figure 4.20: Median electron density difference in the region of interest of carboxyl oxygens of GLU and ASP residues in the DY2A4 radiation damage series on an absolute scale. The lines representing the GLU 155 and ASP 188 residues are marked in bold. The lowest median is marked for each residue. Some residues, such as GLU 106 and ASP 188 reach a lower minimum than others. With the exception of ASP 140 and ASP 143 all residues reach their minimum at doses below 40 MGy.
The electron densities of all GLU and ASP residues at 14.1 MGy (grey dashed line) are shown in the next figure, Figure 4.21.

It has been demonstrated earlier that the median electron density difference of ASP 188 reaches a lower minimum at a lower dose compared to GLU 155 (Figure 4.15, page 177). When the median electron density difference is plotted for all GLU and ASP residues, shown in Figure 4.20, it appears that most of the residues, including ASP 188, reach their minimum at doses near or below 20 MGy. A few residues have their minimum at higher doses, particularly high doses are for ASP 140 and ASP 143.

To determine whether there is a physical meaning attached to these determined minima it is useful to look at the electron densities of the residues at a dose where some residues have already reached their minimum, while others have not yet reached theirs. One such point is the

third dataset with an associated Diffraction Weighted Dose of 14.1 MGy, which is marked with a vertical dashed grey line in Figure 4.20. The electron densities of all residues at this dose are shown in Figure 4.21. Although ASP 140 and ASP 143 have particularly high associated doses for their minima, some radiation damage is clearly already visible at ASP 140 on the carboxyl carbon and one carboxyl oxygen, while ASP 143 appears to be undamaged at 14.1 MGy. Inspection of the ASP 140 residue in the 15th dataset, at its electron density difference minimum (89.3 MGy, not shown), reveals that no electron density remains at the entire carboxyl group at a 1 σ level, and the negative electron density difference, although large in absolute terms, has fallen below 2.2 σ.

These results indicate that a metric based on a median electron density, either relative to T* or to the first dataset, will show a larger variation or noise component than one based on an electron density maximum. Both maximum electron densities may be suitable candidates for a future metric. To design such a metric a parametrised model for the maximum electron density decay is required. Such a model does not currently exist.

The median electron density difference on the other hand could form the basis of a parameter-free model due to it having a minimum for each residue. It however also contains a sizeable noise component, which may limit its potential use. The effects of experimental noise could be reduced by relying on a larger number of observations. In the following section the median electron density difference is applied to the full range of collected datasets from all RhoGDI crystals in an attempt to identify effects of crystal contacts on residue stability.

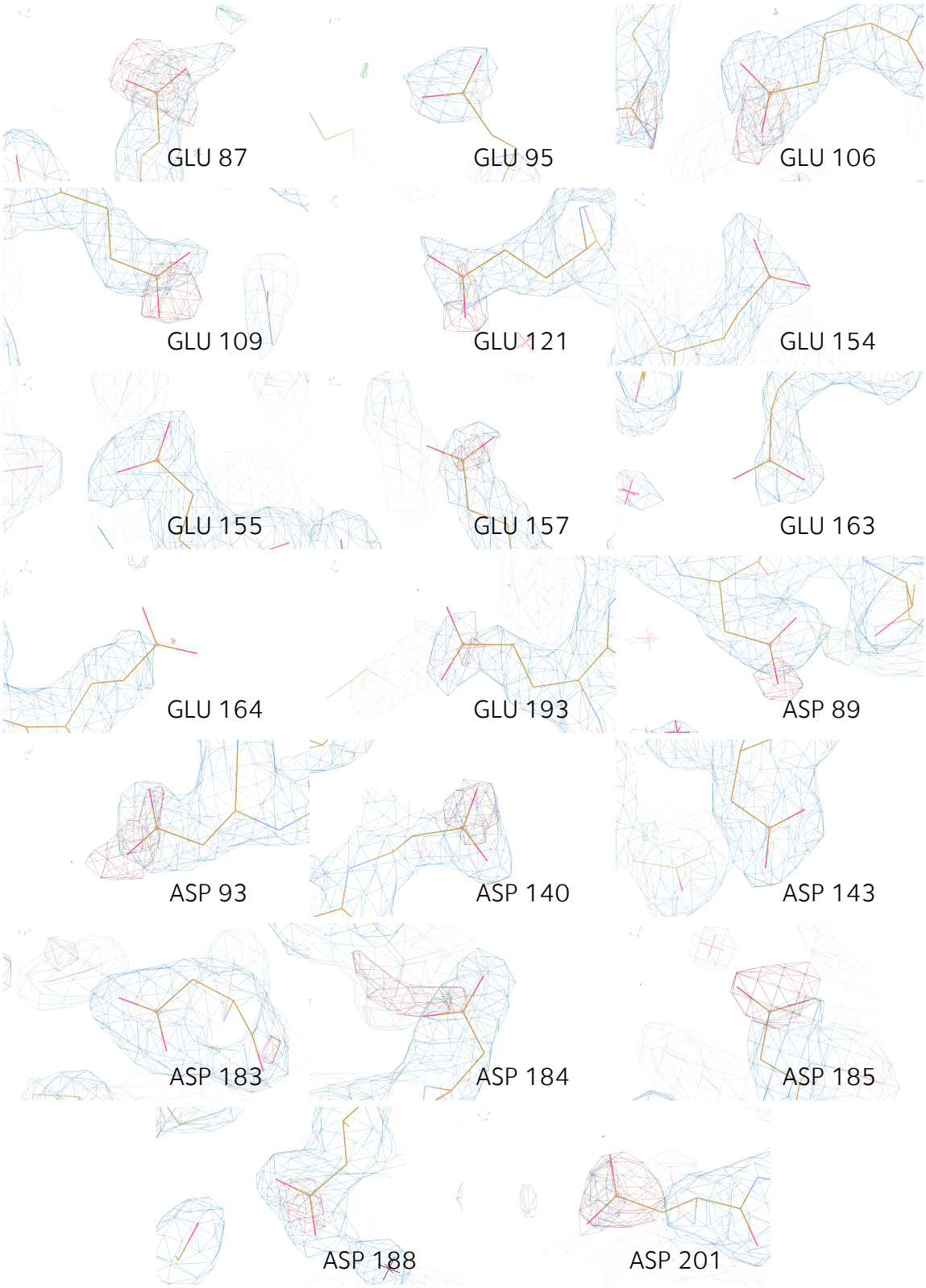


Figure 4.21: Electron density maps (blue, 1.0 σ level) of the 20 GLU and ASP residues in the DY2A4 series at 14.1 MGy. F_{obs}-F_{calc} difference electron density (red negative, green positive difference) plotted at 3.0 σ.
ASP 140, one of the two residues that are far from the dose at which they would reach the maximum negative electron density difference (Figure 4.20), shows clear signs of radiation damage on the carboxyl carbon and one carboxyl oxygen. ASP 143 shows no signs of radiation damage.

4.3.9 Effects of crystal contacts on residue stability

The dose at which the median electron density difference reaches its minimum (Minimum Density Difference Dose; MDDD) is a non-parametric statistic. Unlike the other electron density decay metrics which have been presented, the MDDD can be calculated without fitting a function or choosing a parameter.

The MDDD was calculated for all GLU and ASP carboxyl oxygens over all radiation damage series. The residue-specific doses were grouped together by the RhoGDI mutant and its crystallisation space group, which, as the RhoGDI EA mutant crystallises in two different forms with different crystal contacts (Figure 4.10), results in a total of four groups.

The resultant plot (see Figure 4.22), shows that the interquartile ranges of the MDDD distributions for the four different crystal groups overlap for most GLU and ASP residues. This indicates that, according to the MDDD, there are no significant differences between the different mutants in general.

The MDDD of ASP 140 of the EA mutants crystallising in the R $3\,2$ space group appear to be consistently higher than for the EA mutants which crystallise in the P $1\,2_1\,1$ space group. ASP 140 forms a crystal contact for the RhoGDI EA protein in its R $3\,2$ space group arrangement, but only in a few cases in the P $1\,2_1\,1$ arrangement[8]. This seems to indicate that the ASP 140 residue appears to be more stable when it is involved in a crystal contact. However this is only observed for one residue, and only between these two groups. Other similarly large differences in the distributions (e.g. ASP 183 between the EA mutant groups, or GLU 95 between the EA and CH mutants) can not be explained by the presence or absence of crystal contacts.

[8] 83 % of the protein chains in the R $3\,2$ space group, compared to 25 % of the protein chains in the P $1\,2_1\,1$ space group. The full list of identified crystal contacts can be found in Appendix D (page 223).

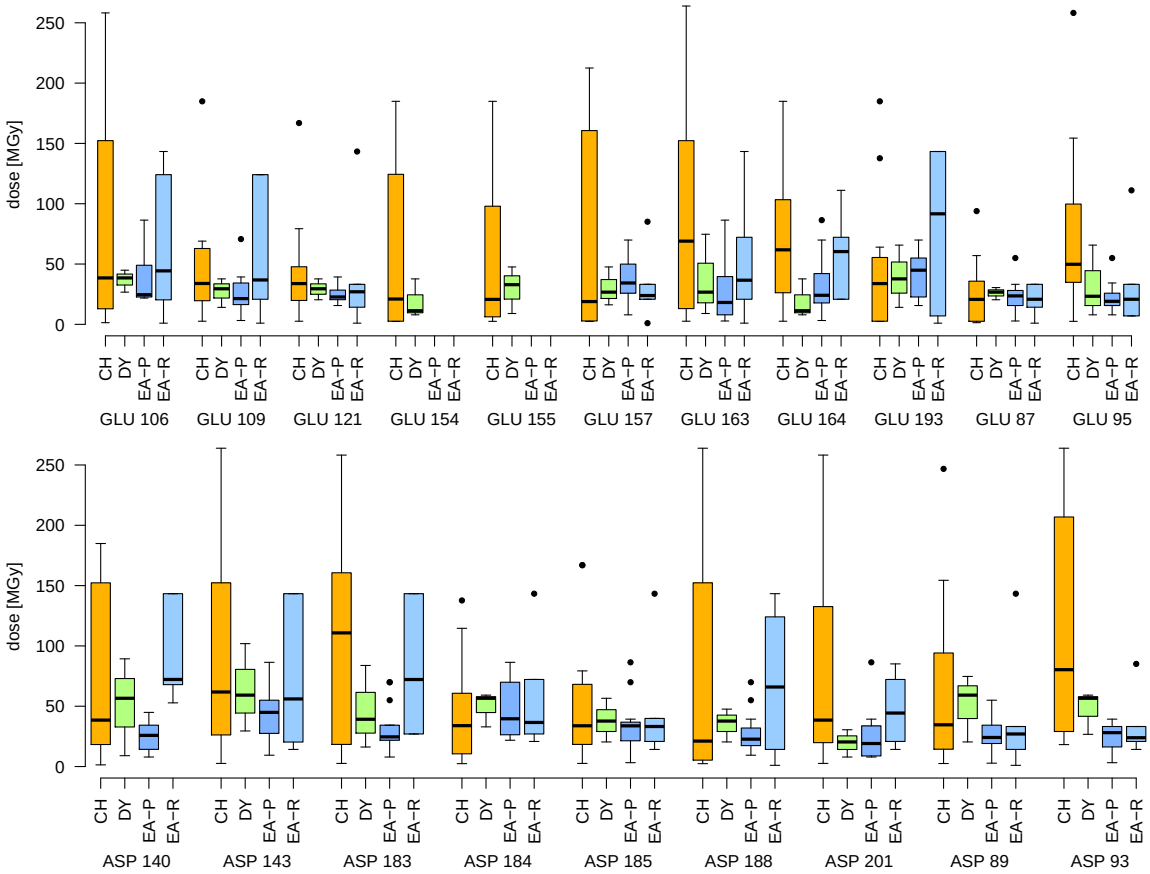


Figure 4.22: For each oxygen of each GLU and ASP residue for all RhoGDI radiation damage series the dose is identified at which the median F_{obs}–F_{calc} electron density difference reaches its minimum. A lower Minimum Density Difference Dose (MDDD) suggests a more susceptible residue. The doses are grouped together for the RhoGDI CH and DY mutants, and, separately, the EA mutants crystallising in the P $1\,2_1\,1$ and the R $3\,2$ space group.

The interquartile ranges (coloured boxes) overlap for most residues, which indicates that there are no significant differences between the groups. The difference between the two EA mutant forms for ASP 140 is discussed in the main text.

4.3.10 Is electron density based damage quantification feasible?

So far, the suitability of real space electron density distributions to quantitatively measure specific radiation damage for atoms lighter than sulphur appears to be inconclusive. To a large extent this is due to the lack of an explaining model. The median and especially the maximum electron density distributions look promising, but it is unclear how exactly they could be used to convert the observable electron density disappearance into useful radiation damage statistics. While the median electron density difference could be used directly without such a model, its strong noise component will limit its applicability.

The method could be improved further independently of an electron density decay model. In the work shown so far, both carboxyl oxygens are treated as one unit. However, as is clear from Figure 4.21 the decay in both glutamic and aspartic acids is not necessarily symmetric. In fact, seven of the 20 residues (GLU 106, GLU 121, GLU 157, ASP 89, ASP 140, ASP 184, and ASP 188) show a definite asymmetry in the position of the negative electron density difference. A clearer signal may thus be obtained if the oxygen atoms, and the carboxyl carbon, are considered separately.

It may be feasible to eventually base a metric on the real space electron density decay. A simple non-parametric statistic, such as the presented and discussed MDDD, is unlikely to suffice, and a more thorough understanding of the electron density decay process may be required. The required theoretical underpinning for a reliable real space electron density decay metric might be provided by a model that can explain the observed disappearance of the electron density difference at very high doses. This disappearance may foreshadow yet another limit for data collection, following the Henderson and Garman limits (Section 3.1, page 82), an indicator that once the difference is

disappearing at multiple sites the data should not be trusted or used in reduction.

In the absence of a rigorous model, it could also prove useful to either consider the σ-level of the median electron density differences, or a weighting scheme that discounts observations at higher doses.

4.4 Time-resolved specific damage observation

Any refined automated damage quantification metric analysing datasets is fundamentally limited by the dose-'distance' between the datasets. A very obvious example of this was shown in Figure 4.20: here, a minimum can only be found at the dose of an existing dataset, and thus it can only take on values in steps of, in the case of the DY2A4 series, ~6.2 MGy.

Datasets are sometimes intuitively understood as a coherent unit, not to be broken up. However there is nothing special about a group of diffraction images – with the sole exception that they were obtained at continuous rotation angles from the same crystal. It is, for example, commonplace to exclude diffraction images, or parts of images, during indexing and integration, when for instance the crystal is rotated out of the beam, or when the beam path is obscured by experimental equipment.

Datasets can also be broken up and rejoined. Berglund *et al.* (2002) demonstrated an interesting method for obtaining datasets at very low doses by combining data from multiple crystals. A full diffraction dataset (for example 90°) is collected from a number of identical protein crystals. The crystals are aligned before collection, such that the data collection starts at different rotation angles relative to the first crystal. When 9 crystals are available, the data collection on the second crystal starts at a rotational angle of 10° relative to the initial angle on the first crystal, on the

second crystal 20°, etc. Data are then collected from each crystal. When a rotational angle of 90° is reached, the crystal is rotated back to 0° – the initial position of the first crystal. So for the first crystal data are collected from 0–90°, from the second crystal data are collected from 10–90° and then from 0–10°. For the third crystal: 20–90° followed by 0–20°, etc.

Data from the crystals can then be combined by merging the first 10° wedge obtained from each crystal to get a complete ultra-low dose dataset. Combining the second 10° wedge collected from each crystal yields a dataset with a somewhat higher equivalent dose, and so on. This data collection protocol requires careful indexing and alignment of crystals, which need to be as identical as possible. It has been successfully used in radiation damage research (Ravelli & Garman, 2006).

Merging data from different crystals is a commonly used technique, particularly with crystals that rapidly succumb to global radiation damage, such as virus, membrane, and very small crystals, and in any experiment involving an X-ray Free Electron Laser. When data are collected from many crystals the crystal-to-crystal variation in protein structure needs to be considered. The software BLEND for example applies a clustering algorithm to try and identify isomorphous datasets that merge well together (Foadi *et al.*, 2013).

A novel technique in time-resolved crystallography is using the Hadamard transform (Yorke *et al.*, 2014). Data are collected from multiple crystals at different pump-probe time delays. Because the exposure at any one time delay would be too short to obtain a suitable diffraction pattern the crystal is exposed at a combination of time intervals in a predetermined pulse sequence. The time-dependent components can then be extracted from the collected data by combining them according to the inverse Hadamard matrix.

In radiation-damage induced phasing (RIP) the segmentation of one large dataset into smaller sub-datasets has been proposed by de Sanctis & Nanao (2012). With RIP the phasing signal stems from structural changes within the protein crystal occurring between two datasets due to radiation damage. For an optimised RIP solution a maximum signal (worst possible specific radiation damage) needs to be reconciled with the need for a suitably damaged dataset, that otherwise is still of sufficient quality and completeness. With the segmented RIP method, de Sanctis & Nanao (2012) recommend collecting one large oscillation dataset and then identifying sensible boundaries for the two RIP datasets within this large dataset.

4.4.1 The sliding window technique

The sliding window data acquisition and processing technique depends on a large number of continuously collected diffraction image from a single crystal. A window of fixed size is then moved along the list of collected images to generate a large number of subsets, Figure 4.23. These datasets are then indexed and integrated independently. Each subsequent dataset corresponds to a dataset obtained at a slightly higher overall dose than the previous ones. Therefore a much more detailed and fine-grained image of the development of any radiation damage processes can be obtained.

In the previous Section 4.3, the DY2A4 radiation damage series was extensively covered[9]. Depending on the dataset size, a large number of sub-datasets can be generated: a sliding window size of 180° (m = 1,800 images) results in 63,001 potential sub-datasets.

[9] The radiation damage series CH2A1H1, DY2A4, M1C4, and M2C4 are revisited in this section. See Table 4.4, page 154 for the crystallisation conditions, Table 4.5, page 155 for the data collection protocols, and Table 4.6, page 156 for the integration statistics of their primary datasets.

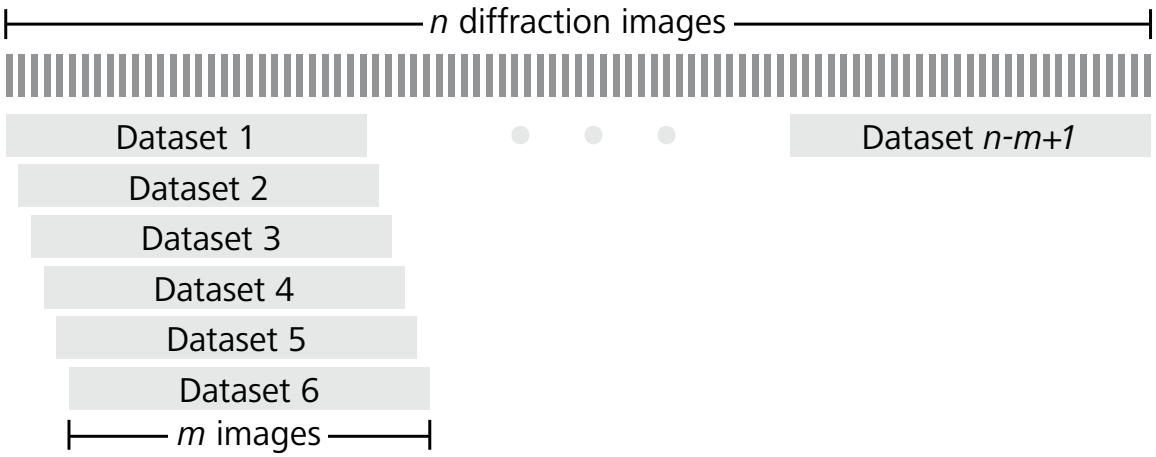


Figure 4.23: With the sliding window data collection and integration protocol a large number of diffraction images is collected continuously. Individual datasets are created by sliding a window of fixed size over the set of collected diffraction images.

