CHAPTER 6

Methods to Assess Radiation Damage in SAXS

6.1 Introduction

As outlined in the introduction of this thesis, SAXS is a method to obtain a molecular envelope of a macromolecule in solution. However, during X-ray exposure the sample undergoes radiation induced changes which progress throughout the experiment. Ultimately these changes lead to aggregation of the molecules in the sample which is observed as a progressive dissimilarity of subsequent 1D scattering curves (frames).

This chapter presents a quantitative study of the possible mitigation of radiation damage by addition of various radioprotectant compounds in SAXS experiments. The compounds were chosen for several reasons:

- they have shown good protection ability in MX (**?**) and (**?**).
- they are commonly used already as radioprotectants in SAXS (**?**).
- the majority of their atomic constituents have relatively small atomic numbers and hence they will not significantly contribute to the dose absorbed by the sample.

6.2 Extending RADDOSE-3D for SAXS

To perform comparative analysis of the extent of radiation damage in SAXS experiments it is necessary to calculate the dose absorbed by the sample. Currently dose estimates in SAXS are calculated using

$$D = \frac{tfE(1-T)}{V\rho}, \tag{6.2.1}$$

where D is the dose, t is the exposure time, f is the X-ray flux, E is the energy of the incident photons, T is the transmission factor, V is the illuminated volume and ρ is the mass density of the sample (**??**). To simplify the calculation of the dose, the explicit geometry of the sample is not taken into account. A more accurate calculation can be performed if the SAXS experiment is simulated in three dimensions, providing a spatially resolved dose field. This dose field can then be interpreted with the various dose metrics already developed for MX (**??**).

This section presents the extensions written into RADDOSE-3D to perform simulations of SAXS experiments for improved dose calculations.

6.2.1 RADDOSE-3D architecture

At its core, RADDOSE-3D takes a description of a crystal and exposes it to a computational beam model via a user specified exposure strategy (wedge) (**?**). The crystal is computationally modelled as a 3D voxel grid, where each grid element stores information about x, y, z coordinates, the dose and the fluence. The dose is calculated using the beam intensity at the leading edge of the voxel and the proportion of the beam that is absorbed by the voxel, which is governed by the absorption coefficient. The beam is described by the intensity profile, photon energy and the total flux. Finally, the wedge specifies the total angular rotation of the sample, exposure time and any rotation offsets or translations. The structure of the program is illustrated in Figure 6.1. A single RADDOSE-3D job performs the MX simulation on a single crystal, but it can be exposed to multiple beams with multiple wedges. Output files giving information about the crystal state are updated after the crystal is exposed to a beam via a single wedge.

At an abstract level, the crystal is defined by several properties that are typically associated