4.4.2 Data processing considerations

For obvious reasons it is impractical to manually process each individual dataset. The practicability of the sliding window technique therefore depends on the automation of the entire processing stage from indexing and integration up to the production of electron density maps and protein structure models without the necessity for human intervention. A suitable data processing pipeline has already been demonstrated in the previous section (Figure 4.11, page 167) based on software such as xia2 and REFMAC5.

The data processing requirements also need to be considered. Processing 63,001 datasets is a non-trivial task: assuming a modern machine is capable of processing a dataset within 10 minutes and can run 8 processing jobs in parallel, handling all datasets will still take 55 days. Similarly, there is considerable demand on storage capabilities. The processing requirements can be alleviated by reducing the number of sub-datasets. When instead of every possible dataset with 0.1° resolution only every 10^{th} dataset is considered, the data processing task becomes much more tractable.

To gain a quick overview of the behaviour of metrics throughout the radiation damage series it is recommended to not process the sub-datasets in order[10], but to either randomly sample the sub-dataset space, or to process the datasets in such an order that the maximum distance between any two already processed sets is kept at a minimum[11]. These orderings allow the early analysis of overall trends, with successively more detail becoming available as more and more datasets finish processing.

4.4.3 Window sizes and crystallographic integration statistics

Statistics available at the integration step, such as the number of observed reflections, the resolution, unit cell volume, and the Wilson B factor are all adversely affected by global radiation damage. With the sliding window method the dose-dependent changes of these statistics can be explored in more detail at finer resolution.

The M1C4 radiation damage series (Table 4.5, page 155) was processed using a sliding window of 90° and 180° (equivalent to 450 and 900 images respectively). The M1C4 diffraction data were collected in three successive data collection sweeps (360°, 1800°, and 180°). Some sliding window start angles may therefore cover diffraction data from more than one data collection sweep.

The integration summary statistics, such as the number of observed reflections, oscillates when a sliding window with a size smaller than 180° is used, Figure 4.24. This oscillating behaviour is caused by the exposure of different crystal volume as well as a different amount of crystal volume within neighbouring sub-datasets. Assuming a cuboid crystal, it is clear that for a data

[10] i.e. sub-dataset 1, 2, 3, ...

[11] A suitable ordering for 100 sub-datasets could be to process the first and the last sub-dataset (1, 100), and then to successively fill in the gaps (50; 25, 75; 12, 37, 62, 87; ...) until all sub-datasets are processed.

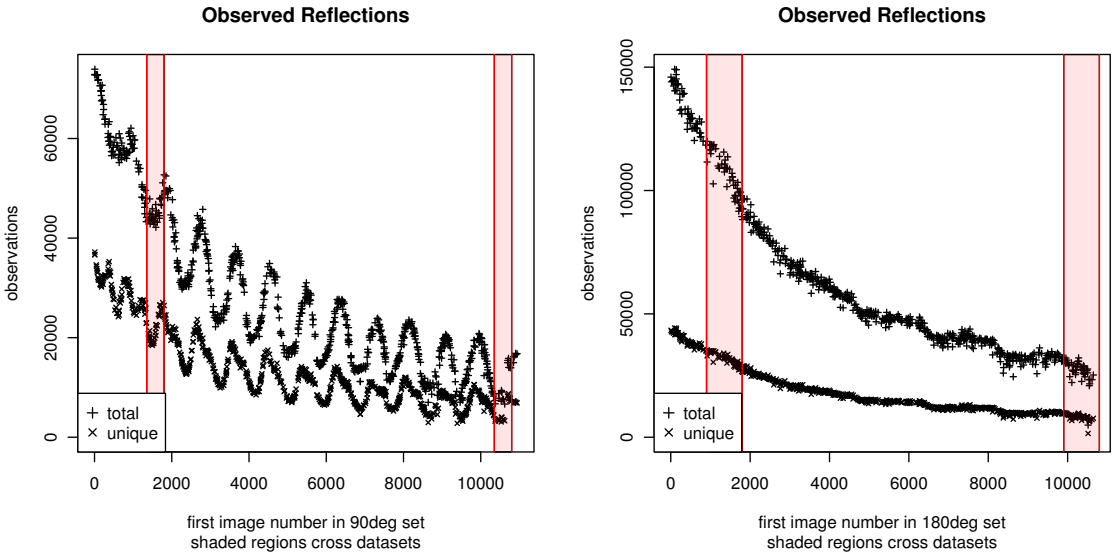


Figure 4.24: Number of observed reflections in the M1C4 radiation damage series (Table 4.5, page 155) for sliding windows of 90° (**left**) and 180° (**right**).
With a window size of 90° a clear oscillation is visible in the data. No oscillation is observed with a 180° window. Markings within the red shaded areas indicate that this sub-dataset crosses the boundary of two separate data collection sweeps.

collection of 90° the exposed crystal volume would vary depending on the starting angle. Intra-crystal variance of the protein structure may also contribute to the varying results (Pozharski, 2012). Because the rotation and the beam axis intersected in the data collection for the M1C4 series, all sliding window datasets of size 180° cover exactly the same crystal volume[12].

The Diffraction Weighted Dose (DWD) was calculated with RADDOSE-3D for a sliding window of 180° on the M1C4 radiation damage series, see Figure 4.25. Even under ideal circumstances, a small oscillation is to be expected. The amplitude of this oscillation in DWD increases with more complex crystal shapes, more uneven beam profiles, and longer experiments (data not shown).

These oscillations can be observed not only in the number of reflections and in the resolution, which directly relate to the exposed crystal volume, but also in the unit cell volume, Figure

[12] A very small oscillation may still be expected if a strong anomalous signal is present. This could then be eliminated by increasing the window size to 360° or by adding a corrective term (half the Bijvoet difference) to the structure factor amplitudes.

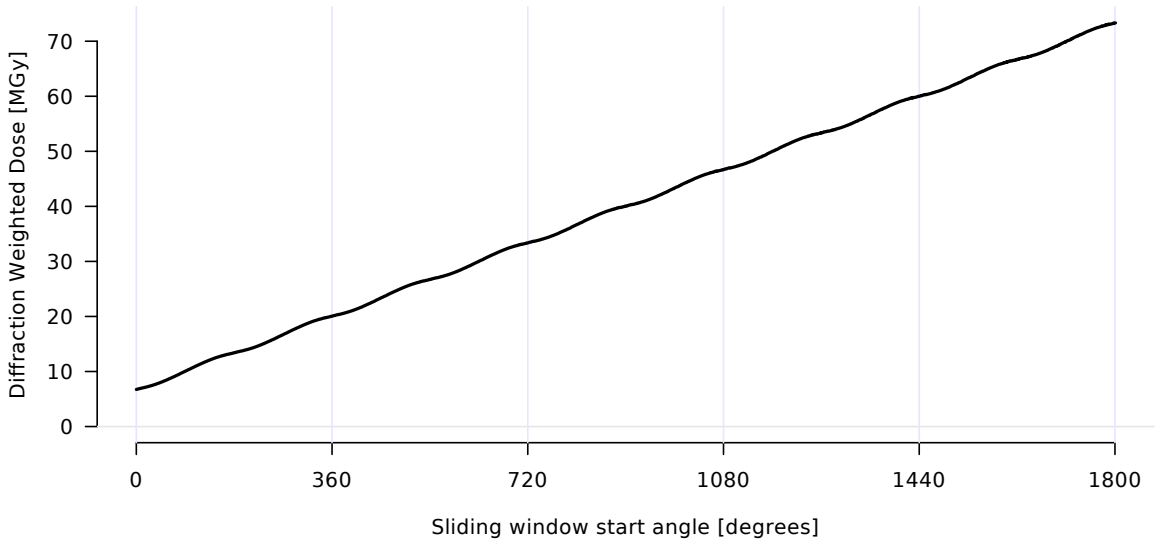


Figure 4.25: The Diffraction Weighted Dose (DWD) was calculated with RADDOSE-3D for a sliding window dataset of 180° on the M1C4 radiation damage series. The DWD increases nearly linearly, but not monotonically. Rotating previously exposed crystal volume in and out of the beam causes an unequal dose distribution within the crystal and a small oscillation in DWD.

4.26. The unit cell size is known to change, usually to increase, with increasing dose in a sample-dependent fashion (Ravelli *et al.*, 2002; Murray & Garman, 2002).

Another consideration before the data acquisition is the possibility of software processing problems at the boundaries between collected datasets. It may not be desirable (or even possible) to set up a single data collection of over 50.000 images at a beamline. On the beamline the data may therefore be collected in several separate sweeps. Even when neither the crystal, the beam, nor the detector is repositioned between these data collection runs, indexing and integration software may still treat a sub-dataset crossing this data collection run boundary as a merged dataset from two different crystals. The sub-dataset may then be treated differently than neighbouring sub-datasets. Some possible artefacts are shown in Figure 4.27, where with a 180° sliding window on the M2C4 radiation damage series some outliers with very low Wilson B factor appear when data from different collection runs are merged.

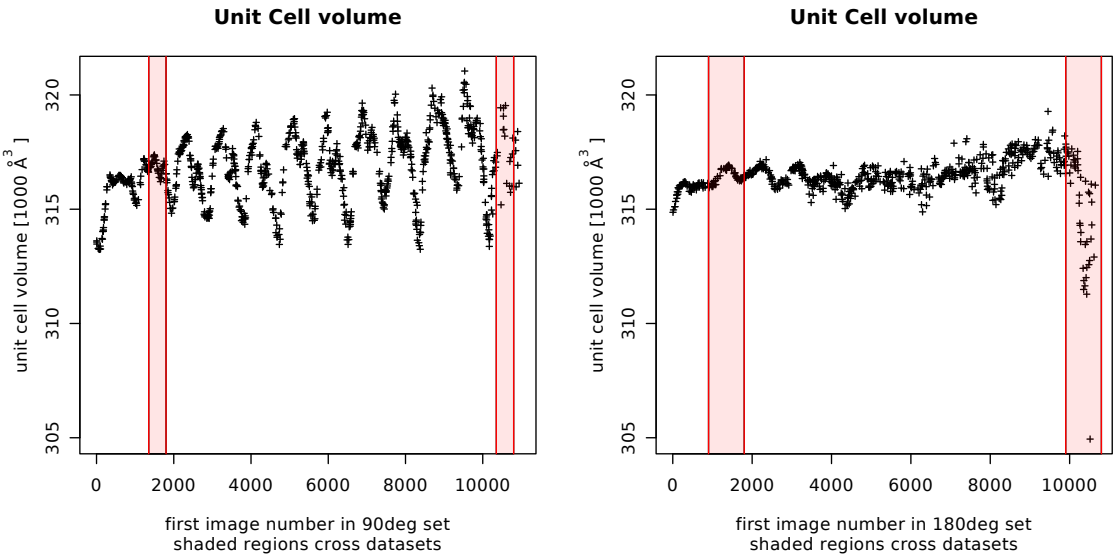


Figure 4.26: The calculated unit cell volume for the M1C4 radiation damage series for sliding windows of 90° (**left**) and 180° (**right**).
The unit cell volume is calculated from the unit cell axes, which, with a window size of 90°, are themselves subject to oscillations (not shown). The variance in the unit cell size for a 180° window is not as pronounced. Markings within the red shaded areas indicate that this sub-dataset crosses the boundary of two separate data collection sweeps.

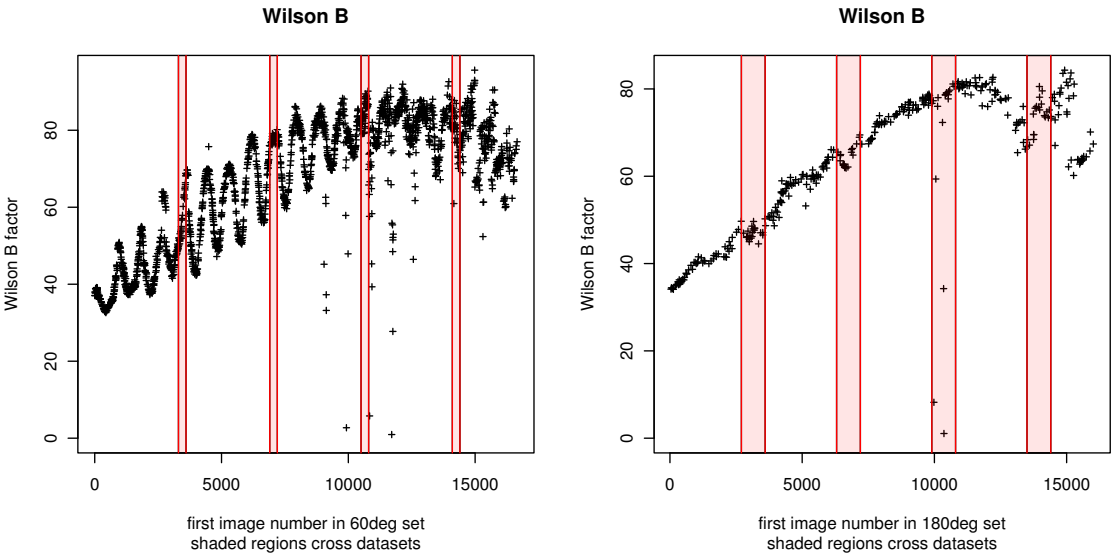


Figure 4.27: The calculated Wilson B factor for the M2C4 radiation damage series (Table 4.5, page 155) for sliding windows of 60° (**left**) and 180° (**right**).
Markings within the red shaded areas indicate that this sub-dataset crosses the boundary of two separate data collection sweeps. A total of five outliers are visible in the region around image number 10,000, and may be caused by the processing software.

To be able to identify potential artefacts introduced by data processing software, it is advised to collect data in 720° segments. This way any processing artefact would be visible only every second full rotation, and could be identified as such.

4.4.4 When should data collection be stopped?

As diffraction data are collected from a protein crystal structural changes due to specific radiation damage accumulate in the crystal. One particular research question is, how one can determine an optimal dataset size (preferably without processing all possible dataset sizes) that reduces or limits the effect of specific radiation damage. With this aim in mind, the suitability of a newly developed quantity, $CC_{1/2}$ (Evans, 2011; Karplus & Diederichs, 2012), was examined using the sliding window technique.

$CC_{1/2}$ has been shown to be a useful statistic in assessing data quality. When the observed diffracted intensities are randomly split into two half datasets, then $CC_{1/2}$ is the Pearson correlation coefficient between those. It is an internal measure of consistency of the collected diffraction data.

Specific radiation damage can be observed in the diffraction pattern as changes in the reflection intensities. If the intensities of a diffraction pattern change over time this may negatively impact the internal consistency and thus cause a drop in the $CC_{1/2}$ statistic.

The radiation damage series CH2A1H1 (Table 4.5, page 155) was processed with xia2 and XDS and POINTLESS as before. The $CC_{1/2}$ was obtained using AIMLESS for a sliding window of 360° and the entire dataset from the start up to a given rotation angle.

AIMLESS returns the resolution at which $CC_{1/2} = 0.5$. The sliding window datasets show an increasing resolution (equivalent to worse data) for later sub-datasets, Figure 4.28. However for

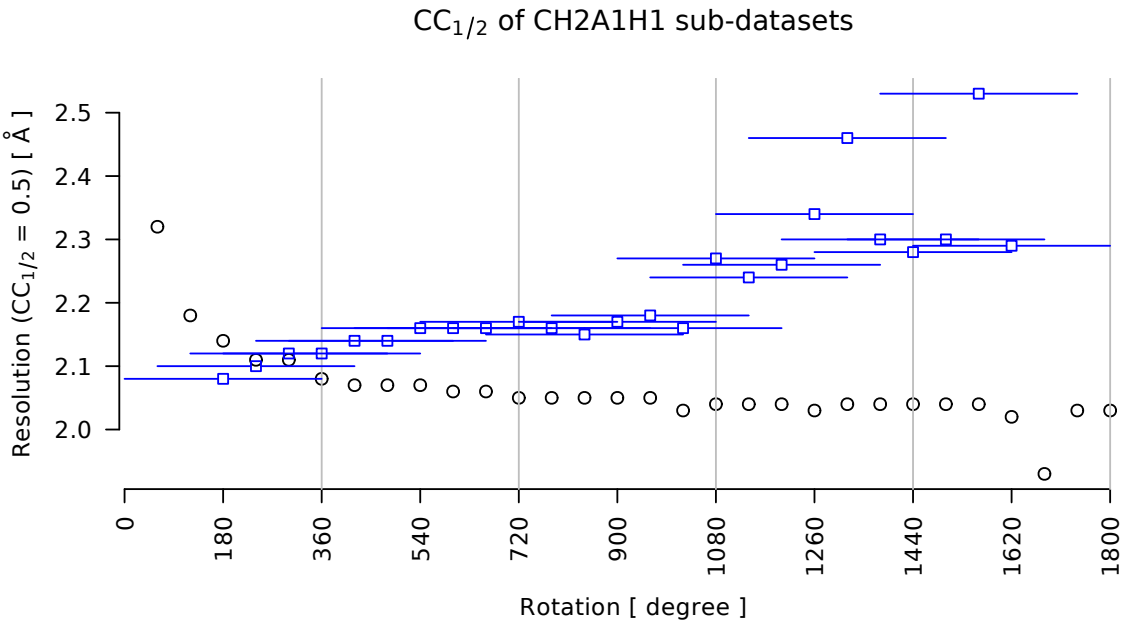


Figure 4.28: Resolution at which $CC_{1/2} = 0.5$ for sub-datasets of the CH2A1H1 radiation damage series (Table 4.5, page 155).
The resolution at which $CC_{1/2} = 0.5$ increases for 360° sliding window sub-datasets (blue) over time. The sliding window extents are marked with a horizontal blue line. When integrating the entire dataset from 0° up to a given rotation angle, the $CC_{1/2}$ statistics continuously improve (black).

the full datasets from 0° up to a given rotation angle this resolution asymptotically decreases, indicating better data. $CC_{1/2}$ of the entire dataset is therefore no suitable indicator for the size of a radiation-damage optimal dataset.

4.4.5 Conclusions

The sliding window technique increases the dose-resolution of observations and provides insight into the variance of key figures such as the unit cell volume and the Wilson B factor. To avoid oscillations in the data, the window size should be appropriately chosen. 180° is a sensible choice for a standard diffraction experiment in which the rotation axis intersects the beam axis.

It is impossible to process the large number of resulting sub-datasets manually, so the sliding

window technique relies on automatable processing. For dataset reduction summary statistics, the only limitation to the technique is currently computing power.

Gaining insight into real space changes requires an automated pipeline, such as the one presented in Figure 4.11, page 167. Together with an automated radiation damage quantification metric, such as the ones explored in Section 4.3, the sliding window technique has the potential to become an essential tool for radiation damage research.

Chapter 5

Conclusions

Conclusions

It has been shown throughout this study that a more detailed experimental understanding of preferential specific radiation damage directly depends on our capability to quantify exactly what is meant by 'radiation damage' and 'susceptibility'. Meaningful analyses rely on combined observations from multiple experiments, and, for example, involve more than one crystal or multiple datasets from the same crystal. Variations of experimental parameters, for example crystal composition and size, can only be taken into account through precise dose estimates. In this concluding chapter the fundamental results on each of the issues addressed in this thesis will be put into context. Possible avenues for further work will be suggested.

5.1 Quantifying specific radiation damage in single datasets

Ideally, dose estimates should be part of any deposited protein structure model. While dose information does give some indication of the generally expected level of specific radiation damage, it neither identifies the level nor location of actual specific radiation damage present in a given protein structure model.

The first attempt at a general solution for this challenge is demonstrated with the new metric B_{Damage}, presented in Chapter 2. By relying on the atomic B factors given in a protein model, B_{Damage} can recognise regions that deviate from the rest of the model in their expected clarity.

On the basis of B_{Damage}, a metric can be envisaged that would diagnose the level of specific radiation damage in protein structure models upon their deposition. For this to happen, a number of open research questions need to be first addressed. It has not yet been shown how the

presence of alternative conformations and the absence of unmodelled atoms of residues should be treated. The detailed information available in anisotropic atomic B factors is, so far, untapped, and provides room for further improvement. Another option to refine B_{Damage} could be by stratifying it for residues and/or different atom types. For example sulphur atoms tend to have a different distribution of associated B_{Damage} values as their basic atomic B factors are different to those of lighter atoms. The damage states of different residues could be made comparable by addressing this implicit difference in B_{Damage} distributions.

Even in its current form, B_{Damage} already provides insight into probable causes of specific and preferential radiation damage, particularly in highlighting the relevance of solvent accessibility promoting radiation damage susceptibility. Its most important contribution arguably lies in opening up the vast body of published protein structure models for radiation damage research. Being able to identify signs of specific radiation damage in a large sample of models puts any observations on a much firmer statistical footing.

B_{Damage} thus provides a powerful tool in the field of radiation damage research. It allows the causes of preferential specific radiation damage to be uncovered through the testing of hypotheses on specific radiation damage mechanisms and symptoms, and the identification of sensitising and stabilising factors.

5.2 The need for precise dose estimates

Systematic studies of specific radiation damage depend on the availability of realistic dose values. Any observed radiation damage effects can only be compared across experiments when the dose values are accurate and reliable. For this, the development of RADDOSE-3D, detailed in Chapter 3,

is an important milestone since it allows the modelling of the absorbed dose within a crystal for the first time as a scalar field. The treatment of the crystal as a set of independent small volume elements has already enabled the development of new dose metrics, such as Diffraction Weighted Dose, which describe the dose state of collected datasets much more accurately than previous dose state summary figures.

However, there is still much room for improvement. All currently obtained dose values include uncertainties in the experimental parameters such as the crystal size and the beam shape and flux. Particularly the latter two can cause a sizeable variation in dose values as described in Section 4.3.4 (page 163). This variation is avoidable: the technology to characterise X-ray beams is not only widely available, but already routinely used at beamlines to monitor beam quality. However even basic beam characterisation data, such as the current beam flux, are not routinely available to the experimenters. The reliability of dose values could be improved even without the installation of new hardware, simply by recording the readings of already available in-beam ionisation chambers as a proxy value for the beam flux in the headers of detector image files. These readings could be regularly calibrated against silicon pin diode measurements, e.g. at the beginning and end of each beamline run, and could also prove useful in the integration and scaling stages of diffraction data processing.

The long-term aim is to integrate dose calculation software such as RADDOSE-3D directly into beamline data acquisition software. This would allow the experimenter to obtain an expected dataset dose based on the crystal composition (given in advance by the experimenter) and size (measured at the beamline using image analysis software or manually marked by the experimenter), the known beam profile, beam conditioning through slits and apertures, the current

beam flux, and the proposed data collection strategy. All necessary calculations would be carried out in the background, and the beamline user could be presented with a simple dose value next to the *'Start data collection'*-button. With this level of integration, the optimisation for low dose could become a relevant factor in selecting the data collection strategy even for experimenters not involved the field of researching radiation damage.

More widespread dose-aware data collection could result in more reliable structure models, ultimately benefiting all users of protein structure models. Readily available dose values would also encourage their publication, which in turn would make systematic studies of radiation damage in published structures much more viable.

5.3 Observing specific radiation damage over time

When diffraction data are collected specifically for radiation damage research, and multiple datasets are available from the same protein crystal along with precise dose values, more complex analyses become possible.

In Chapter 4 it was demonstrated that the monitoring of real space electron density provides an objective criterion by which to measure site-specific radiation damage. By tracking the decay of real space electron density over time, this damage can be identified and quantified. Residues can be assigned with an electron density half-life which describes its particular susceptibility to radiation damage. The information from the diffraction dataset which has been collected can be combined using the sliding window technique to track the decay with very high resolution against the absorbed dose.

Observations from multiple similar crystals can be compared to identify correlations in the

local environment of atoms or residues with their change in susceptibility to preferential specific radiation damage. These correlations, once established, could then serve as a foundation towards theories of radiation damage mechanisms.

The real space electron density quantification, as presented here, depends on a high quality first dataset. If a residue is already partially disordered during the first dataset, then it will not necessarily decay as quickly as a perfectly ordered residue. It may be useful to calculate the electron density decay relative to the idealised 'perfect' electron density expected from the first coordinate model. Another challenge is how alternative conformations present in the first dataset are best accommodated. Should they be tracked separately or should their electron density observations be combined in some way?

The sliding window technique is compelling in that it promises to make the decay process visible in much higher resolution over time. It currently puts large demands on computational processing power as each sub-dataset needs to be processed independently. This issue could be addressed by building the technique directly into data processing software such as DIALS (Fuentes-Montero *et al.*, 2014), which would ensure that the sub-datasets are treated independently, but common calculations are not unnecessarily repeated.

5.4 Improved understanding leads to improved structure models

The common theme in the techniques and analyses discussed here is that they are all aimed at obtaining a better understanding of specific radiation damage. Whether it is a more precise dose estimate for a single experiment, the detection of signs of specific damage in a large set of protein structure models, or the identification of a single susceptible residue and tracking its decay in

a number of different local environments, it is only through observation and experiment that a clearer picture of radiation damage mechanisms can be achieved.

A detailed understanding of this radiation chemistry in macromolecular crystallography will help experimenters make informed decisions on the optimised use of the limited crystal volume in diffraction experiments for instance by 'dose spreading', help to identify the signs of specific radiation damage within the resultant structure models, and generally result in more reliable protein structure models.

Chapter 6

Appendices

Appendix A Relational database used for the PDB analysis

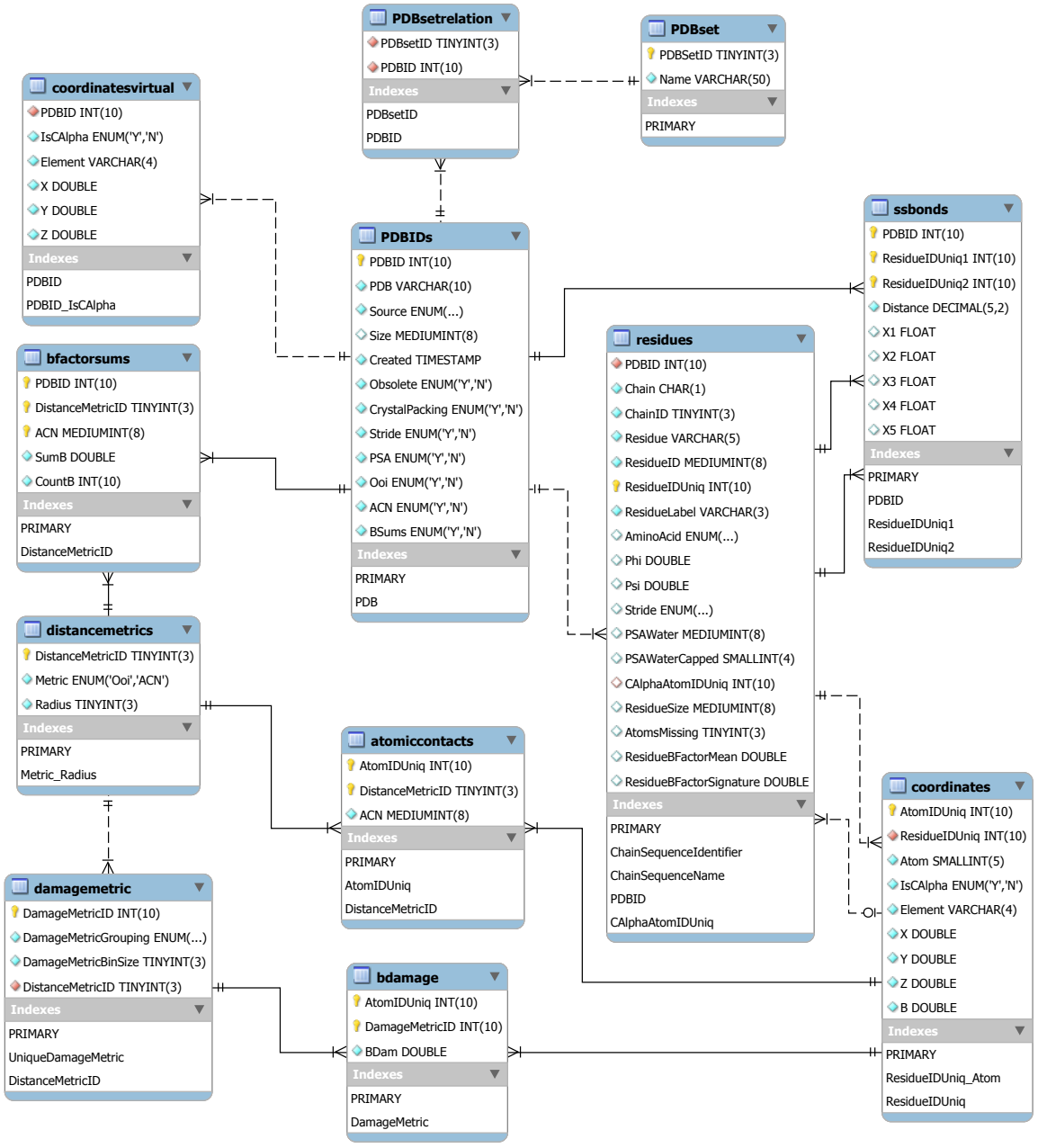


Figure 6.1: Entity-Relationship diagram of the database used for the PDB analysis (Chapter 2, Page 25). Diagram prepared using MySQL Workbench 6.0 (http://mysqlworkbench.org/).

Appendix B List of Protein Data Bank structures used in the statistical survey

Complete list of the 2,704 Protein Data Bank structures which were used in the statistical survey of Chapter 2, Page 25.

1A3A 1A3C 1A4I 1A7D 1AKY 1AL3 1ATZ 1B2L 1B2P 1B4V 1B4Z 1B5E 1B6A 1B8O 1BQC 1BRT 1BW9 1BX4
1C7S 1CJW 1CQX 1CZF 1D2N 1D8W 1DBW 1DEO 1DGW 1DJO 1DJE 1DK8 1DLW 1DPS 1DQG 1DQZ 1DWK 1DYP
1DYQ 1E19 1E2R 1E2W 1E30 1E42 1E4C 1E5M 1E6Y 1E71 1E7R 1EDG 1EEX 1EJD 1ELK 1ELW 1EOK 1EQM
1EWF 1EYE 1EYH 1EZW 1FOL 1F46 1F5V 1F60 1F74 1F7L 1FCQ 1FJ2 1FJJ 1FN9 1FNL 1FO9 1FQT 1FT5
1FTR 1FXO 1G2Q 1G58 1G60 1G6S 1G8M 1GDO 1GGZ 1GKB 1GKK 1GNY 1GP6 1GTJ 1GTK 1GU7 1GUQ 1GV2
1GVG 1GXN 1GXY 1GZJ 1HO3 1H14 1H16 1H2B 1H41 1H4C 1H4R 1H4Y 1H6L 1H72 1H9M 1HBZ 1HD5 1HFE
1HH8 1HT6 1HTW 1HW1 1HX6 1HXI 1IOS 1I1N 1I2H 1I3K 1I86 1IAB 1IDO 1IEO 1IIB 1IJB 1IJT 1ILK
1INL 1IOM 1IQQ 1IV2 1J1M 1J1Y 1J27 1J2W 1J7G 1J83 1J8B 1J8U 1J97 1JA9 1JB3 1JG9 1JH6 1JNE
1JNR 1JP4 1JQ5 1JQC 1JSR 1JWQ 1JY1 1JYH 1K1E 1K2O 1K4N 1K4V 1K6X 1K75 1K77 1K92 1K9Z 1KAP
1KDG 1KG6 1KGS 1KIO 1KIC 1KLL 1KM4 1KOE 1KP4 1KPT 1KSC 1KTG 1KUI 1LOG 1L2H 1L6P 1L7A 1L9X
1LC8 1LJ8 1LK9 1LMI 1LN4 1LO7 1LOV 1LQV 1LV7 1M0W 1M2A 1M2X 1M4I 1M5T 1M65 1M6K 1M6S 1M7J
1M8U 1MG4 1MIX 1MPG 1MR7 1MSK 1MTY 1MXI 1N45 1N57 1N5Q 1N5W 1N7H 1N8U 1N93 1N9E 1NBC 1NBU
1NC5 1NEP 1NF9 1NH8 1NKG 1NLN 1NN5 1NNH 1NOG 1NPY 1NQD 1NRO 1NS5 1NSZ 1NTH 1NTY 1NUO 1NU3
1NVK 1NWA 1NXC 1NYT 1NZJ 1O1Y 1O1Z 1O3Y 1O4R 1O6V 1O9G 1OA1 1OA4 1OAD 1OC6 1OCY 1OD6 1OF8
1OFL 1OFN 1OFZ 1OGO 1OGQ 1OI2 1OJK 1ONW 1OOE 1OQ1 1OQ5 1ORR 1ORU 1OSY 1OW4 1OWL 1OXS 1OYG
1OYW 1OZ2 1POZ 1P1J 1P3C 1P3D 1P5D 1P5Z 1P7S 1P90 1P9H 1PB1 1PBY 1PDO 1PG4 1PKK 1PM4 1PMJ
1PMY 1POA 1PU6 1PV5 1PVM 1PX5 1PXV 1PZ3 1QOG 1Q11 1Q1A 1Q1F 1Q4U 1Q5Z 1Q7E 1Q8D 1Q9U 1QAZ
1QCZ 1QD1 1QF8 1QG8 1QGI 1QH8 1QJP 1QMG 1QNS 1QQ9 1QQJ 1QSA 1QTO 1QTP 1QVY 1QWD 1QWG 1QWO
1QWR 1QZO 1ROU 1R2R 1R3S 1R4P 1R5L 1R5M 1R7A 1R8M 1R9L 1R9Y 1RCV 1RKI 1RLI 1ROC 1RU4 1RV9
1RW7 1RWG 1RWR 1RXQ 1SOI 1S1D 1S1Z 1S5A 1S99 1S9R 1SBX 1SDI 1SFX 1SG6 1SH8 1SHU 1SK2 1SNG
1ST9 1SYY 1TOB 1TOT 1T1U 1T2W 1T3I 1T61 1T6C 1T6E 1T92 1TAG 1TCA 1TE2 1THG 1TJ6 1TJP 1TLU
1TOA 1TP6 1TP9 1TRO 1TU7 1TWD 1TXG 1TY9 1TZW 1UOK 1U14 1U5D 1U6O 1U69 1U6K 1U6R 1U8V 1UAQ
1UAS 1UEK 1UFO 1UGX 1UH4 1UIO 1UJ2 1UKK 1UO9 1UPG 1URR 1US5 1USG 1UT1 1UUQ 1UUX 1UV7 1UX7
1UXL 1UXO 1UZA 1VOT 1V2B 1V3H 1V5D 1V5V 1V7W 1V9F 1VCL 1VDK 1VHF 1VHN 1VIA 1VJF 1VJV 1VK5
1VKA 1VKE 1VKF 1VKN 1VL1 1VL7 1VLA 1VLJ 1VMO 1VMB 1VMJ 1VPB 1VPD 1VPR 1VRM 1VSR 1VYB 1VYI
1WOH 1W1D 1W2W 1W2Y 1W3E 1W4S 1W4W 1W5P 1W6G 1W9I 1WAB 1WBH 1WCU 1WDV 1WEH 1WHI 1WKU 1WLG
1WNY 1WO8 1WOL 1WS8 1WTA 1WTJ 1WUL 1WV3 1WWI 1WWR 1WX4 1WXI 1WZZ 1X1N 1X38 1X3K 1X6O 1X6Q
1X74 1X7D 1X7Y 1X82 1X91 1X9H 1XB3 1XBY 1XC1 1XCR 1XFF 1XFI 1XKP 1XOC 1XOV 1XQL 1XQP 1XSV
1XTE 1XTT 1XV5 1XY7 1Y07 1YOB 1YOH 1YOK 1YOP 1Y21 1Y2T 1Y7B 1Y7R 1YAC 1YB6 1YBI 1YGT 1YJ7
1YKI 1YLE 1YLL 1YN9 1YO3 1YOC 1YQH 1YRK 1YRW 1YS2 1YSQ 1YT3 1YTL 1YUO 1YUM 1YV1 1YWF 1YX1
1ZOB 1Z3E 1Z3X 1Z6N 1Z7A 1Z9N 1ZBF 1ZD7 1ZHS 1ZI9 1ZJC 1ZLD 1ZMT 1ZRS 1ZVT 1ZW6 1ZXU 1ZY7
1ZZ1 1ZZW 2AOB 2A14 2A2K 2A32 2A35 2A3M 2A4D 2A4O 2A5L 2A65 2A6B 2A6V 2A7B 2A94 2A9D 2A9I
2A9S 2AAL 2ABW 2AD6 2AEX 2AFW 2AG4 2AH5 2AHF 2AII 2AJ6 2AJ7 2AL1 2AML 2APJ 2AR1 2ARC 2ASF
2ATF 2AUO 2AVD 2AXC 2AXQ 2BOP 2BOV 2B2H 2B3L 2B4V 2B5H 2B5W 2B61 2B65 2B7U 2B8M 2BBA 2BDR
2BEK 2BEM 2BHY 2BJF 2BJN 2BJQ 2BJV 2BKM 2BKV 2BMW 2BNM 2BP6 2BRJ 2BSY 2BT6 2BUE 2BV9 2BWR
2BZ1 2COH 2C2I 2C2P 2C2T 2C54 2C61 2C6Q 2C6U 2C79 2C8B 2C92 2C9Q 2CBO 2CBZ 2CCM 2CD7 2CDO
2CF7 2CFE 2CHC 2CI3 2CIH 2CIS 2CIU 2CJ2 2CJ4 2CJS 2CKI 2CMW 2CO3 2CWL 2CWR 2CX1 2CXA 2CXH
2CXY 2CYJ 2CZ6 2CZL 2D1G 2D1R 2D29 2D4P 2D5B 2D81 2DBY 2DC3 2DE3 2DEJ 2DF8 2DFY 2DG1 2DG5
2DQA 2DRI 2DSK 2DTC 2DTJ 2DUR 2DVM 2DVN 2DVT 2DWU 2DXA 2DY1 2EOP 2E11 2E1N 2E2O 2E3R 2E7A
2E83 2E8G 2EAE 2EB6 2EB9 2EFF 2EG6 2EHG 2EI1 2EIE 2EJO 2EJN 2ELC 2ELF 2EPN 2ERB 2ET1 2ETJ
2ETV 2EV1 2EVR 2EWR 2EXO 2EX4 2EZ9 2FOC 2F1F 2F1K 2F1N 2F24 2F2B 2F4Q 2F57 2F5X 2F62 2F6D
2F6R 2F7V 2F8A 2F9F 2F9H 2FA1 2FC3 2FCT 2FDV 2FEX 2FGR 2FH1 2FHP 2FI9 2FLU 2FNU 2FOM 2FOZ
2FP1 2FQX 2FR2 2FSR 2FSU 2FSX 2FTN 2FUL 2FUR 2FWO 2FY7 2FYF 2FYG 2FZB 2GOW 2G1U 2G2C 2G40
2G50 2G62 2G82 2G8S 2GA8 2GAI 2GEY 2GFF 2GHS 2GI3 2GIB 2GIY 2GJ4 2GKJ 2GMQ 2GNP 2GS5 2GS8
2GU3 2GUI 2GUY 2GVK 2GWG 2GWM 2GX5 2GZ4 2H1C 2H1T 2H2R 2H6F 2H7J 2H7M 2H8G 2H98 2HA8 2HE2
2HEG 2HHC 2HHJ 2HHP 2HJE 2HKV 2HLY 2HMJ 2HNG 2HPO 2HPL 2HPS 2HQS 2HQY 2HSJ 2HUH 2HXO 2HX5
2HXW 2HY5 2HYT 2HZK 2I3D 2I3H 2I48 2I53 2I5I 2I5R 2I6H 2I74 2I7A 2I7G 2I8D 2I8E 2I9W 2IA1
2IBA 2IBD 2ICA 2ICG 2ICU 2IDL 2IEQ 2IF6 2IG7 2IGI 2IGP 2IIH 2IJA 2IK9 2IKB 2IKK 2IM9 2IMD
2IMH 2IMJ 2IML 2IMZ 2INU 2ION 2IP1 2IPR 2ISB 2IT9 2ITE 2IU5 2IUQ 2IW1 2IWR 2IXD 2IYA 2IZ6
2J0A 2J0P 2J12 2J1S 2J2J 2J43 2J5S 2J6A 2J6G 2J8K 2J97 2J9B 2J9E 2J9N 2J9O 2JB7 2JC5 2JC9