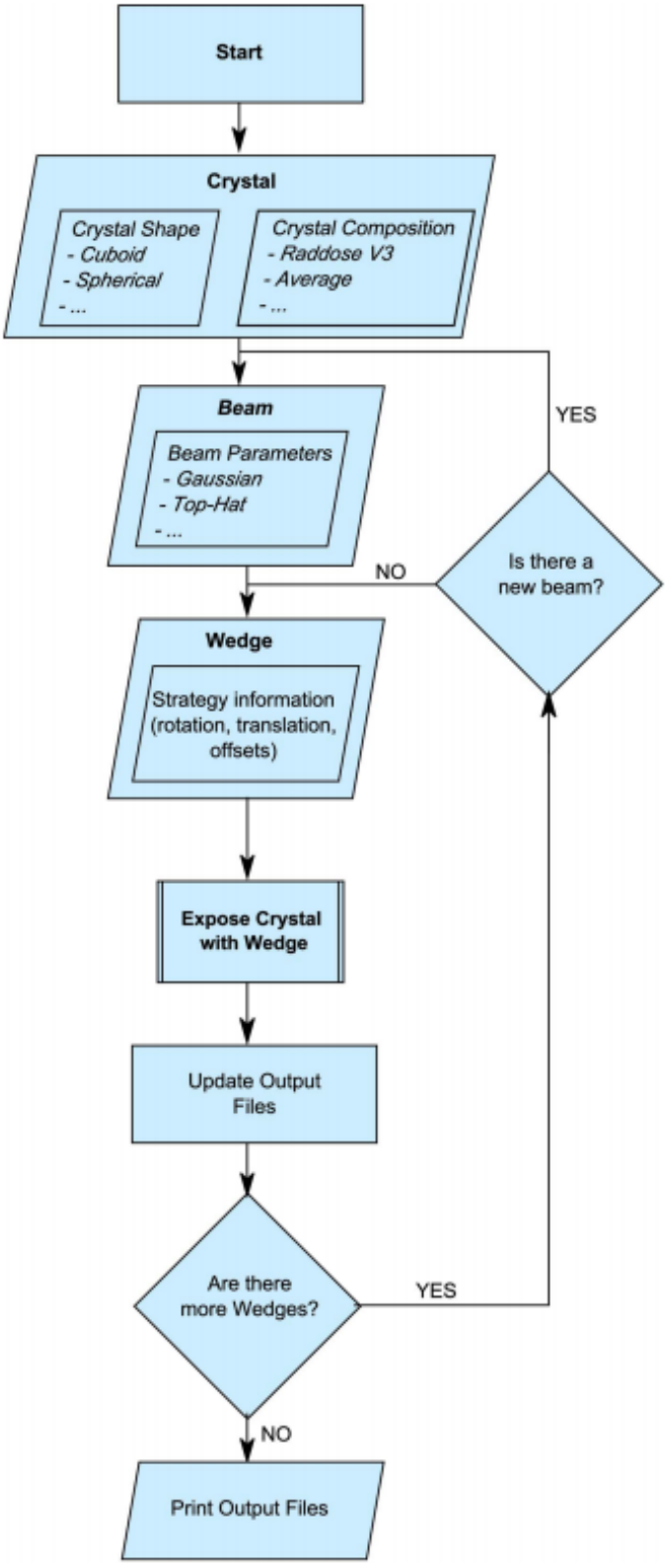


Figure 6.1: Flow chart illustrating the structure of the RADDOSE-3D code. Reproduced from (**?**).

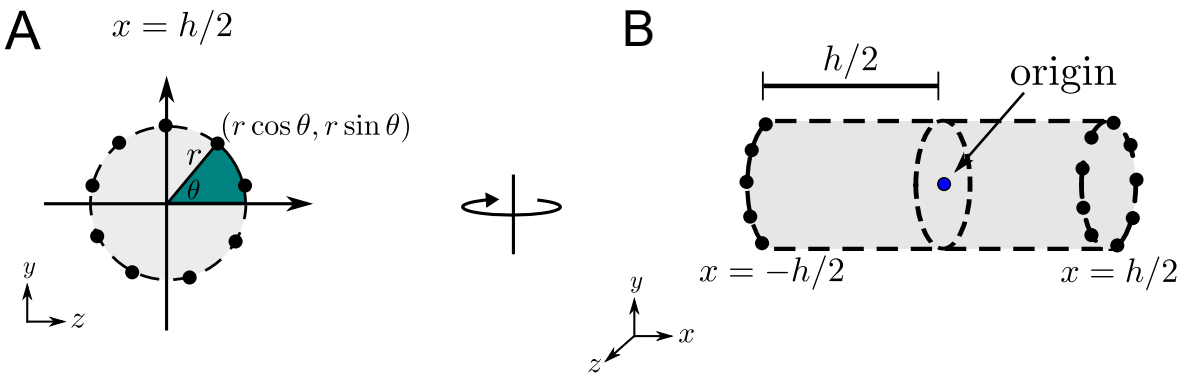


Figure 6.2: Implementation of the cylindrical sample geometry in RADDOSE-3D given user defined diameter, d, and height, h. (A) evenly spaced points around a circle are generated given the radius $r = d/2$ of the circular cross-section. RADDOSE-3D defaults to 32 points. (B) In three dimensions the points represent the circles at each end of the cylinder at a distance of $h/2$ from the origin located at the centre of the cylinder. The connectivity of these vertices is hard-coded into the RADDOSE-3D source code.

with crystals in MX. These include the unit cell volume and the number of molecules in the unit cell. RADDOSE-3D uses this information to determine the composition of the crystal so that an absorption coefficient can be calculated. However, there is no reason that the composition has to be defined by information specific to the crystal, in which case it is more appropriate to refer to what has been termed the "crystal" so far as the "sample". If the sample is a liquid, as is the case in SAXS experiments, then the sample composition should be determined with different inputs. The modular design of RADDOSE-3D enables this type of extension into be easily incorporated to the existing functionality.