2JCB 2JCN 2JDC 2JE3 2JE8 2JFG 2JG0 2JHM 2JIC 2JK9 2JLP 2NLI 2NML 2NNU 2NOO 2NPT 2NRK 2NRT
2NT0 2NTP 2NUO 2NVF 2NW0 2NW8 2NXF 2NXW 2NYH 2NYI 2OO2 2OOM 2O1Q 2O2I 2O2P 2O2X 2O3F 2O3S
2O5V 2O62 2O6P 2O6Y 2O70 2O7I 2O8Q 2OA9 2OB5 2OC3 2OC5 2OD4 2OD5 2OEB 2OFK 2OIT 2OKF 2OKM
2OKO 2OLR 2OMK 2ONS 2OO1 2OOC 2OOK 2OPL 2OQG 2OQZ 2OR7 2ORW 2OS1 2OSA 2OSV 2OT4 2OU5 2OU6
2OUI 2OX6 2OX7 2OY2 2OY7 2OYA 2OYK 2OYO 2OZJ 2OZV 2P0S 2P12 2P14 2P17 2P1F 2P1M 2P3H 2P3P
2P58 2P6W 2P97 2PA7 2PBD 2PBK 2PET 2PFZ 2PKF 2PLR 2PN0 2POR 2PQV 2PRX 2PS1 2PU3 2PWY 2PXR
2PYQ 2PYX 2QO3 2QOI 2QOS 2Q35 2Q62 2Q7D 2Q8K 2Q99 2QA1 2QAC 2QAP 2QB2 2QB7 2QEE 2QEU 2QF9
2QG6 2QGD 2QGI 2QGU 2QGY 2QHF 2QHP 2QHQ 2QHS 2QIB 2QIS 2QK1 2QKP 2QL8 2QLW 2QML 2QNI 2QNK
2QNL 2QRU 2QSA 2QSI 2QSW 2QUB 2QUD 2QUO 2QVU 2QW5 2QWU 2QXF 2QZC 2QZT 2QZU 2R1I 2R2C 2R4I
2R4Q 2R78 2R7G 2R85 2R8Q 2R9F 2RBD 2RBG 2RBW 2RCI 2RDC 2RDG 2RDS 2RFG 2RFQ 2RH3 2RIQ 2RJ2
2RKQ 2RKV 2RLD 2SAK 2SGA 2UUR 2UVJ 2UVK 2UVP 2UWA 2UY2 2UYK 2UYQ 2UYT 2UZ1 2VOH 2V2B 2V2G
2V38 2V3Z 2V4V 2V4X 2V5J 2V6A 2V6U 2V75 2V76 2V7K 2VAC 2VAP 2VB9 2VBP 2VBU 2VCL 2VFK 2VFO
2VFT 2VGO 2VG8 2VK6 2VLQ 2VLZ 2VMH 2VOW 2VPN 2VQ2 2VQP 2VRS 2VSM 2VUW 2VV6 2VVE 2VVP 2VXZ
2VY7 2VYW 2WOB 2WOI 2W31 2W3E 2W3P 2W3X 2W52 2W5T 2W61 2W6K 2W7Z 2W86 2W87 2W92 2W9Y 2WAN
2WAO 2WAW 2WBX 2WCM 2WCR 2WCW 2WDC 2WFO 2WFP 2WH6 2WHE 2WHM 2WI8 2WJ9 2WJR 2WKK 2WMF 2WN3
2WN9 2WNB 2WOP 2WOY 2WQK 2WR8 2WTE 2WTG 2WTP 2WU9 2WUQ 2WVF 2WW5 2WXU 2WY8 2WYA 2WYK 2WZ8
2WZC 2WZO 2X1D 2X21 2X2S 2X32 2X3G 2X3N 2X49 2X4J 2X4L 2X4W 2X5C 2X5F 2X85 2X8R 2X8S 2X9O
2X9X 2XBG 2XBU 2XC3 2XDP 2XE4 2XF3 2XFN 2XGR 2XHG 2XJ4 2XLG 2XMI 2XMX 2XOV 2XPP 2XRH 2XRY
2XSU 2XTP 2XU9 2XVS 2XVY 2XWS 2XXF 2XXN 2XXP 2XXZ 2XZ9 2XZE 2XZK 2YO9 2YOG 2Y1E 2Y1Q 2Y27
2Y2M 2Y2X 2Y3B 2Y3Q 2Y44 2Y4R 2Y51 2Y6X 2Y71 2Y7D 2Y7O 2Y8G 2Y8N 2YAV 2YB6 2YC5 2YET 2YFD
2YFU 2YG2 2YH6 2YHC 2YJG 2YL6 2YMM 2YMV 2YNO 2YNT 2YOA 2YOG 2YR4 2YV9 2YVT 2YWI 2YWV 2YXM
2YXN 2YYK 2Z0J 2Z0T 2Z0X 2Z14 2Z1E 2Z2N 2Z3H 2Z6O 2Z6R 2Z84 2Z8F 2Z9B 2Z9W 2ZA4 2ZAC 2ZAD
2ZB4 2ZBL 2ZBT 2ZCA 2ZCO 2ZDP 2ZFG 2ZFI 2ZFY 2ZGW 2ZHI 2ZIN 2ZKM 2ZM9 2ZQO 2ZQO 2ZUI 2ZVY
2ZWA 2ZYC 2ZYT 2ZZJ 2ZZX 3A0Z 3A2Z 3A3D 3A47 3A54 3A57 3A77 3A9I 3A9S 3A9Z 3AAO 3AAM 3AB8
3ACG 3AHC 3AHN 3AI9 3AIN 3AJ3 3AJR 3AKC 3ALL 3AML 3AOW 3AQ2 3ARQ 3ASQ 3ATS 3ATV 3AV3 3AWM
3AXD 3B2X 3B49 3B53 3B5E 3B6E 3B7C 3B7I 3B9T 3BA3 3BBO 3BB7 3BB9 3BCW 3BCY 3BEX 3BFM 3BFP
3BGY 3BHD 3BI1 3BI7 3BIQ 3BIY 3BJN 3BL9 3BLZ 3BN6 3BND 3BNG 3BOD 3BOF 3BOH 3BPK 3BPT 3BPZ
3BQC 3BRC 3BS4 3BS6 3BVF 3BWV 3BXP 3BY4 3BY9 3BYQ 3BZT 3C1Q 3C4B 3C5K 3C5N 3C7X 3C8C 3C8E
3C8W 3C9A 3C9Q 3CA1 3CA8 3CAN 3CCG 3CE7 3CEC 3CH4 3CHJ 3CHM 3CI6 3CIN 3CJ1 3CJM 3CJY 3CKC
3CKJ 3CKK 3CL5 3CLA 3CLS 3CNM 3COV 3CPG 3CQL 3CRY 3CT1 3CTZ 3CUI 3CV1 3CVO 3CWR 3CXK 3CXM
3CXN 3CYP 3CZ1 3CZX 3DOF 3DOJ 3D33 3D34 3D40 3D59 3D7I 3D8T 3D97 3D9N 3DAN 3DCZ 3DD7 3DDA
3DDO 3DE8 3DEO 3DF8 3DFF 3DFG 3DG9 3DGT 3DHP 3DI4 3DJ8 3DJE 3DKR 3DLU 3DMN 3DN7 3DNU 3DRJ
3DS8 3DT2 3DWG 3DXT 3DXY 3DZA 3E0I 3E0Z 3E11 3E2V 3E3U 3E4W 3E7H 3E9K 3E9T 3EBV 3EC4 3EC9
3ECD 3EDF 3EDN 3EEA 3EF2 3EF3 3EF8 3EFY 3EH1 3EHG 3EJF 3EJN 3EK3 3EKI 3EKL 3ELG 3EMI 3ENO
3EO7 3EOI 3EQN 3EQX 3ER7 3ERJ 3ERP 3ES4 3ESL 3EU3 3EVO 3EW1 3EW8 3EWN 3EZI 3EZM 3F0H 3F0M
3F0P 3F2E 3F43 3F44 3F47 3F4M 3F4S 3F52 3F5R 3F67 3F6J 3F7Q 3F8X 3F95 3F9M 3F9S 3FBU 3FCX
3FD3 3FDH 3FDJ 3FDX 3FF1 3FFR 3FG8 3FGE 3FGR 3FH1 3FIQ 3FJ1 3FKA 3FKC 3FL2 3FLA 3FM2 3FMC
3FN5 3FO8 3FOJ 3FOT 3FP5 3FPF 3FPK 3FPW 3FRQ 3FRR 3FSU 3FT1 3FUN 3FUW 3FVS 3FW2 3FW8 3FWY
3FWZ 3FYB 3GO2 3GOM 3GOT 3G3H 3G48 3G7U 3G7X 3G89 3GAE 3GAX 3GBY 3GDO 3GF3 3GF6 3GFP 3GG7
3GK6 3GKJ 3GKM 3GKR 3GMF 3GMG 3GMV 3GN6 3GNL 3GO9 3GOC 3GOQ 3GRS 3GS9 3GT3 3GVO 3GWI 3GWN
3GY9 3GYK 3GZA 3H05 3H09 3HOU 3H2S 3H3L 3H46 3H4O 3H4W 3H4Y 3H51 3H5L 3H68 3H6J 3H6Q 3H7C
3H7H 3H7I 3H7V 3H8G 3H8T 3H93 3H9B 3H9M 3HA2 3HAP 3HBM 3HD4 3HDO 3HDX 3HID 3HJ4 3HJB 3HJR
3HKW 3HLO 3HLZ 3HMZ 3HNO 3HN5 3HNA 3HNH 3HO6 3HOI 3HP7 3HQ1 3HQM 3HQX 3HRP 3HSY 3HTN 3HU5
3HV2 3HVV 3HYQ 3HZ6 3I07 3I0Y 3I0Z 3I1A 3I24 3I47 3I48 3I4Z 3I6I 3I96 3IAR 3IBZ 3IDU 3IEE
3IGS 3IHV 3IJM 3IKB 3ILF 3ILR 3ILS 3ILW 3IMO 3IM1 3IOX 3IQ1 3IRB 3IRP 3IRS 3ISQ 3IT3 3IU5
3IU6 3IUO 3IV4 3IVZ 3JRV 3JS8 3JSC 3JSY 3JUO 3JXS 3JYZ 3JZ9 3K00 3K0B 3K0X 3K1U 3K1W 3K1Z
3K26 3K3C 3K40 3K4I 3K6M 3K6Q 3K7G 3K8W 3KB9 3KBG 3KBY 3KDW 3KEO 3KGO 3KG9 3KGW 3KGY 3KI6
3KIZ 3KLO 3KMH 3KMT 3KMV 3KOR 3KQR 3KS6 3KSN 3KSX 3KT7 3KU3 3KV1 3KVH 3KVS 3KWS 3L00 3L1F
3L1W 3L34 3L3X 3L40 3L4A 3L4E 3L4H 3L4N 3L51 3L6B 3L8Q 3LAT 3LB4 3LCC 3LD7 3LEW 3LFJ 3LFR
3LFU 3LGB 3LGI 3LHE 3LHO 3LI3 3LID 3LIM 3LIO 3LKM 3LLO 3LLP 3LLX 3LM2 3LOG 3LOP 3LRT 3LSC
3LTJ 3LVU 3LXO 3LX3 3LXS 3LY1 3LYG 3LYH 3LZA 3LZN 3LZW 3MOH 3MOJ 3M1H 3M1T 3M1U 3M66 3M6L
3M76 3M7O 3M84 3M86 3M9Z 3MAK 3MD9 3MDM 3MDQ 3ME7 3MF7 3MHZ 3MIL 3MK1 3MKO 3ML1 3MMG 3MOK
3MPC 3MTO 3MVU 3MW8 3MWL 3MWZ 3MXN 3MXO 3MXU 3MXZ 3MYV 3MZ2 3MZF 3MZQ 3NO3 3NOU 3NOX 3N1E
3N34 3N37 3N6Y 3N79 3N80 3N8I 3N9C 3N9K 3NA5 3NAD 3ND1 3NDD 3NEC 3NEH 3NEU 3NFT 3NFW 3NGV
3NJC 3NJK 3NKQ 3NM6 3NMW 3NNB 3NOH 3NOK 3NPD 3NPF 3NPK 3NR5 3NRE 3NRF 3NRH 3NRS 3NRW 3NRX
3NSL 3NSW 3NSX 3NT1 3NT3 3NUQ 3NV1 3NWE 3NYH 3NZE 3OOQ 3OOY 3O12 3O2G 3O2H 3O48 3O4Q 3OAM
3OBE 3OBU 3OC9 3OCC 3OCV 3OEB 3OEP 3OF5 3OGH 3OHG 3OIO 3OJC 3OKX 3OLL 3OM1 3OMC 3OMD 3ON1
3ON9 3ONH 3ONO 3OO8 3OOO 3OOS 3OOU 3OQI 3OS4 3OS7 3OSE 3OST 3OSX 3OT9 3OTI 3OTM 3OU2 3OUG
3OUI 3OUL 3OXH 3P02 3P1G 3P2C 3P2E 3P2T 3P3G 3P3V 3P4G 3P97 3P9A 3P9V 3PA6 3PBT 3PCX 3PCZ

```
3PDD 3PE7 3PFE 3PFT 3PG6 3PGU 3PI6 3PI7 3PJ0 3PJP 3PJY 3PLN 3PM2 3PMD 3PME 3PMS 3PNA 3PO8
3POF 3POH 3POP 3POW 3PPM 3PR6 3PU9 3PUI 3PVH 3PWK 3PYW 3QOW 3Q18 3Q1C 3Q1F 3Q1N 3Q1P 3Q1X
3Q20 3Q26 3Q2B 3Q3M 3Q49 3Q4R 3Q60 3Q64 3Q6B 3Q7C 3Q7M 3Q7W 3QB8 3QC7 3QF2 3QH6 3QH8 3QHO
3QNM 3QOR 3QOU 3QP8 3QS2 3QSJ 3QSQ 3QUF 3QUT 3QUV 3QWP 3QY0 3QY1 3QY3 3QY9 3QZ4 3R1M 3R26
3R4V 3R4Z 3R5G 3R5S 3R5Z 3R62 3R6H 3R6U 3R8J 3R9Z 3RAY 3RC4 3RD5 3RF0 3RF3 3RGA 3RHG 3RJT
3RJU 3RJV 3RKC 3RLH 3RLK 3RLO 3RLS 3RNL 3ROO 3RPD 3RPP 3RPW 3RPZ 3RQ7 3RQT 3RR6 3RRI 3RT2
3RU6 3RUO 3RUR 3RVC 3RWV 3RZV 3S0E 3S19 3S21 3S3U 3S5B 3S5F 3S5V 3S7P 3S82 3S8M 3S9J 3S9Z
3SAL 3SBF 3SBQ 3SC7 3SCY 3SGH 3SGV 3SH4 3SHG 3SIH 3SJ5 3SK7 3SK9 3SLR 3SM1 3SMJ 3SNS 3SQR
3SS7 3SSX 3ST1 3STD 3SXM 3SZ3 3T01 3T0L 3T0W 3T1K 3T4H 3T4L 3T50 3T63 3T68 3T7A 3T7D 3T7H
3T8G 3T8J 3T8K 3T90 3T92 3T95 3T9G 3T9L 3T9W 3TA5 3TAW 3TBD 3TC2 3TC3 3TC7 3TCQ 3TCV 3TDW
3TE8 3TEF 3TFJ 3TG7 3TGH 3THD 3TIP 3TJ8 3TJL 3TK0 3TKF 3TL1 3TLZ 3TMB 3TNT 3TNY 3TOS 3TQE
3TQL 3TT9 3TUO 3TUT 3TVK 3TW1 3TYT 3U15 3U1D 3U1L 3U26 3U2A 3U2V 3U3L 3U3Z 3U49 3U4V 3U9J
3UB1 3UC1 3UDF 3UF6 3UF8 3UFA 3UFB 3UFE 3UID 3UJD 3UMO 3UOA 3UP0 3UPL 3UPS 3UQ8 3URK 3USH
3UTK 3UUW 3UV9 3UW1 3UW3 3UWS 3UX2 3UXE 3UXF 3V0E 3V2I 3V38 3V46 3V4M 3V5C 3V66 3V68 3V7B
3VAV 3VBC 3VBJ 3VC5 3VEN 3VEU 3VFN 3VGL 3VHE 3VJ9 3VJZ 3VKA 3VKJ 3VL1 3VLV 3VM1 3VMN 3VMV
3VNB 3VNR 3VNY 3VOT 3VPB 3VPI 3VPQ 3VRD 3VRP 3VSY 3VTC 3VTU 3VTV 3VTX 3VU9 3VV1 3VV3 3VVY
3VWC 3VXG 3VXI 3VZ3 3VZH 3VZX 3W08 3WOK 3WOO 3WOQ 3WOS 3W15 3W2G 3W4S 3W52 3W7T 3W9K 3W9S
3WA7 3ZBO 3ZFN 3ZHF 3ZJH 3ZJI 3ZJL 3ZJM 3ZN6 3ZQI 3ZQK 3ZR0 3ZRI 3ZRV 3ZSU 3ZT9 3ZUI 3ZUZ
3ZWF 3ZWT 3ZX4 3ZYH 3ZYL 3ZYP 3ZYU 4A0D 4A2B 4A30 4A35 4A37 4A42 4A4Y 4A5K 4A5S 4A6Q 4A6U
4A6X 4A84 4A8T 4AC7 4ACY 4ADI 4ADN 4ADZ 4AE2 4AE4 4AEK 4AFV 4AGH 4AGP 4AHW 4AIW 4AJY 4AMF
4AML 4AO5 4AO6 4AO9 4AQ4 4AQJ 4AR9 4ARO 4ASC 4ASM 4ATM 4AUJ 4AVA 4AVD 4AWU 4AWY 4AXX 4AY0
4AZ7 4AZJ 4B0Z 4B1L 4B1V 4B2N 4B2O 4B4C 4B4P 4B72 4B89 4BA1 4BC3 4BCD 4BE3 4BGO 4BGP 4BH5
4BJ4 4BJT 4BJZ 4BK1 4BK3 4BOL 4BOU 4BQH 4BR5 4BR7 4BRA 4BRD 4BRK 4BWR 4D9B 4D9S 4DAM 4DB5
4DBC 4DDP 4DF3 4DH2 4DKK 4DL8 4DLN 4DM5 4DMC 4DMI 4DMK 4DMV 4DN7 4DNX 4DNY 4DO7 4DOI 4DOM
4DOY 4DQ6 4DQA 4DQD 4DRR 4DTH 4DUH 4DV8 4DWD 4DWN 4DY5 4DYQ 4DZO 4E15 4E1A 4E1B 4E2U 4E32
4E4T 4E5V 4E70 4EAD 4EBJ 4EBY 4ECF 4ECP 4EE7 4EF0 4EF8 4EHS 4EHU 4EKX 4EL6 4EMT 4EO0 4EPZ
4EQB 4EQG 4EQL 4ES5 4ES8 4ESQ 4ESU 4ESW 4EU3 4EUS 4EW1 4EW7 4EWR 4EX7 4EYS 4F0J 4F1J 4F3N
4F54 4F67 4F6T 4F8L 4FBS 4FCJ 4FCS 4FDB 4FDM 4FDV 4FE3 4FFL 4FGQ 4FKZ 4FNV 4FO7 4FOJ 4FP1
4FQS 4FR7 4FUS 4FVC 4FVG 4FVY 4FX5 4FYP 4G1J 4G1R 4G22 4G26 4G29 4G38 4G3V 4G48 4G4K 4G4P
4G4X 4G54 4G55 4G79 4G9M 4G9P 4G9Q 4GB5 4GBM 4GCY 4GD6 4GE3 4GEI 4GEK 4GF0 4GFT 4GGJ 4GH9
4GHG 4GHI 4GHN 4GKB 4GLQ 4GMM 4GNB 4GNF 4GNM 4GNO 4GNV 4GNY 4GOB 4GPV 4GQ6 4GT6 4GTA 4GUJ
4GVR 4GWK 4GXB 4GYF 4GYW 4GZJ 4GZK 4HO4 4HO8 4HOC 4H14 4H2G 4H41 4H59 4H5G 4H5S 4H6R 4H7U
4H7Y 4H8N 4HBA 4HBZ 4HD9 4HDK 4HF7 4HFS 4HGI 4HH3 4HHJ 4HHR 4HJF 4HKF 4HL1 4HLB 4HLC 4HMM
4HMU 4HMW 4HNG 4HPA 4HPU 4HQ1 4HRY 4HST 4HV3 4HW4 4HW6 4HXY 4HYQ 4HZ2 4HZ4 4HZA 4I1I 4I1K
4I3F 4I4V 4I56 4I6X 4I70 4I74 4I79 4I8I 4I90 4IA6 4IA9 4IAB 4IBN 4IDD 4IDH 4IHZ 4IIK 4IJI
4IJN 4IKC 4IKD 4IMQ 4IN9 4IND 4IO1 4IPB 4IPI 4IQM 4IQN 4IQY 4IR0 4IRG 4IRT 4IVA 4IVK 4IZB
4J27 4J2K 4J32 4J4H 4J4M 4J5F 4J5S 4J6O 4J79 4J7H 4J7N 4J8P 4J8S 4JA8 4JB3 4JBU 4JC1 4JDF
4JG3 4JGI 4JIF 4JKC 4JMD 4JN3 4JND 4JNF 4JPX 4JQF 4JQP 4JTF 4JTI 4JTJ 4JVB 4JWO 4JXC 4K05
4K0X 4K37 4K3F 4K3S 4K45 4K4K 4K60 4K73 4K8W 4KA2 4KAY 4KFN 4KH8 4KIL 4KK3 4KK4 4KK7 4KKV
4KM6 4KMY 4KNU 4KO1 4KP6 4KQ7 4KQC 4KSJ 4KUA 4KV1 4KVP 4KVS 4KWH 4KYQ 4KYV 4KZK 4L4G 4L5H
4L83 4L8E 4LBO 4LBA 4LC2 4LD6 4LFG 4LFY 4LG1 4LHR 4LLE 4LLS 4LMI 4LQB 4LRD 4LRU 4LSM 4LUI
4LW2 5CYT 5NUL 7ODC
```

Appendix C RADDOSE-3D user guide

The RADDOSE-3D user guide describes the syntax of the input files understood by RADDOSE-3D, and is included here for reference. The most current version of the user guide is always available online at http://raddo.se/rd3d/RD3D-User-Guide.pdf.

RADDOSE-3D Command Reference

version 1.0.950 – 15 July 2013

Please cite the following publication when using RADDOSE-3D:

RADDOSE-3D: time- and space-resolved modelling of dose in macromolecular crystallography
Zeldin, O. B.; Gerstel, M. & Garman, E. F. (2013). *J. Appl. Cryst.* **46**
doi:10.1107/S0021889813011461

Contents

RADDOSE-3D can take input from one or more files and/or from standard input (STDIN). Any input will be processed by the InputParser class and the RADDOSE-3D ANTLR parser. This section describes the syntax of accepted input. Advanced users of RADDOSE-3D can create their own input method that need not rely on the InputParser class or the RADDOSE-3D ANTLR parser. This feature will not be covered in this reference.

The simplest use case of RADDOSE-3D will involve only one file describing the entire experiment. In some instances it may be desired to split up the input into a number of files, e.g. one file describing the crystal, one automatically updated file describing the current beam on the beamline, and one file chosen from a set of possible wedge strategies. Each file can contain an arbitrary number (including none) of Crystal, Beam and Wedge block (henceforth called *blocks*). However, splitting up blocks across multiple files is not allowed.

The parser will read the input sequentially, and, when multiple sources are given, one source after the other in the specified order. While the parser may accept Crystal, Beam and Wedge blocks in any order, the exposure of a wedge can only take place if both the crystal and the beam have been set either in an earlier file or before the Wedge block within the same file.

1 General syntax considerations

Any keywords specified below are case-insensitive. Upper (**CRYSTAL**), lower (**crystal**) and mixed case (**CrYsTaL**) are equivalent.

The characters **#**, **!** and the character sequence **//** denote the start of a comment. Any text from that position until the end of the current line is ignored.

Tabular and newline characters are treated as white space. They can therefore by freely used to format the file for increased readability.

The order of statements within a Crystal, Beam and Wedge block generally is not relevant. There are two exceptions to this rule: The leading keyword (**CRYSTAL**, **BEAM**, **WEDGE**) **must** be the first keyword of the block. If a keyword is repeated within the same block, then the latter will always override the former.

Every block must be self-contained, e.g. the energy set for the previous Beam is not remembered when setting up the following Beam, and must be repeated.

Numeric values can be given in scientific notation ($2.0\text{e}2 = 2e{+}2 = 200$), negative values may not have a space between the sign ('-') and the value ($-1.9\text{e}-1 = -.19 = -0.19$).

2 Crystal block

A Crystal block must begin with the keyword **CRYSTAL**. At least the **TYPE** and **DIMENSION** must be specified. Depending on the chosen **TYPE** further declarations may be required.

2.1 TYPE

With the keyword **TYPE** the underlying crystal implementation is chosen. Currently two distinct crystal implementations exist:

TYPE CUBOID defines a solid crystal with a cuboid shape.

TYPE SPHERICAL defines a solid crystal with a spherical shape.

2.2 DIMENSION

DIMENSION specifies the size of the crystal. Dimensions are given in micrometres (µm). The keyword **DIMENSION** can take either one or three parameters:

DIMENSION *D* with a single number (see section 1) as parameter is used for specifying the crystal dimensions for spherical crystals. The parameter sets the crystal diameter. This syntax cannot be used for cuboid crystals.

DIMENSION *X Y Z* with three numbers as parameters X, Y and Z is used to set the dimensions for cuboid crystals (**TYPE CUBOID**). X defines the length of the crystal orthogonal to both the beam and the goniometer at L=P=0, (see below) Y defines the length along the goniometer axis at L=P=0 and Z defines the length along the beam axis.

If three parameters are given for a spherical crystal (**TYPE SPHERICAL**) the value for X sets the diameter of the crystal while the values of Y and Z are ignored.

2.3 PIXELSPERMICRON

PIXELSPERMICRON *F* specifies the resolution of the voxel grid used to represent the crystal in voxels/µm. Defaults to 0.5 voxels/µm.

2.4 ANGLEP

ANGLEP *F* sets the angle in the plane of the loop between the crystal Y axis and the goniometer axis. The angle is to be given in degrees, but without the degree symbol (°). The default P ('plane') angle is 0°.

The rotation angle to be applied to the crystal in the plane of the loop (right handed rotation about Z axis applied to all voxels, as shown in figure 1).

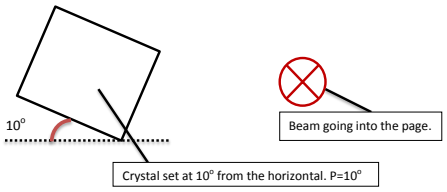


Figure 1: Schematic of **ANGLEP**. Figure courtesy of John Bremridge.

2.5 ANGLEL

ANGLEL *F* sets the loop angle between the plane of the crystal loop and the goniometer axis. The angle is to be given in degrees, but without the degree symbol (°). The default L ('loop') angle is 0°.

The rotation angle to be applied to the angle of the crystal in the loop (right handed rotation about X axis applied to all voxels, as shown in figure 2).

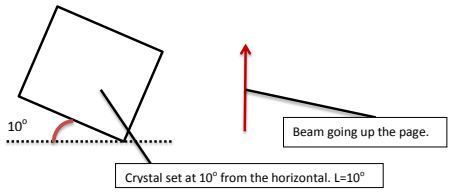


Figure 2: Schematic of **ANGLEL**. Figure courtesy of John Bremridge.

2.6 ABSCOEFCALC

This keyword specifies whether the program should use average absorption and attenuation coefficients, or whether it should calculate them from input crystal parameters.

ABSCOEFCALC AVERAGE
ABSCOEFCALC DUMMY

These two commands are equivalent. Each will cause RADDOSE-3D to assume an absorption coefficient of 0.237 mm^{-1} and an attenuation coefficient of 0.281 mm^{-1}. These values are representative of an average crystal at an incident X-ray beam energy of 12.4 keV (1Å). Please see Section 3 in the main paper for more details. Crystal composition keywords will have no effect.

ABSCOEFCALC RD
ABSCOEFCALC RDV2
ABSCOEFCALC RDV3

These three commands are equivalent. RADDOSE-3D will call a previous version of RADDOSE to estimate absorption and attenuation coefficients.

The composition of the crystal has to be described using the keywords **UNITCELL**, **NUMMONOMERS**, **NUMRESIDUES**, **NUMRNA**, **NUMDNA**, **PROTEINHEAVYATOMS**, **SOLVENTHEAVYCONC** and **SOLVENTFRACTION**. The use of these keywords is described in the sections 2.7–2.14 below.

2.7 UNITCELL

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

UNITCELL *A B C*
UNITCELL *A B C* α β γ

Dimensions and angles of the unit cell a, b, c, α, β, γ

The first three numbers specify the unit cell size in Angstroms. The second three numbers optionally specify the unit cell angles alpha, beta and gamma.

The (optional) angles are to be given in degrees, but without the degree symbol (°). If no angles are specified RADDOSE-3D assumes default angles of 90°.

2.8 NUMMONOMERS

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

NUMMONOMERS ***I*** specifies the number of monomers in the unit cell. Only integer numbers I should be used. This number should not be confused with the number of monomers in the asymmetric unit.

2.9 NUMRESIDUES

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

NUMRESIDUES ***I*** specifies the number of amino acid residues per monomer. Only integer numbers I should be used. Using this keyword the number and types of atoms are calculated according to the formula

$$\text{amino acid} = 5C + 1.35N + 1.5O + 8H$$

Sulfur atoms, e.g. from CYS and MET residues, should be added explicitly with the **PROTEINHEAVYATOMS** keyword.

The default value for I is 0.

2.10 NUMRNA

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

NUMRNA ***I*** specifies the number of RNA nucleotides per monomer. Only integer numbers I should be used. Using this keyword the number and types of atoms are calculated assuming an average nucleotide content defined as

$$\text{mean nucleotide} = 11.25H + 9.5C + 3.75N + 7O + 1P$$

If a more accurate estimate is required, individual atoms may be entered explicitly with the **PROTEINHEAVYATOMS** keyword.

The default value for I is 0.

2.11 NUMDNA

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

NUMDNA ***I***

specifies the number of DNA deoxynucleotides per monomer. Only integer numbers I should be used. Using this keyword the number and types of atoms are calculated assuming an average deoxynucleotide content defined as

$$\text{mean nucleotide} = 11.75H + 9.75C + 4N + 6O + 1P$$

If a more accurate estimate is required, individual atoms may be entered explicitly with the **PROTEINHEAVYATOMS** keyword.

The default value for I is 0.

2.12 PROTEINHEAVYATOMS

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

PROTEINHEAVYATOMS ***El I (El I (El I ..))*** defines a list of atoms to add to the protein part of the absorption. Each species is defined by a two character string *El* for the elemental symbol, and an integer number *I* of atoms of that species per monomer.

The command **PROTEINHEAVYATOMS S 10 Se 2** would add 10 sulfur and 2 selenium atoms per monomer.

2.13 SOLVENTHEAVYCONC

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

SOLVENTHEAVYCONC ***El I (El I (El I ..))*** defines the concentration of elements (not including water) in the solvent in millimoles per litre. Oxygen and lighter elements should not be specified.

The command **SOLVENTHEAVYCONC Na 1000 Cl 1000** specifies 1M sodium chloride in the solvent.

2.14 SOLVENTFRACTION

This keyword only has an effect when the absorption and attenuation coefficients are estimated using a legacy version of RADDOSE (see section 2.6).

SOLVENTFRACTION ***F***

The fraction of the unit cell that is occupied by solvent. If not given explicitly, this value is estimated from **NUMRESIDUES**, **NUMRNA** and **NUMDNA** using 1.35 g/ml for protein, and 2.0 g/ml for RNA and DNA.

3 Beam block

A `Beam` block must begin with the keyword **BEAM**. At least the **TYPE** must be specified. Depending on the chosen **TYPE**, further declarations may be required.

3.1 TYPE

With the keyword **TYPE**, the underlying beam implementation is chosen. Currently two distinct beam implementations exist:

TYPE TOPHAT defines a beam with uniform flux.

TYPE GAUSSIAN defines a beam with a 2-dimensional Gaussian flux profile. The full-width half-maximum must be specified with the **FWHM** keyword (see section 3.3).

3.2 FLUX

FLUX ***F*** specifies the total beam flux in photons per second. The flux parameter F can be specified in scientific notation (e.g. **1.3e12**).

3.3 FWHM

FWHM *X Y*

The FWHM of the beam (vertical), (horizontal). Not needed if a Top-Hat beam is used. This defines the X and Y FWHM of the beam respectively in the RADDOSE coordinate system.

3.4 ENERGY

ENERGY *F*
ENERGY *F* KEV

specifies the incident photon energy in keV. The optional keyword **KEV** can be appended for human readability of the input file.

3.5 COLLIMATION

COLLIMATION RECTANGULAR *X Y*

specifies the horizontal and vertical collimation of the beam. Delimits where the beam has non-zero intensity. This is defined by the slits. For an uncollimated Gaussian beam, set to $\approx 3\times$ FWHM.

4 Wedge block

A `Wedge` block must begin with the keyword **WEDGE**.

WEDGE *A B*

A and B define the start and end angle of the rotation in degrees (°). At 0° the front face of the crystal (X-Y plane) is normal to the beam. Rotation is right handed about the Y axis, as shown in figure 3).

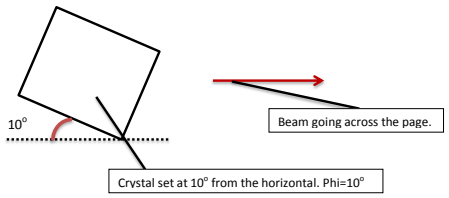


Figure 3: Schematic of angles for **WEDGE**. Figure courtesy of John Bremridge.

4.1 EXPOSURETIME

EXPOSURETIME *F*

specifies the total exposure time for this wedge in seconds.

4.2 ANGULARRESOLUTION

ANGULARRESOLUTION *F*

specifies the angular step size used for wedge iterations in degrees (°). Defaults to 2°.

4.3 STARTOFFSET

STARTOFFSET *X Y Z*

offset translation in µm applied to the crystal relative to the origin (defined as the intersection of the beam and the aligned goniometer axis) for the starting position of the wedge. Defaults to 0 0 0.

4.4 TRANSLATEPERDEGREE

TRANSLATEPERDEGREE *X Y Z*

translation of the goniometer during exposure in µm/° for helical scanning, leading to improvements in dose distribution. Defaults to 0 0 0.

4.5 ROTAXBEAMOFFSET

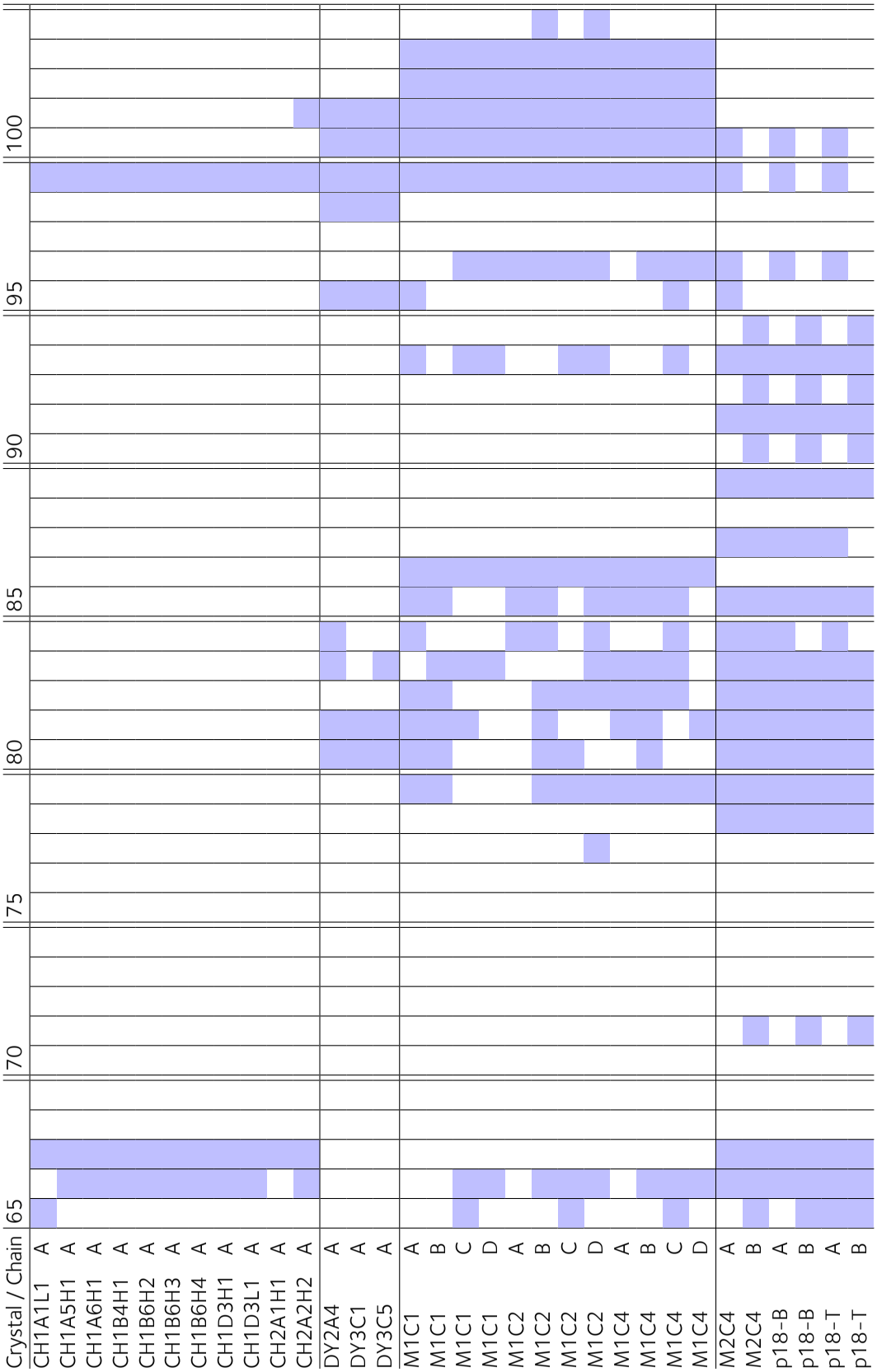
ROTAXBEAMOFFSET *F*

the offset in µm along X (vertical in most set-ups) between the beam axis and the rotation axis. Used to create 'offset' scanning for improvements in dose distribution. Defaults to 0 µm.