6.2.2 Cylindrical sample geometry

In a typical SAXS experiment, liquid samples are generally contained in, or flowed through, a cylindrical capillary during the X-ray exposure. Therefore it is necessary for RADDOSE-3D to be able to model cylindrical sample shapes. RADDOSE-3D had already been extended to handle polygonal shapes (**?**), which meant that it was already capable of modelling cylindrical shapes, since any 3D shape can be modelled by a series of polygons, However, the implementation for polyhedral crystals requires the user to define the sample geometry in a non-user friendly manner. A file specifying the geometry of the shape using a set of vertex positions, and their connectivity is required. This is much more complex than defining the diameter of the circular cross section and the length/height of a cylinder. Therefore a module was written that accepts a diameter and height as input and converts this into a polyhedral description (a set of vertices and faces) of a cylinder within RADDOSE-3D.

The cylindrical implementation, which specifies the geometry of the sample alone, not the capillary in which it is contained, is graphically depicted in Figure 6.2 (the effect of the capillary is dealt with separately in section 6.2.4). First the points around a circle are generated using the user defined diameter of the circular cross section. RADDOSE-3D uses 32 points around the circle by default, no matter what dimensions are specified. The points are evenly spaced around the circle with y, z coordinates $(r\cos\theta, r\sin\theta)$. The angle (in radians) between any two consecutive points is $2\pi/32$. A cylinder can be defined by the circles at either end of the shape so this is done using the final coordinate x. Depending on which end a particular point is, it will have coordinates $(x, y, z) = (-h/2, r\cos\theta, r\sin\theta)$ or $(x, y, z) = (h/2, r\cos\theta, r\sin\theta)$. Note that this assumes the origin of the system is located at the centre of the cylinder.

Regardless of the dimensions of the cylinder, the connectivity of the vertices remains the same because the number of vertices and their orientation with respect to one another is constant. Therefore the connectivity has been hard-coded into RADDOSE-3D. It will only ever need to be changed if the number of points is altered. However, this parameter is not exposed to the user and hence it would only change if a developer modified the source code.

The geometry defined to create the cylinder is rotated 90° about the y-axis compared to the usual RADDOSE-3D simulation geometry. This means that the beam would irradiate the sample along the x-axis (or directly into the page looking at Figure 6.2 A). In a typical SAXS experiment the beam direction is along the z-axis defined in Figure 6.2 (i.e. perpendicular to the axial dimension of the cylinder). So whenever a user specifies a SAXS experiment in RADDOSE-3D, the sample is rotated by an additional 90° on the angle which the user specifies as the initial orientation of the sample to the beam (the sample geometry does not have to be specified as cylindrical).

6.2.3 Determining the sample composition

Knowledge of the atomic composition of the sample is necessary to be able to calculate the dose absorbed upon X-ray irradiation. This is because the overall absorption coefficient of

the sample, μ_{abs}, is calculated from the individual atomic absorption coefficients, σ_j as

$$\mu_{\text{abs}} = (1/V_c) \sum_{j=1}^{N} \sigma_j, \tag{6.2.2}$$

where, V_c is the volume of the unit cell, N is the number of atoms in the unit cell and $\sigma_j = \sigma_j^{\text{Thompson}} + \sigma_j^{\text{Compton}} + \sigma_j^{\text{Photoelectric}}$ (**?**). (The previous RADDOSE versions assumed that the absorption coefficient was equal to the attenuation coefficient, $\mu_{\text{att}} i.e. \mu_{\text{abs}} = \mu_{\text{att}}$, so equation 1 in Murray *et al.* (2004) , which is analogous to equation 6.2.2 here, writes 'μ_{att}' instead of 'μ_{abs}'). In MX the crystal composition is calculated from the contents of the unit cell. In SAXS the samples are liquids as opposed to crystals, and hence the notion of a unit cell does not apply. Thus instead, the approach to determine the atomic composition of the sample is to define a volume of liquid and estimate the contents given its protein concentration and buffer composition.

First the molarity of the solution is calculated using the formula

$$\text{Molarity (moles/litre)} = \frac{\text{sample concentration (grams/litre)}}{\text{molecular mass (grams/mole)}}. \tag{6.2.3}$$

The sample concentration is provided by the user in units of grams per litre ($\equiv$ mg/ml). The molecular mass of the molecule is calculated from other parameters provided in the user input. If the sequence file is given for the protein (the sample can also contain DNA and RNA) then the molecular mass can be determined accurately by summing the molecular mass of each residue in the file. Otherwise an average molecular weight is used for each residue (110 Da for protein residues, 339.5 Da for RNA residues and 327 Da for DNA residues) and the user has to specify the type and number of residues for the sample.