Appendix D RhoGDI crystal contacts

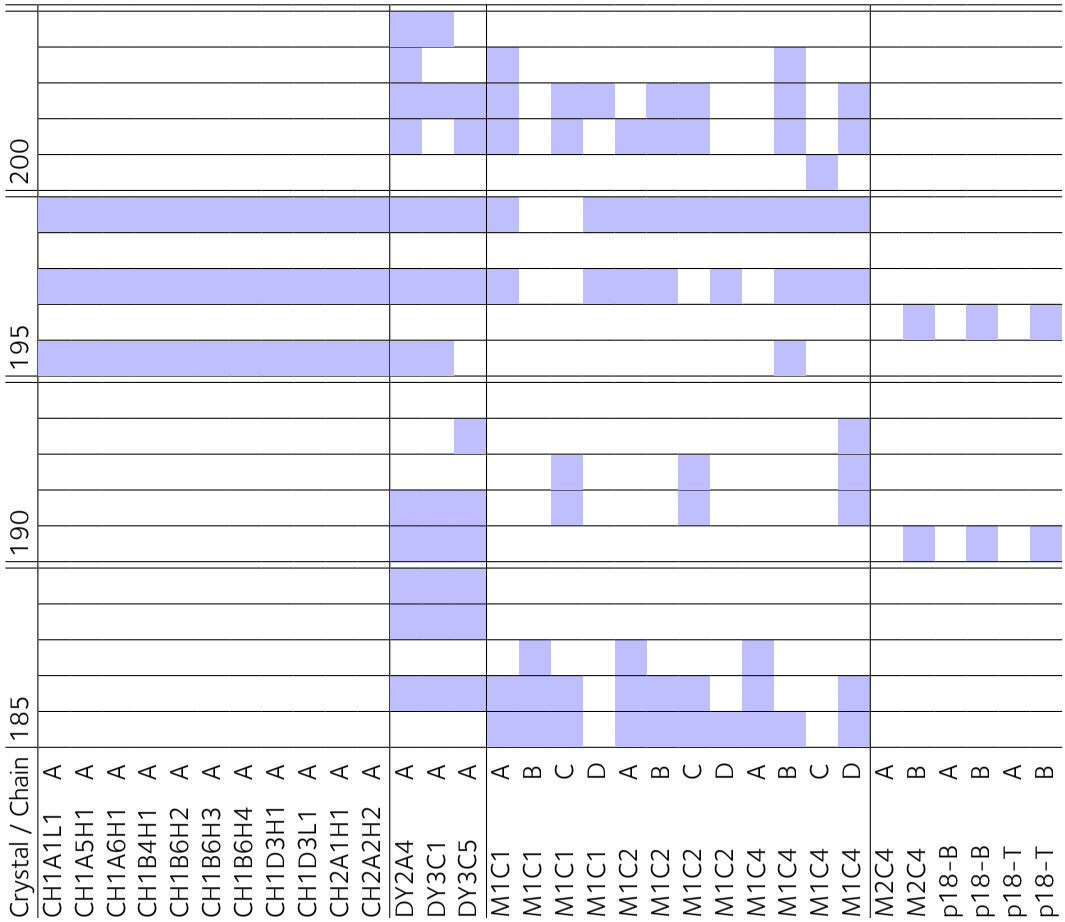
The crystal contacts were identified for each primary dataset using NCONT, which is part of the CCP4 software suite (Winn *et al.*, 2011). Residues that have one atom within 4 Å of another protein molecule atom are identified as crystal contact residue.

The next four pages contain an overview of all residues in all chains of all crystals. Crystal contact residues are marked in blue. For each crystal, the residues mutated by Surface Entropy Reduction (Table 4.3, page 153) are marked with a black dot (●).


Crystal / Chain
65
70
75
80
85
90
95
100
CH1A1L1 A
CH1A5H1 A
CH1A6H1 A
CH1B4H1 A
CH1B6H2 A
CH1B6H3 A
CH1B6H4 A
CH1D3H1 A
CH1D3L1 A
CH2A1H1 A
CH2A2H2 A
DY2A4 A
DY3C1 A
DY3C5 A
M1C1 A
M1C1 B
M1C1 C
M1C1 D
M1C2 A
M1C2 B
M1C2 C
M1C2 D
M1C4 A
M1C4 B
M1C4 C
M1C4 D
M2C4 A
M2C4 B
p18-B A
p18-B B
p18-T A
p18-T B


Crystal / Chain
105
110
115
120
125
130
135
140
CH1A1L1 A
CH1A5H1 A
CH1A6H1 A
CH1B4H1 A
CH1B6H2 A
CH1B6H3 A
CH1B6H4 A
CH1D3H1 A
CH1D3L1 A
CH2A1H1 A
CH2A2H2 A
DY2A4 A
DY3C1 A
DY3C5 A
M1C1 A
M1C1 B
M1C1 C
M1C1 D
M1C2 A
M1C2 B
M1C2 C
M1C2 D
M1C4 A
M1C4 B
M1C4 C
M1C4 D
M2C4 A
M2C4 B
p18-B A
p18-B B
p18-T A
p18-T B


Crystal / Chain
145
150
155
160
165
170
175
180
CH1A1L1 A
CH1A5H1 A
CH1A6H1 A
CH1B4H1 A
CH1B6H2 A
CH1B6H3 A
CH1B6H4 A
CH1D3H1 A
CH1D3L1 A
CH2A1H1 A
CH2A2H2 A
DY2A4 A
DY3C1 A
DY3C5 A
M1C1 A
M1C1 B
M1C1 C
M1C1 D
M1C2 A
M1C2 B
M1C2 C
M1C2 D
M1C4 A
M1C4 B
M1C4 C
M1C4 D
M2C4 A
M2C4 B
p18-B A
p18-B B
p18-T A
p18-T B


Crystal / Chain
185
190
195
200
CH1A1L1 A
CH1A5H1 A
CH1A6H1 A
CH1B4H1 A
CH1B6H2 A
CH1B6H3 A
CH1B6H4 A
CH1D3H1 A
CH1D3L1 A
CH2A1H1 A
CH2A2H2 A
DY2A4 A
DY3C1 A
DY3C5 A
M1C1 A
M1C1 B
M1C1 C
M1C1 D
M1C2 A
M1C2 B
M1C2 C
M1C2 D
M1C4 A
M1C4 B
M1C4 C
M1C4 D
M2C4 A
M2C4 B
p18-B A
p18-B B
p18-T A
p18-T B

Bibliography

ADLER DANIEL. *vioplot: Violin plot*, 2005. R package version 0.2.
http://wsopuppenkiste.wiso.uni-goettingen.de/~dadler.
referenced on page 142.

ALLAN ELIZABETH G., KANDER MELISSA C., CARMICHAEL IAN, AND GARMAN ELSPETH F. To scavenge or not to scavenge, that is STILL the question. *Journal of Synchrotron Radiation*, 20(1): 23–36, 2013.
doi: 10.1107/S0909049512046237.
referenced on page 10.

AMBE K. S., KUMTA U. S., AND TAPPEL A. L. Radiation Damage to Cytochrome c and Hemoglobin. *Radiation research*, 15(6):709–719, 1961.
doi: 10.2307/3571106.
referenced on page 26.

ARNDT U. W. Optimum X-ray wavelength for protein crystallography. *Journal of Applied Crystallography*, 17(2):118–119, 1984.
doi: 10.1107/S0021889884011092.
referenced on page 8.

AYEWAH NATHANIEL, HOVEMEYER DAVID, MORGENTHALER J. DAVID, PENIX JOHN, AND PUGH WILLIAM. Using Static Analysis to Find Bugs. *Software, IEEE*, 25(5):22–29, 2008.
doi: 10.1109/MS.2008.130.
referenced on page 114.

BARKER ADAM I., SOUTHWORTH-DAVIES ROBERT J., PAITHANKAR KARTHIK S., CARMICHAEL IAN, AND GARMAN ELSPETH F. Room-temperature scavengers for macromolecular crystallography: increased lifetimes and modified dose dependence of the intensity decay. *Journal of Synchrotron Radiation*, 16(2):205–216, 2009.
doi: 10.1107/S0909049509003343.
referenced on pages 10, 25, and 139.

BEITLICH THORSTEN, KÜHNEL KARIN, SCHULZE-BRIESE CLEMENS, SHOEMAN ROBERT L., AND SCHLICHTING ILME. Cryoradiolytic reduction of crystalline heme proteins: analysis by UV-Vis spectroscopy and X-ray crystallography. *Journal of Synchrotron Radiation*, 14(1):11–23, 2007.
doi: 10.1107/S0909049506049806.
referenced on page 86.

BERGLUND GUNNAR I., CARLSSON GUNILLA H., SMITH ANDREW T., SZÖKE HANNA, HENRIKSEN ANETTE, AND HAJDU JANOS. The catalytic pathway of horseradish peroxidase at high resolution. *Nature*, 417(6887):463–468, 2002.
doi: 10.1038/417463a.
referenced on page 189.

BERMAN HELEN, HENRICK KIM, AND NAKAMURA HARUKI. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980–980, 2003.
doi: 10.1038/nsb1203-980.
referenced on pages 32, 36, 132, and 138.

Beust Cédric and Suleiman Hani. *Next Generation Java Testing*. Addison-Wesley Professional, first edition, 2007.
ISBN 978-0321503107.
referenced on page 109.

Blake C.C.F. and Phillips D.C. Effects of X-irradiation on single crystals of myoglobin. *Proceedings of the Symposium on the Biological Effects of Ionizing Radiation at the Molecular Level*, pages 183–191, 1962.
referenced on pages 5 and 82.

Borshchevskiy Valentin I., Round Ekaterina S., Popov Alexandr N., Büldt Georg, and Gordeliy Valentin I. X-ray-radiation-induced changes in bacteriorhodopsin structure. *Journal of molecular biology*, 409(5):813–825, 2011.
doi: 10.1016/j.jmb.2011.04.038.
referenced on page 31.

Bourenkov Gleb P. and Popov Alexander N. Optimization of data collection taking radiation damage into account. *Acta Crystallographica Section D*, 66(4):409–419, 2010.
doi: 10.1107/S0907444909054961.
referenced on pages 83 and 86.

Bowler Matthew W., Guijarro Matias, Petitdemange Sebastien, Baker Isabel, Svensson Olof, Burghammer Manfred, Mueller-Dieckmann Christoph, Gordon Elspeth J., Flot David, McSweeney Sean M., and Leonard Gordon A. Diffraction cartography: applying microbeams to macromolecular crystallography sample evaluation and data collection. *Acta Crystallographica Section D*, 66(8):855–864, 2010.
doi: 10.1107/S0907444910019591.
referenced on page 85.

Bragg William Henry. The Reflection of X-rays by Crystals. (II.). *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 89 (610):246–248, 1913a, http://www.jstor.org/stable/93487.
referenced on page 4.

Bragg William Lawrence. The diffraction of short electromagnetic waves by a crystal. In *Proceedings of the Cambridge Philosophical Society*, volume 17, pages 43–57, 1913b.
referenced on page 3.

Brändén Carl-Ivar and Jones T. Alwyn. Between objectivity and subjectivity. *Nature*, 343 (6260):687–689, 1990.
doi: 10.1038/343687a0.
referenced on page 145.

Bricogne Gérard. [23] Bayesian statistical viewpoint on structure determination: Basic concepts and examples. In Charles W. Carter Jr., editor, *Macromolecular Crystallography Part A*, volume 276 of *Methods in Enzymology*, pages 361–423. Academic Press, 1997.
doi: 10.1016/S0076-6879(97)76069-5.
referenced on page 132.

Brooks-Bartlett Jonathan, Zeldin Oliver B., and Garman Elspeth F. Exploring relationships between diffracted intensity decay and a new dose metric. *Acta Crystallographica Section A*, 69(a1):s410, 2013.
doi: 10.1107/S0108767313096426.
referenced on page 120.

Brown Morton B. and Forsythe Alan B. Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.
doi: 10.1080/01621459.1974.10482955.
referenced on pages 58, 68, and 73.

Brünger Axel T. [19] Free *R* value: Cross-validation in crystallography. In Charles W. Carter Jr. Robert M. Sweet, editor, *Macromolecular Crystallography Part B*, volume 277 of *Methods in Enzymology*, pages 366–396. Academic Press, 1997.
doi: 10.1016/S0076-6879(97)77021-6.
referenced on pages 135 and 158.

Bulheller Benjamin M. and Hirst Jonathan D. Dichrocalc – circular and linear dichroism online. *Bioinformatics*, 25(4):539–540, 2009.
doi: 10.1093/bioinformatics/btp016.
referenced on pages 49 and 66.

Burmeister Wilhelm P. Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Crystallographica Section D*, 56(3):328–341, 2000.
doi: 10.1107/S0907444999016261.
referenced on pages 15, 18, 26, and 31.

Canfield Robert E. and Liu Anne K. The Disulfide Bonds of Egg White Lysozyme (Muramidase). *Journal of Biological Chemistry*, 240(5):1997–2002, 1965.
referenced on pages 136 and 138.

Carugo Kristina Djinović, Helliwell John R., Stuhrmann Heinrich, and Weiss Manfred S. Softer and soft X-rays in macromolecular crystallography. *Journal of Synchrotron Radiation*, 12(4):410–419, 2005.
doi: 10.1107/S0909049504025762.
referenced on page 8.

Carugo Oliviero and Argos Patrick. Reliability of atomic displacement parameters in protein crystal structures. *Acta Crystallographica Section D*, 55(2):473–478, 1999.
doi: 10.1107/S0907444998011688.
referenced on page 66.

Chapman Henry N., Fromme Petra, Barty Anton, White Thomas A., Kirian Richard A., Aquila Andrew, Hunter Mark S., Schulz Joachim, DePonte Daniel P., Weierstall Uwe R., et al. Femtosecond X-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.
doi: 10.1038/nature09750.
referenced on page 86.

Choi Young Ah. The use of free radical scavengers in Macromolecular Crystallography. Part II Project report, Molecular and Cellular Biochemistry, University of Oxford, April 2013.
referenced on pages 129, 136, 139, 142, 143, 144, and 145.

Compton Arthur H. A Quantum Theory of the Scattering of X-rays by Light Elements. *Phys. Rev.*, 21:483–502, 1923.
doi: 10.1103/PhysRev.21.483.
referenced on page 5.

Cooper David R., Boczek Tomasz, Grelewska Katarzyna, Pinkowska Malgorzata, Sikorska Malgorzata, Zawadzki Michal, and Derewenda Zygmunt. Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallographica Section D*, 63(5):636–645, 2007.
doi: 10.1107/S0907444907010931.
referenced on pages 50, 53, 151, 152, and 153.

Copeland Tom. *PMD Applied*. Centennial Books, November 2005.
ISBN 0-9762214-1-1.
referenced on page 112.

Cowan John A. and Nave Colin. The optimum conditions to collect X-ray data from very small samples. *Journal of Synchrotron Radiation*, 15(5):458–462, 2008.
doi: 10.1107/S0909049508014623.
referenced on page 12.

Crick F. H. C. and Magdoff Beatrice S. The Theory of the Method of Isomorphous Replacement for Protein Crystals. I. *Acta Crystallographica*, 9(11):901–908, 1956.
doi: 10.1107/S0365110X56002552.
referenced on pages 15 and 28.

Dauter Zbigniew and Jaskolski Mariusz. How to read (and understand) Volume A of *International Tables for Crystallography*: an introduction for nonspecialists. *Journal of Applied Crystallography*, 43(5 Part 2):1150–1171, 2010.
doi: 10.1107/S0021889810026956.
referenced on page 34.

Mora Eugenio, Carmichael Ian, and Garman Elspeth F. Effective scavenging at cryotemperatures: further increasing the dose tolerance of protein crystals. *Journal of Synchrotron Radiation*, 18(3):346–357, 2011.
doi: 10.1107/S0909049511007163.
referenced on pages 10, 25, and 31.

Sanctis Daniele and Nanao Max H. Segmenting data sets for RIP. *Acta Crystallographica Section D*, 68(9):1152–1162, 2012.
doi: 10.1107/S0907444912023475.
referenced on page 191.

Derewenda Zygmunt S. It's all in the crystals... *Acta Crystallographica Section D*, 67(4): 243–248, 2011.
doi: 10.1107/S0907444911007797.
referenced on pages 129 and 150.

Einstein Albert. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, 322(6):132–148, 1905.
doi: 10.1002/andp.19053220607.
referenced on page 5.

ELLIS PAUL J., COHEN AINA E., AND SOLTIS S. MICHAEL. Beamstop with integrated X-ray sensor. *Journal of Synchrotron Radiation*, 10(3):287–288, 2003.
doi: 10.1107/S0909049503003285.
referenced on page 166.

EMSLEY P., LOHKAMP B., SCOTT W. G., AND COWTAN K. Features and development of *Coot*. *Acta Crystallographica Section D*, 66(4):486–501, 2010.
doi: 10.1107/S0907444910007493.
referenced on pages 52, 53, 141, 146, and 158.

ENNIFAR E., CARPENTIER P., FERRER J.-L., WALTER P., AND DUMAS P. X-ray-induced debromination of nucleic acids at the Br*K* absorption edge and implications for MAD phasing. *Acta Crystallographica Section D*, 58(8):1262–1268, 2002.
doi: 10.1107/S0907444902009526.
referenced on page 26.

EVANS PHILIP. Scaling and assessment of data quality. *Acta Crystallographica Section D*, 62(1): 72–82, 2006.
doi: 10.1107/S0907444905036693.
referenced on page 157.

EVANS PHILIP R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D*, 67(4):282–292, 2011.
doi: 10.1107/S090744491003982X.
referenced on pages 168 and 197.

EVANS PHILIP R. AND MURSHUDOV GARIB N. How good are my data and what is the resolution? *Acta Crystallographica Section D*, 69(7):1204–1214, 2013.
doi: 10.1107/S0907444913000061.
referenced on page 157.

FIORAVANTI EMANUELA, VELLIEUX FRÉDÉRIC M. D., AMARA PATRICIA, MADERN DOMINIQUE, AND WEIK MARTIN. Specific radiation damage to acidic residues and its relation to their chemical and structural environment. *Journal of Synchrotron Radiation*, 14(1):84–91, 2007.
doi: 10.1107/S0909049506038623.
referenced on pages 26, 29, 30, 31, and 76.

FISCHER HANNES, POLIKARPOV IGOR, AND CRAIEVICH ALDO F. Average protein density is a molecular-weight-dependent function. *Protein Science*, 13(10):2825–2828, 2004. ISSN 1469-896X.
doi: 10.1110/ps.04688204.
referenced on page 101.

FLOT DAVID, MAIRS TREVOR, GIRAUD THIERRY, GUIJARRO MATIAS, LESOURD MARC, REY VICENTE, BRUSSEL DENIS, MORAWE CHRISTIAN, BOREL CHRISTINE, HIGNETTE OLIVIER, CHAVANNE JOEL, NURIZZO DIDIER, MCSWEENEY SEAN, AND MITCHELL EDWARD. The ID23-2 structural biology microfocus beamline at the ESRF. *Journal of Synchrotron Radiation*, 17(1):107–118, 2010.
doi: 10.1107/S0909049509041168.
referenced on pages 12, 13, 88, 89, and 102.

Foadi James, Aller Pierre, Alguel Yilmaz, Cameron Alex, Axford Danny, Owen Robin L., Armour Wes, Waterman David G., Iwata So, and Evans Gwyndaf. Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. *Acta Crystallographica Section D*, 69(8):1617–1632, 2013.
doi: 10.1107/S0907444913012274.
referenced on page 190.

Fourme Roger, Girard Eric, Dhaussy Anne-Claire, Medjoubi Kadda, Prangé Thierry, Ascone Isabella, Mezouar Mohamed, and Kahn Richard. A new paradigm for macromolecular crystallography beamlines derived from high-pressure methodology and results. *Journal of Synchrotron Radiation*, 18(1):31–36, 2011.
doi: 10.1107/S0909049510041695.
referenced on page 11.

Fourme Roger, Honkimäki Veijo, Girard Eric, Medjoubi Kadda, Dhaussy Anne-Claire, and Kahn Richard. Reduction of radiation damage and other benefits of short wavelengths for macromolecular crystallography data collection. *Journal of Applied Crystallography*, 45(4):652–661, 2012.
doi: 10.1107/S0021889812019164.
referenced on pages 8 and 11.

Frankaer Christian Grundahl, Mossin Susanne, Ståhl Kenny, and Harris Pernille. Towards accurate structural characterization of metal centres in protein crystals: the structures of Ni and Cu T_6 bovine insulin derivatives. *Acta Crystallographica Section D*, 70(1):110–122, 2014.
doi: 10.1107/S1399004713029040.
referenced on page 31.

Frazão Carlos, Sieker Larry, Sheldrick George, Lamzin Victor, LeGall Jean, and Carrondo Maria A. *Ab initio* structure solution of a dimeric cytochrome c_3 from *Desulfovibrio gigas* containing disulfide bridges. *JBIC Journal of Biological Inorganic Chemistry*, 4(2):162–165, 1999.
doi: 10.1007/s007750050299.
referenced on page 131.

Frishman Dmitrij and Argos Patrick. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995. ISSN 1097-0134.
doi: 10.1002/prot.340230412.
referenced on pages 43, 66, and 68.

Fuentes-Montero Luis, Parkhurst James, Winter Graeme, Waterman David, Gildea Richard, Brewster Aaron, Hattne Johan, Sauter Nicholas, and Evans Gwyndaf. DIALS – a toolbox for diffraction data analysis. *Acta Crystallographica Section A*, 70(a1):C1440, 2014.
doi: 10.1107/S2053273314085593.
referenced on page 207.

Gabanyi Margaret J., Adams Paul D., Arnold Konstantin, Bordoli Lorenza, Carter Lester G., Flippen-Andersen Judith, Gifford Lida, Haas Juergen, Kouranov Andrei, McLaughlin William A., Micallef David I., Minor Wladek, Shah Raship, Schwede Torsten, Tao Yi-Ping, Westbrook John D., Zimmerman Matthew, and Berman Helen M. The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *Journal of Structural and Functional Genomics*, 12(2):45–54, 2011.
doi: 10.1007/s10969-011-9106-2.
referenced on page 33.

Galton F. Cutting a Round Cake on Scientific Principles. *Nature*, 75:173, 1906.
doi: 10.1038/075173c0.
referenced on page 89.

Gamma Erich, Helm Richard, Johnson Ralph, and Vlissides John. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1st edition, 1994.
ISBN 978-0201633610.
referenced on pages 95 and 102.

Garman Elspeth F. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallographica Section D*, 66(4):339–351, 2010.
doi: 10.1107/S0907444910008656.
referenced on pages 7, 37, and 69.

Garman Elspeth F. and Schneider Thomas R. Macromolecular Cryocrystallography. *Journal of Applied Crystallography*, 30(3):211–237, 1997.
doi: 10.1107/S0021889897002677.
referenced on page 9.

Genick Ulrich K., Soltis S. Michael, Kuhn Peter, Canestrelli Ilona L., and Getzoff Elizabeth D. Structure at 0.85 Å resolution of an early protein photocycle intermediate. *Nature*, 392(6672):206–209, 1998.
doi: 10.1038/32462.
referenced on page 27.

Gerstel Markus, Olekhnovich Natalya, Brooks-Bartlett Jonathan, Derewenda Zygmunt S., Deane Charlotte M., and Garman Elspeth F. Quantitative radiation damage studies in macromolecular X-ray crystallography. *Acta Crystallographica Section A*, 69(a1): s407–s408, 2013.
doi: 10.1107/S010876731309644X.
referenced on pages 25 and 129.

Gerstel Markus, Deane Charlotte, and Garman Elspeth. BDamage: Quantifying radiation damage in MX structures. *Acta Crystallographica Section A*, 70(a1):C317, 2014.
doi: 10.1107/S205327331409682X.
referenced on page 25.

Gerstel Markus, Deane Charlotte M., and Garman Elspeth F. Identifying and quantifying radiation damage at the atomic level. *Journal of Synchrotron Radiation*, 22(2):201–212, 2015.
doi: 10.1107/S1600577515002131.
referenced on pages 25 and 47.

Glazer A. M. The first paper by W.L. Bragg - what and when? *Crystallography Reviews*, 19(3): 117–124, 2013.
doi: 10.1080/0889311X.2013.813494.
referenced on page 4.

Gumiero Andrea, Metcalfe Clive L., Pearson Arwen R., Raven Emma Lloyd, and Moody Peter C. E. Nature of the Ferryl Heme in Compounds I and II. *Journal of Biological Chemistry*, 286(2):1260–1268, 2011.
doi: 10.1074/jbc.M110.183483.
referenced on page 86.

Haas Juergen, Roth Steven, Arnold Konstantin, Kiefer Florian, Schmidt Tobias, Bordoli Lorenza, and Schwede Torsten. The Protein Model Portal – a comprehensive resource for protein structure and model information. *Database*, 2013.
doi: 10.1093/database/bat031.
referenced on page 33.

Hakim Aaron, Nguyen Jennifer B., Basu Koli, Zhu Darren F., Thakral Durga, Davies Peter L., Isaacs Farren J., Modis Yorgo, and Meng Wuyi. Crystal Structure of an Insect Antifreeze Protein and Its Implications for Ice Binding. *Journal of Biological Chemistry*, 288 (17):12295–12304, 2013.
doi: 10.1074/jbc.M113.450973.
referenced on page 131.

Halle Bertil. Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences*, 99(3):1274–1279, 2002.
doi: 10.1073/pnas.032522499.
referenced on page 39.

Henderson Richard. Cryo-Protection of Protein Crystals against Radiation Damage in Electron and X-Ray Diffraction. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 241(1300):6–8, 1990.
doi: 10.1098/rspb.1990.0057.
referenced on page 82.

Hintze Jerry L. and Nelson Ray D. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, 52(2):181–184, 1998.
doi: 10.1080/00031305.1998.10480559.
referenced on page 142.

Holton James M. A beginner's guide to radiation damage. *Journal of Synchrotron Radiation*, 16 (2):133–142, 2009.
doi: 10.1107/S0909049509004361.
referenced on pages 9, 11, 30, and 83.

Holton James M. and Frankel Kenneth A. The minimum crystal size needed for a complete diffraction data set. *Acta Crystallographica Section D*, 66(4):393–408, 2010.
doi: 10.1107/S0907444910007262.
referenced on pages 85 and 101.

HOMER CHRISTINA, COOPER LAURA, AND GONZALEZ ANA. Energy dependence of site-specific radiation damage in protein crystals. *Journal of Synchrotron Radiation*, 18(3):338–345, 2011. doi: 10.1107/S0909049511005504.
referenced on pages 11, 26, 29, and 76.

HORTON PETER, COLES SIMON, PITAK MATEUSZ, TIZZARD GRAHAM, AND WILSON CLAIRE. Radiation Damage in Chemical Crystallography. *Acta Crystallographica Section A*, 70(a1):C1699, 2014.
doi: 10.1107/S2053273314083004.
referenced on page 125.

JOHANSSON RENZO, TORRENTS EDUARD, LUNDIN DANIEL, SPRENGER JANINA, SAHLIN MARGARETA, SJÖBERG BRITT-MARIE, AND LOGAN DEREK T. High-resolution crystal structures of the flavoprotein NrdI in oxidized and reduced states – an unusual flavodoxin. *FEBS Journal*, 277 (20):4265–4277, 2010. ISSN 1742-4658.
doi: 10.1111/j.1742-4658.2010.07815.x.
referenced on page 86.

JONES GEORGE D. D., LEA JEREMY S., SYMONS MARTYN C. R., AND TAIWO FATAI A. Structure and mobility of electron gain and loss centres in proteins. *Nature*, 330(6150):772–773, 1987.
doi: 10.1038/330772a0.
referenced on page 27.

JOOSTEN ROBBIE P., LONG FEI, MURSHUDOV GARIB N., AND PERRAKIS ANASTASSIS. The *PDB_REDO* server for macromolecular structure model optimization. *IUCrJ*, 1(4):213–220, 2014.
doi: 10.1107/S2052252514009324.
referenced on page 33.

JUERS DOUGLAS H. AND WEIK MARTIN. Similarities and differences in radiation damage at 100K *versus* 160K in a crystal of thermolysin. *Journal of Synchrotron Radiation*, 18(3):329–337, 2011.
doi: 10.1107/S0909049511007631.
referenced on pages 10, 29, 31, and 61.

KABSCH WOLFGANG. *XDS*. *Acta Crystallographica Section D*, 66(2):125–132, 2010.
doi: 10.1107/S0907444909047337.
referenced on page 157.

KARPLUS P. ANDREW AND DIEDERICHS KAY. Linking crystallographic model and data quality. *Science*, 336(6084):1030–1033, 2012.
doi: 10.1126/science.1218231.
referenced on page 197.

KEEP NICHOLAS H, BARNES MARIA, BARSUKOV IGOR, BADII RAMIN, LIAN LU-YUN, SEGAL ANTHONY W, MOODY PETER CE, AND ROBERTS GORDON CK. A modulator of rho family G proteins, rhoGDI, binds these G proteins via an immunoglobulin-like domain and a flexible N-terminal arm. *Structure*, 5(5):623–633, 1997.
doi: 10.1016/S0969-2126(97)00218-9.
referenced on pages 150 and 152.

Kekilli Demet, Dworkowski Florian S. N., Pompidor Guillaume, Fuchs Martin R., Andrew Colin R., Antonyuk Svetlana, Strange Richard W., Eady Robert R., Hasnain S. Samar, and Hough Michael A. Fingerprinting redox and ligand states in haemprotein crystal structures using resonance Raman spectroscopy. *Acta Crystallographica Section D*, 70 (5):1289–1296, 2014.
doi: 10.1107/S1399004714004039.
referenced on page 91.

Kern Jan, Alonso-Mori Roberto, Hellmich Julia, Tran Rosalie, Hattne Johan, Laksmono Hartawan, Glöckner Carina, Echols Nathaniel, Sierra Raymond G., Sellberg Jonas, Lassalle-Kaiser Benedikt, Gildea Richard J., Glatzel Pieter, Grosse-Kunstleve Ralf W., Latimer Matthew J., McQueen Trevor A., DiFiore Dörte, Fry Alan R., Messerschmidt Marc, Miahnahri Alan, Schafer Donald W., Seibert M. Marvin, Sokaras Dimosthenis, Weng Tsu-Chien, Zwart Petrus H., White William E., Adams Paul D., Bogan Michael J., Boutet Sébastien, Williams Garth J., Messinger Johannes, Sauter Nicholas K., Zouni Athina, Bergmann Uwe, Yano Junko, and Yachandra Vittal K. Room temperature femtosecond X-ray diffraction of photosystem II microcrystals. *Proceedings of the National Academy of Sciences*, 109(25):9721–9726, 2012.
doi: 10.1073/pnas.1204598109.
referenced on page 86.

Kern Jan, Alonso-Mori Roberto, Tran Rosalie, Hattne Johan, Gildea Richard J., Echols Nathaniel, Glöckner Carina, Hellmich Julia, Laksmono Hartawan, Sierra Raymond G., Lassalle-Kaiser Benedikt, Koroidov Sergey, Lampe Alyssa, Han Guangye, Gul Sheraz, DiFiore Dörte, Milathianaki Despina, Fry Alan R., Miahnahri Alan, Schafer Donald W., Messerschmidt Marc, Seibert M. Marvin, Koglin Jason E., Sokaras Dimosthenis, Weng Tsu-Chien, Sellberg Jonas, Latimer Matthew J., Grosse-Kunstleve Ralf W., Zwart Petrus H., White William E., Glatzel Pieter, Adams Paul D., Bogan Michael J., Williams Garth J., Boutet Sébastien, Messinger Johannes, Zouni Athina, Sauter Nicholas K., Yachandra Vittal K., Bergmann Uwe, and Yano Junko. Simultaneous Femtosecond X-ray Spectroscopy and Diffraction of Photosystem II at Room Temperature. *Science*, 340(6131):491–495, 2013.
doi: 10.1126/science.1234273.
referenced on page 14.

Kleywegt Gerard J., Harris Mark R., Zou Jin-yu, Taylor Thomas C., Wählby Anders, and Jones T. Alwyn. The Uppsala Electron-Density Server. *Acta Crystallographica Section D*, 60 (12 Part 1):2240–2249, 2004.
doi: 10.1107/S0907444904013253.
referenced on page 132.

Kmetko Jan, Husseini Naji S., Naides Matthew, Kalinin Yevgeniy, and Thorne Robert E. Quantifying X-ray radiation damage in protein crystals at cryogenic temperatures. *Acta Crystallographica Section D*, 62(9):1030–1038, 2006.
doi: 10.1107/S0907444906023869.
referenced on page 37.

KMETKO JAN, WARKENTIN MATTHEW, ENGLICH ULRICH, AND THORNE ROBERT E. Can radiation damage to protein crystals be reduced using small-molecule compounds? *Acta Crystallographica Section D*, 67(10):881–893, 2011.
doi: 10.1107/S0907444911032835.
referenced on pages 10 and 25.

KNUTH DONALD E. Structured programming with `go to` statements. *ACM Comput. Surv.*, 6(4): 261–301, 1974. ISSN 0360-0300.
doi: 10.1145/356635.356640.
referenced on page 114.

KORT REMCO, HELLINGWERF KLAAS J., AND RAVELLI RAIMOND B. G. Initial events in the photocycle of photoactive yellow protein. *Journal of Biological Chemistry*, 279(25):26417–26424, 2004.
doi: 10.1074/jbc.M311961200.
referenced on page 27.

KUZAY TUNCER M., KAZMIERCZAK MICHAEL, AND HSIEH B. J. X-ray beam/biomaterial thermal interactions in third-generation synchrotron sources. *Acta Crystallographica Section D*, 57(1): 69–81, 2001.
doi: 10.1107/S0907444900013299.
referenced on page 87.

LEBEDEV ANDREY A. AND ISUPOV MICHAIL N. Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program *Zanuda*. *Acta Crystallographica Section D*, 70(9):2430–2443, 2014.
doi: 10.1107/S1399004714014795.
referenced on page 157.

LEIROS HANNA-KIRSTI S., TIMMINS JOANNA, RAVELLI RAIMOND B. G., AND MCSWEENEY SEÁN M. Is radiation damage dependent on the dose rate used during macromolecular crystallography data collection? *Acta Crystallographica Section D*, 62(2):125–132, 2006.
doi: 10.1107/S0907444905033627.
referenced on page 84.

LESLIE ANDREW G. W. AND POWELL HAROLD R. Processing diffraction data with mosflm. In Read Randy J. and Sussman Joel L., editors, *Evolving Methods for Macromolecular Crystallography*, volume 245 of *NATO Science Series*, pages 41–51. Springer Netherlands, 2007.
doi: 10.1007/978-1-4020-6316-9_4.
referenced on page 157.

LIU WEI, WACKER DANIEL, GATI CORNELIUS, HAN GYE WON, JAMES DANIEL, WANG DINGJIE, NELSON GARRETT, WEIERSTALL UWE, KATRITCH VSEVOLOD, BARTY ANTON, ZATSEPIN NADIA A., LI DIANFAN, MESSERSCHMIDT MARC, BOUTET SÉBASTIEN, WILLIAMS GARTH J., KOGLIN JASON E., SEIBERT M. MARVIN, WANG CHONG, SHAH SYED T. A., BASU SHIBOM, FROMME RAIMUND, KUPITZ CHRISTOPHER, RENDEK KIMBERLEY N., GROTJOHANN INGO, FROMME PETRA, KIRIAN RICHARD A., BEYERLEIN KENNETH R., WHITE THOMAS A., CHAPMAN HENRY N., CAFFREY MARTIN, SPENCE JOHN C. H., STEVENS RAYMOND C., AND CHEREZOV VADIM. Serial Femtosecond Crystallography of G Protein–Coupled Receptors. *Science*, 342(6165):1521–1524, 2013.
doi: 10.1126/science.1244142.
referenced on page 14.

Longenecker Kenton L., Garrard Sarah M., Sheffield Peter J., and Derewenda Zygmunt S. Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. *Acta Crystallographica Section D*, 57(5):679–688, 2001.
doi: 10.1107/S0907444901003122.
referenced on pages 129, 150, and 151.

Louridas Panagiotis. Static code analysis. *Software, IEEE*, 23(4):58–61, 2006.
doi: 10.1109/MS.2006.114.
referenced on page 109.

Markley John L., Ulrich Eldon L., Berman Helen M., Henrick Kim, Nakamura Haruki, and Akutsu Hideo. BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *Journal of Biomolecular NMR*, 40(3):153–155, 2008.
doi: 10.1007/s10858-008-9221-y.
referenced on page 132.

Mateja Agnieszka, Devedjiev Yancho, Krowarsch Daniel, Longenecker Kenton, Dauter Zbigniew, Otlewski Jacek, and Derewenda Zygmunt S. The impact of Glu→Ala and Glu→Asp mutations on the crystallization properties of RhoGDI: the structure of RhoGDI at 1.3Å resolution. *Acta Crystallographica Section D*, 58(12):1983–1991, 2002.
doi: 10.1107/S090744490201394X.
referenced on page 151.

Matsui Yasuhiro, Sakai Keisuke, Murakami Midori, Shiro Yoshitsugu, Adachi Shin-ichi, Okumura Hideo, and Kouyama Tsutomu. Specific damage induced by X-ray radiation and structural changes in the primary photoreaction of bacteriorhodopsin. *Journal of molecular biology*, 324(3):469–481, 2002.
doi: 10.1016/S0022-2836(02)01110-5.
referenced on page 27.

Matthews B.W. Solvent content of protein crystals. *Journal of Molecular Biology*, 33(2):491 – 497, 1968. ISSN 0022-2836.
doi: 10.1016/0022-2836(68)90205-2.
referenced on page 157.

Mayer Jürgen, Khairy Khaled, and Howard Jonathon. Drawing an elephant with four complex parameters. *American Journal of Physics*, 78(6):648–649, 2010.
doi: 10.1119/1.3254017.
referenced on page 135.

McCoy Airlie J., Grosse-Kunstleve Ralf W., Adams Paul D., Winn Martyn D., Storoni Laurent C., and Read Randy J. *Phaser* crystallographic software. *Journal of Applied Crystallography*, 40(4):658–674, 2007.
doi: 10.1107/S0021889807021206.
referenced on page 157.

Meents Alke, Gutmann Sascha, Wagner Armin, and Schulze-Briese Clemens. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proceedings of the National Academy of Sciences*, 107(3):1094–1099, 2010.
doi: 10.1073/pnas.0905481107.
referenced on pages 10 and 29.

Mizuguchi Kenji, Deane Charlotte M., Blundell Tom L., Johnson Mark S., and Overington John P. Joy: protein sequence-structure representation and analysis. *Bioinformatics*, 14(7):617–623, 1998.
doi: 10.1093/bioinformatics/14.7.617.
referenced on pages 61, 63, and 66.

Moon P. B. Developments in Gamma-Ray Optics. *Nature*, 185:427–429, 1960.
doi: 10.1038/185427a0.
referenced on page 5.

Murray James and Garman Elspeth. Investigation of possible free-radical scavengers and metrics for radiation damage in protein cryocrystallography. *Journal of Synchrotron Radiation*, 9(6):347–354, 2002.
doi: 10.1107/S0909049502014632.
referenced on pages 15, 37, 83, and 195.

Murray James W., Garman Elspeth F., and Ravelli Raimond B. G. X-ray absorption by macromolecular crystals: the effects of wavelength and crystal composition on absorbed dose. *Journal of Applied Crystallography*, 37(4):513–522, 2004.
doi: 10.1107/S0021889804010660.
referenced on pages 8, 20, 86, and 99.

Murray James W., Rudiño-Piñera Enrique, Owen Robin L., Grininger Martin, Ravelli Raimond B. G., and Garman Elspeth F. Parameters affecting the X-ray dose absorbed by macromolecular crystals. *Journal of Synchrotron Radiation*, 12(3):268–275, 2005.
doi: 10.1107/S0909049505003262.
referenced on pages 9, 83, and 85.

Murshudov Garib N., Skubák Pavol, Lebedev Andrey A., Pannu Navraj S., Steiner Roberto A., Nicholls Robert A., Winn Martyn D., Long Fei, and Vagin Alexei A. *REFMAC*5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D*, 67(4):355–367, 2011.
doi: 10.1107/S0907444911001314.
referenced on pages 131, 158, 168, and 169.

Nanao Max H., Sheldrick George M., and Ravelli Raimond B. G. Improving radiation-damage substructures for RIP. *Acta Crystallographica Section D*, 61(9):1227–1237, 2005.
doi: 10.1107/S0907444905019360.
referenced on pages xi, 8, 31, 51, 54, 55, 56, 57, 59, 60, 61, 62, 63, 64, 70, 72, 74, 75, and 76.

Nave Colin and Garman Elspeth F. Towards an understanding of radiation damage in cryocooled macromolecular crystals. *Journal of Synchrotron Radiation*, 12(3):257–260, 2005.
doi: 10.1107/S0909049505007132.
referenced on pages 10 and 25.

Nave Colin and Hill Mark A. Will reduced radiation damage occur with very small crystals? *Journal of Synchrotron Radiation*, 12(3):299–303, 2005.
doi: 10.1107/S0909049505003274.
referenced on pages 12 and 88.

Neutze Richard, Wouts Remco, Spoel David, Weckert Edgar, and Hajdu Janos. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, 406(6797):752–757, 2000.
doi: 10.1038/35021099.
referenced on pages 5 and 14.