Secondly, a suitable volume needs to be specified to calculate the atomic composition. A suitable volume is one that is large enough to contain at least one complete molecule. By default this volume is defined to be (1000 Å)^3 but this can be changed by specifying the length, width and height dimensions of the volume in the input file using the *unit cell* input keyword.

The number of monomers/molecules in the volume can now be calculated by multiplying the molarity, volume (converted to litres) and Avogadro's number (N_A = 6.022 $\times$ 10^{23} mol^{-1}), which is then rounded to the nearest integer. The absorption coefficient can then be com-

puted in the usual way as is described in Paithankar *et al.* (2009). If there is less than 1 molecule in the volume then this is flagged up and the user is advised to increase the volume.

6.2.4 Attenuation of X-ray flux due to capillary

In a typical MX experiment, a crystal is exposed directly to the X-ray beam. In contrast, samples from SAXS experiments are held inside a quartz capillary. This means that the attenuation of the X-ray flux due to the capillary needs to be taken into account before calculating the dose absorbed by the sample.

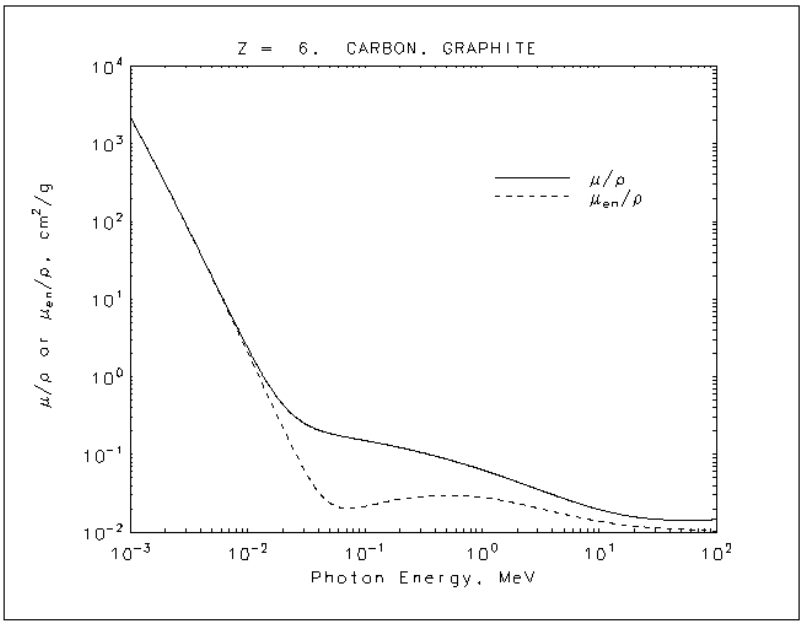
The transmission fraction of an X-ray beam due to a material with mass thickness x and density ρ is given by

$$I/I_0 = \exp\left[-(\mu/\rho)x\right], \tag{6.2.4}$$

where I is the emergent intensity of the beam after penetrating the material, I_0 is the incident intensity and μ/ρ is defined as the mass attenuation coefficient (**?**). The mass thickness, x, is defined as the mass per unit area and is given by $x = \rho t$ where t is the thickness of the material. The attenuation fraction by the capillary can hence be calculated as $1 - I/I_0$.

RADDOSE-3D requires the user to supply the thickness, density and material composition of the capillary to calculate the attenuation fraction. The mass thickness can be directly calculated using the density and thickness as described above. The mass attenuation coefficient is dependent on the atomic composition of the material as well as the energy of the incident photons. The relevant values can be found online via the National Institute of Science and Technology (NIST) tables (**??**). A section of the webpage for the mass attenuation coefficient table of carbon is shown in Figure 6.3. The tabulated energy values do not explicitly include the typical energies used in crystallograpy and SAXS (around 12 keV), so the mass attenuation coefficient is linearly interpolated between the closest values. Mass attenuation coefficient data for various mixtures including air, borosilicate glass, water, bone and soft tissue are also tabulated in NIST table 4 (**?**). These mixtures can be used directly by RADDOSE-3D.