Nishikawa Ken and Ooi Tatsuo. Prediction of the surface-interior diagram of globular proteins by an empirical method. *International Journal of Peptide and Protein Research*, 16(1):19–32, 1980. ISSN 1399-3011.
doi: 10.1111/j.1399-3011.1980.tb02931.x.
referenced on page 39.

Nishikawa Ken and Ooi Tatsuo. Radial locations of amino acid residues in a globular protein: Correlation with the sequence. *Journal of Biochemistry*, 100(4):1043–1047, 1986.
referenced on page 39.

Novak J., Krajnc A., and Žontar R. Taxonomy of static code analysis tools. In *MIPRO, 2010 Proceedings of the 33rd International Convention*, pages 418–422, May 2010.
ISBN 978-1-4244-7763-0.
referenced on page 109.

Nowak Elzbieta, Brzuszkiewicz Anna, Dauter Miroslawa, Dauter Zbigniew, and Rosenbaum Gerd. To scavenge or not to scavenge: that is the question. *Acta Crystallographica Section D*, 65(9):1004–1006, 2009.
doi: 10.1107/S0907444909026821.
referenced on page 11.

O'Neill Peter, Stevens David L., and Garman Elspeth F. Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective. *Journal of Synchrotron Radiation*, 9(6):329–332, 2002.
doi: 10.1107/S0909049502014553.
referenced on page 5.

Owen Robin L., Rudiño-Piñera Enrique, and Garman Elspeth F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proceedings of the National Academy of Sciences of the United States of America*, 103(13):4912–4917, 2006.
doi: 10.1073/pnas.0600973103.
referenced on pages 37, 82, and 85.

Owen Robin L., Holton James M., Schulze-Briese Clemens, and Garman Elspeth F. Determination of X-ray flux using silicon pin diodes. *Journal of Synchrotron Radiation*, 16(2):143–151, 2009.
doi: 10.1107/S0909049508040429.
referenced on pages 85 and 163.

Owen Robin L., Axford Danny, Nettleship Joanne E., Owens Raymond J., Robinson James I., Morgan Ann W., Doré Andrew S., Lebon Guillaume, Tate Christopher G., Fry Elizabeth E., Ren Jingshan, Stuart David I., and Evans Gwyndaf. Outrunning free radicals in room-temperature macromolecular crystallography. *Acta Crystallographica Section D*, 68(7):810–818, 2012.
doi: 10.1107/S0907444912012553.
referenced on pages 9, 14, and 85.

Owen Robin L., Paterson Neil, Axford Danny, Aishima Jun, Schulze-Briese Clemens, Ren Jingshan, Fry Elizabeth E., Stuart David I., and Evans Gwyndaf. Exploiting fast detectors to enter a new dimension in room-temperature crystallography. *Acta Crystallographica Section D*, 70(5):1248–1256, 2014.
doi: 10.1107/S1399004714005379.
referenced on page 14.

Painter Jay and Merritt Ethan A. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallographica Section D*, 62(4):439–450, 2006.
doi: 10.1107/S0907444906005270.
referenced on page 34.

Paithankar Karthik S. and Garman Elspeth F. Know your dose: *RADDOSE*. *Acta Crystallographica Section D*, 66(4):381–388, 2010.
doi: 10.1107/S0907444910006724.
referenced on pages 20, 86, and 87.

Paithankar Karthik S., Owen Robin L., and Garman Elspeth F. Absorbed dose calculations for macromolecular crystals: improvements to *RADDOSE*. *Journal of Synchrotron Radiation*, 16(2):152–162, 2009.
doi: 10.1107/S0909049508040430.
referenced on pages 8, 20, 86, 87, 99, and 139.

Parr Terence. *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf, 2nd revised edition, 2013.
ISBN 978-1934356999.
referenced on page 95.

Petrova Tatiana, Ginell Stephan, Mitschler Andre, Kim Youngchang, Lunin Vladimir Y., Joachimiak Grazyna, Cousido-Siah Alexandra, Hazemann Isabelle, Podjarny Alberto, Lazarski Krzysztof, and Joachimiak Andrzej. X-ray-induced deterioration of disulfide bridges at atomic resolution. *Acta Crystallographica Section D*, 66(10):1075–1091, 2010.
doi: 10.1107/S0907444910033986.
referenced on pages 17, 18, 29, 31, and 75.

Pintar Alessandro, Carugo Oliviero, and Pongor Sándor. CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics*, 18(7):980–984, 2002.
doi: 10.1093/bioinformatics/18.7.980.
referenced on pages 39 and 41.

Pozharski Edwin. On the variability of experimental data in macromolecular crystallography. *Acta Crystallographica Section D*, 68(9):1077–1087, 2012.
doi: 10.1107/S0907444912020100.
referenced on pages 11, 13, and 194.

Pozharski Edwin, Weichenberger Christian X., and Rupp Bernhard. Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallographica Section D*, 69(2):150–167, 2013.
doi: 10.1107/S0907444912044423.
referenced on page 37.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. http://www.R-project.org/.
referenced on pages 47, 57, 69, and 93.

Rack A., Weitkamp T., Riotte M., Grigoriev D., Rack T., Helfen L., Baumbach T., Dietsch R., Holz T., Krämer M., Siewert F., Meduna M., Cloetens P., and Ziegler E. Comparative study of multilayers used in monochromators for synchrotron-based coherent hard X-ray imaging. *Journal of Synchrotron Radiation*, 17(4):496–510, 2010.
doi: 10.1107/S0909049510011623.
referenced on page 84.

Ramagopal Udupi A., Dauter Zbigniew, Thirumuruhan Radhakannan, Fedorov Elena, and Almo Steven C. Radiation-induced site-specific damage of mercury derivatives: phasing and implications. *Acta Crystallographica Section D*, 61(9):1289–1298, 2005.
doi: 10.1107/S0907444905022316.
referenced on page 26.

Ravelli Raimond B. G. and Garman Elspeth F. Radiation damage in macromolecular cryocrystallography. *Current Opinion in Structural Biology*, 16(5):624–629, 2006.
doi: 10.1016/j.sbi.2006.08.001.
referenced on pages 15 and 190.

Ravelli Raimond B. G. and McSweeney Sean M. The 'fingerprint' that X-rays can leave on structures. *Structure*, 8(3):315–328, 2000.
doi: 10.1016/S0969-2126(00)00109-X.
referenced on pages 15, 26, 30, 138, 139, and 144.

Ravelli Raimond B. G., Theveneau Pascal, McSweeney Sean, and Caffrey Martin. Unit-cell volume change as a metric of radiation damage in crystals of macromolecules. *Journal of Synchrotron Radiation*, 9(6):355–360, 2002.
doi: 10.1107/S0909049502014541.
referenced on pages 15, 37, and 195.

Ravelli Raimond B. G., Leiros Hanna-Kirsti Schrøder, Pan Baocheng, Caffrey Martin, and McSweeney Sean. Specific Radiation Damage Can Be Used to Solve Macromolecular Crystal Structures. *Structure*, 11(2):217–224, 2003.
doi: 10.1016/S0969-2126(03)00006-6.
referenced on pages 28, 54, and 131.

Read Randy J., Adams Paul D., Arendall W. Bryan III, Brunger Axel T., Emsley Paul, Joosten Robbie P., Kleywegt Gerard J., Krissinel Eugene B., Lütteke Thomas, Otwinowski Zbyszek, et al. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure*, 19(10):1395–1412, 2011.
doi: 10.1016/j.str.2011.08.006.
referenced on page 77.

Richards Frederic M. The interpretation of protein structures: Total volume, group volume distributions and packing density. *Journal of Molecular Biology*, 82(1):1–14, 1974. ISSN 0022-2836.
doi: 10.1016/0022-2836(74)90570-1.
referenced on page 41.

RINGWALD A. Pair production from vacuum at the focus of an X-ray free electron laser. *Physics Letters B*, 510(1–4):107–116, 2001.
doi: 10.1016/S0370-2693(01)00496-8.
referenced on page 5.

SALDIN E. L., SCHNEIDMILLER E. A., AND YURKOV M. V. Statistical and coherence properties of radiation from x-ray free-electron lasers. *New Journal of Physics*, 12(3):035010, 2010.
doi: 10.1088/1367-2630/12/3/035010.
referenced on page 84.

SANISHVILI RUSLAN, YODER DEREK W., POTHINENI SUDHIR BABU, ROSENBAUM GERD, XU SHENGLAN, VOGT STEFAN, STEPANOV SERGEY, MAKAROV OLEG A., CORCORAN STEPHEN, BENN RICHARD, NAGARAJAN VENUGOPALAN, SMITH JANET L., AND FISCHETTI ROBERT F. Radiation damage in protein crystals is reduced with a micron-sized X-ray beam. *Proceedings of the National Academy of Sciences*, 108(15):6127–6132, 2011.
doi: 10.1073/pnas.1017701108.
referenced on pages 5, 12, 85, 88, 121, and 155.

SAUTER NICHOLAS K., GROSSE-KUNSTLEVE RALF W., AND ADAMS PAUL D. Robust indexing for automatic data collection. *Journal of Applied Crystallography*, 37(3):399–409, 2004.
doi: 10.1107/S0021889804005874.
referenced on page 157.

SCHMIDT BRYAN, HO LORRAINE, AND HOGG PHILIP J. Allosteric disulfide bonds. *Biochemistry*, 45 (24):7429–7433, 2006.
doi: 10.1021/bi0603064.
referenced on pages 70 and 71.

SCHOMAKER V. AND TRUEBLOOD K. N. On the rigid-body motion of molecules in crystals. *Acta Crystallographica Section B*, 24(1):63–76, 1968.
doi: 10.1107/S0567740868001718.
referenced on page 34.

SCHRÖDINGER, LLC . The PyMOL molecular graphics system, version 1.3r1. August 2010.
referenced on page 152.

SCHULZE-BRIESE C., WAGNER A., TOMIZAKI T., AND OETIKER M. Beam-size effects in radiation damage in insulin and thaumatin crystals. *Journal of Synchrotron Radiation*, 12(3):261–267, 2005.
doi: 10.1107/S0909049505003298.
referenced on page 12.

SHIMAZU F., KUMTA U. S., AND TAPPEL A. L. Radiation Damage to Methionine and Its Derivatives. *Radiation research*, 22(2):276–287, 1964.
doi: 10.2307/3571659.
referenced on page 26.

SHIMIZU NOBUTAKA, HIRATA KUNIO, HASEGAWA KAZUYA, UENO GO, AND YAMAMOTO MASAKI. Dose dependence of radiation damage for protein crystals studied at various X-ray energies. *Journal of Synchrotron Radiation*, 14(1):4–10, 2007.
doi: 10.1107/S0909049506049296.
referenced on pages 11 and 26.

Sliz Piotr, Harrison Stephen C., and Rosenbaum Gerd. How does Radiation Damage in Protein Crystals Depend on X-Ray Dose? *Structure*, 11(1):13–19, 2003.
doi: 10.1016/S0969-2126(02)00910-3.
referenced on page 13.

Southworth-Davies Robert J., Medina Melissa A., Carmichael Ian, and Garman Elspeth F. Observation of Decreased Radiation Damage at Higher Dose Rates in Room Temperature Protein Crystallography. *Structure*, 15(12):1531–1541, 2007.
doi: 10.1016/j.str.2007.10.013.
referenced on page 14.

Sutton Kristin A., Black Paul J., Mercer Kermit R., Garman Elspeth F., Owen Robin L., Snell Edward H., and Bernhard William 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 Crystallographica Section D*, 69(12):2381–2394, 2013.
doi: 10.1107/S0907444913022117.
referenced on pages 18, 26, 31, 138, 139, 143, and 144.

Taylor Garry L. Introduction to phasing. *Acta Crystallographica Section D*, 66(4):325–338, 2010.
doi: 10.1107/S0907444910006694.
referenced on pages 131 and 168.

Trofimov A. A., Polyakov K. M., Lazarenko V. A., Popov A. N., Tikhonova T. V., Tikhonov A. V., and Popov V. O. PDB structure entries: 4L3X, 4L3Y, 4L3Z. Investigation of the X-ray-induced nitrite reduction catalysed by cytochrome c nitrite reductase from the bacterium Thioalkalivibrio nitratireducens. 2013.
referenced on page 31.

Turing Alan M. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(2):230–265, 1936.
referenced on page 110.

Turner Isaac, Deane Charlotte M., and Garman Elspeth F. Specific X-ray Radiation Damage to Disulphide Bonds. DTC project report, University of Oxford, Jul 2010.
referenced on page 32.

Upper Dennis. The unsuccessful self-treatment of a case of "writer's block". *Journal of Applied Behavior Analysis*, 7(3):497, 1974.
doi: 10.1901/jaba.1974.7-497a.

Wang Guoli and Dunbrack Roland L. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.
doi: 10.1093/bioinformatics/btg224.
referenced on page 66.

Warkentin Matthew, Badeau Ryan, Hopkins Jesse, and Thorne Robert E. Dark progression reveals slow timescales for radiation damage between $T = 180$ and 240K. *Acta Crystallographica Section D*, 67(9):792–803, 2011.
doi: 10.1107/S0907444911027600.
referenced on page 120.

Warkentin Matthew, Badeau Ryan, Hopkins Jesse B., Mulichak Anne M., Keefe Lisa J., and Thorne Robert E. Global radiation damage at 300 and 260K with dose rates approaching 1 MGy s^{-1}. *Acta Crystallographica Section D*, 68(2):124–133, 2012a.
doi: 10.1107/S0907444911052085.
referenced on page 14.

Warkentin Matthew, Badeau Ryan, Hopkins Jesse B., and Thorne Robert E. Spatial distribution of radiation damage to crystalline proteins at 25–300K. *Acta Crystallographica Section D*, 68(9):1108–1117, 2012b.
doi: 10.1107/S0907444912021361.
referenced on pages 10, 29, and 76.

Weckert E. and Hümmer K. Multiple-Beam X-ray Diffraction for Physical Determination of Reflection Phases and its Applications. *Acta Crystallographica Section A*, 53(2):108–143, 1997.
doi: 10.1107/S0108767396011117.
referenced on page 130.

Weckert Edgar, Hölzer Kerstin, Schroer Klaus, Zellner Johannes, and Hümmer Kurt. Feasibility study for the measurement of a large set of triplet phases from a small protein. *Acta Crystallographica Section D*, 55(7):1320–1328, 1999.
doi: 10.1107/S0907444999005338.
referenced on page 130.

Weik Martin, Ravelli Raimond B. G., Kryger Gitay, McSweeney Sean, Raves Maria L., Harel Michal, Gros Piet, Silman Israel, Kroon Jan, and Sussman Joel L. Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proceedings of the National Academy of Sciences*, 97(2):623–628, 2000.
doi: 10.1073/pnas.97.2.623.
referenced on pages 15, 26, 31, 72, 129, 136, 138, 139, 143, 144, 145, 147, and 149.

Weik Martin, Bergès Jacqueline, Raves Maria L., Gros Piet, McSweeney Sean, Silman Israel, Sussman Joel L., Houée-Levin Chantal, and Ravelli Raimond B. G. Evidence for the formation of disulfide radicals in protein crystals upon X-ray irradiation. *Journal of Synchrotron Radiation*, 9(6):342–346, 2002.
doi: 10.1107/S0909049502014589.
referenced on pages 15, 18, 26, 145, and 146.

Weiss Manfred S. On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures. *Acta Crystallographica Section D*, 63(12):1235–1242, 2007.
doi: 10.1107/S0907444907052146.
referenced on pages 38, 39, and 66.

Weiss Manfred S., Panjikar Santosh, Mueller-Dieckmann Christoph, and Tucker Paul A. On the influence of the incident photon energy on the radiation damage in crystalline biological samples. *Journal of Synchrotron Radiation*, 12(3):304–309, 2005.
doi: 10.1107/S0909049505003328.
referenced on page 11.

White Jim F., Noinaj Nicholas, Shibata Yoko, Love James, Kloss Brian, Xu Feng, Gvozdenovic-Jeremic Jelena, Shah Priyanka, Shiloach Joseph, Tate Christopher G., and Grisshammer Reinhard. Structure of the agonist-bound neurotensin receptor. *Nature*, 490(7421):508–513, 2012.
doi: 10.1038/nature11558.
referenced on page 86.

Winn Martyn D., Ballard Charles C., Cowtan Kevin D., Dodson Eleanor J., Emsley Paul, Evans Phil R., Keegan Ronan M., Krissinel Eugene B., Leslie Andrew G. W., McCoy Airlie, McNicholas Stuart J., Murshudov Garib N., Pannu Navraj S., Potterton Elizabeth A., Powell Harold R., Read Randy J., Vagin Alexei, and Wilson Keith S. Overview of the *CCP*4 suite and current developments. *Acta Crystallographica Section D*, 67(4):235–242, 2011.
doi: 10.1107/S0907444910045749.
referenced on pages 47, 49, 52, 66, 139, 140, 145, 157, 159, and 223.

Winter Graeme. *xia2*: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography*, 43(1), 2010.
doi: 10.1107/S0021889809045701.
referenced on page 157.

Winter Graeme, Lobley Carina M. C., and Prince Stephen M. Decision making in *xia*2. *Acta Crystallographica Section D*, 69(7):1260–1273, 2013.
doi: 10.1107/S0907444913015308.
referenced on page 163.

Yano Junko, Kern Jan, Irrgang Klaus-Dieter, Latimer Matthew J., Bergmann Uwe, Glatzel Pieter, Pushkar Yulia, Biesiadka Jacek, Loll Bernhard, Sauer Kenneth, Messinger Johannes, Zouni Athina, and Yachandra Vittal K. X-ray damage to the mn4ca complex in single crystals of photosystem ii: A case study for metalloprotein crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, 102(34): 12047–12052, 2005.
doi: 10.1073/pnas.0505207102.
referenced on page 26.

Yogavel Manickam, Tripathi Timir, Gupta Ankita, Banday Mudassir Meraj, Rahlfs Stefan, Becker Katja, Belrhali Hassan, and Sharma Amit. Atomic resolution crystal structure of glutaredoxin 1 from *Plasmodium falciparum* and comparison with other glutaredoxins. *Acta Crystallographica Section D*, 70(1):91–100, 2014.
doi: 10.1107/S1399004713025285.
referenced on page 91.

Yorke Briony A., Beddard Godfrey S., Owen Robin L., and Pearson Arwen R. Time-resolved crystallography using the Hadamard transform. *Nature Methods*, 2014.
doi: 10.1038/nmeth.3139.
referenced on page 190.

Zeldin Oliver B., Brockhauser Sandor, Bremridge John, Holton James M., and Garman Elspeth F. Predicting the X-ray lifetime of protein crystals. *Proceedings of the National Academy of Sciences*, 110(51):20551–20556, 2013a.
doi: 10.1073/pnas.1315879110.
referenced on pages 13, 84, 85, 89, 91, 104, 105, 119, 120, and 165.

Zeldin Oliver B., Gerstel Markus, and Garman Elspeth F. RADDOSE-3D: time- and space-resolved modelling of dose in macromolecular crystallography. *Journal of Applied Crystallography*, 46(4):1225–1230, 2013b.
doi: 10.1107/S0021889813011461.
referenced on pages 13, 99, and 165.

Zeldin Oliver B., Gerstel Markus, and Garman Elspeth F. Optimizing the spatial distribution of dose in X-ray macromolecular crystallography. *Journal of Synchrotron Radiation*, 20(1): 49–57, 2013c.
doi: 10.1107/S0909049512044706.
referenced on pages 13, 85, 88, 89, 104, 105, and 125.

Zhai Yujia, Zhang Kai, Huo Yanwu, Zhu Yanshi, Zhou Qiangjun, Lu Jiuwei, Black Isobel, Pang Xiaoyun, Roszak Aleksander W., Zhang Xujia, Isaacs Neil W., and Sun Fei. Autotransporter passenger domain secretion requires a hydrophobic cavity at the extracellular entrance of the β-domain pore. *Biochemical Journal*, 435(3):577–587, 2011.
doi: 10.1042/BJ20101548.
referenced on page 86.

Zhang Zepu, Sauter Nicholas K., Bedem Henry, Snell Gyorgy, and Deacon Ashley M. Automated diffraction image analysis and spot searching for high-throughput crystal screening. *Journal of Applied Crystallography*, 39(1):112–119, 2006.
doi: 10.1107/S0021889805040677.
referenced on page 157.