When mixtures of atomic species are used (the most commonly used capillary material in SAXS is quartz which has elemental composition SiO_2) the mass attenuation coefficient is



Carbon, Graphite
***Z* = 6**

HTML table format

Energy (MeV)	μ/ρ (cm^2/g)	μ_{en}/ρ (cm^2/g)
1.00000E-03	2.211E+03	2.209E+03
1.50000E-03	7.002E+02	6.990E+02
2.00000E-03	3.026E+02	3.016E+02
3.00000E-03	9.033E+01	8.963E+01
4.00000E-03	3.778E+01	3.723E+01
5.00000E-03	1.912E+01	1.866E+01
6.00000E-03	1.095E+01	1.054E+01
8.00000E-03	4.576E+00	4.242E+00
1.00000E-02	2.373E+00	2.078E+00

Carbon, Graphite
***Z* = 6**

ASCII format

Energy (MeV)	μ/ρ (cm^2/g)	μ_{en}/ρ (cm^2/g)
1.00000E-03	2.211E+03	2.209E+03
1.50000E-03	7.002E+02	6.990E+02
2.00000E-03	3.026E+02	3.016E+02
3.00000E-03	9.033E+01	8.963E+01
4.00000E-03	3.778E+01	3.723E+01
5.00000E-03	1.912E+01	1.866E+01
6.00000E-03	1.095E+01	1.054E+01
8.00000E-03	4.576E+00	4.242E+00
1.00000E-02	2.373E+00	2.078E+00
1.50000E-02	8.071E-01	5.627E-01
2.00000E-02	4.420E-01	2.238E-01

Figure 6.3: Section of the X-ray mass attenuation coefficient data for carbon from NIST (**?**). The mass attenuation coefficient data are tabulated beneath the graph. The exact energy of the X-ray photons used in crystallography and SAXS (typically around 12 keV) is not explicitly tabulated therefore the mass attenuation coefficient is obtained by linear interpolation.

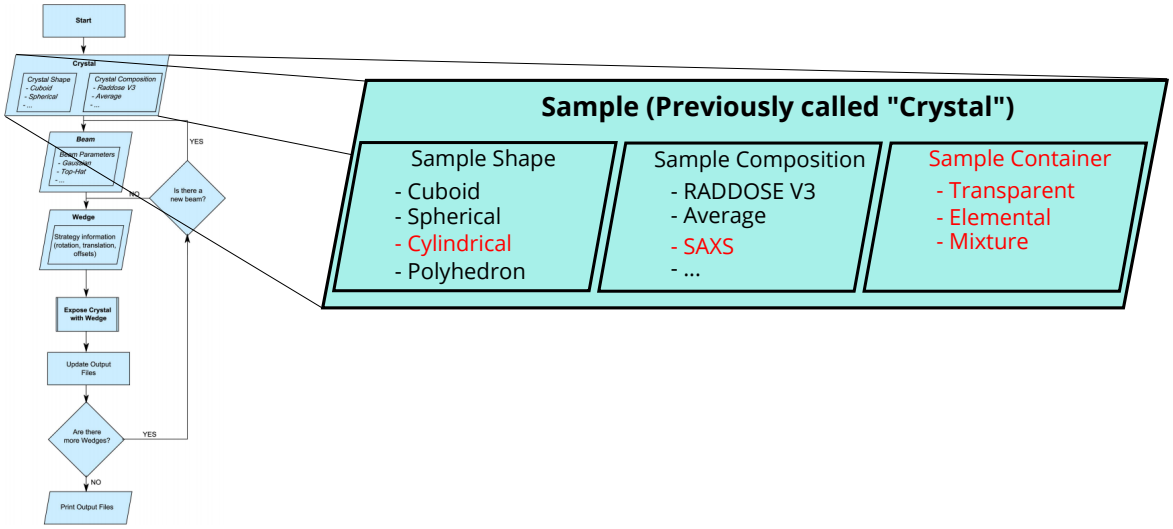


Figure 6.4: RADDOSE-3D flowchart from Figure 6.1 with the updated sample implementation, which extends the previous "crystal" implementation. The text coloured red show the SAXS extensions covered in this section.

obtained using a weighted average given by

$$\mu/\rho = \sum_i w_i(\mu/\rho)_i, \tag{6.2.5}$$

where w_i and $(\mu/\rho)_i$ are the fraction by weight and mass attenuation coefficient of the i^{th} atomic constituent respectively.

6.2.5 Summary of SAXS extensions

Figure 6.4 puts into context how the SAXS extensions fit into the RADDOSE-3D program design. All of the changes have been made in the "crystal" definition which has been referred to as the *sample* in this section.

Figure 6.5 shows explicitly how the extensions described above combine to allow accurate dose calculations for SAXS experiments. The cylindrical sample geometry is defined first from the specified height and diameter of the sample in the capillary. Then the atomic sample composition is defined from the protein concentration and buffer components. Once the container material is specified, the attenuated X-ray beam flux can then be calculated. Finally the SAXS sample is exposed to the attenuated X-ray beam according to the 'wedge' parameters.

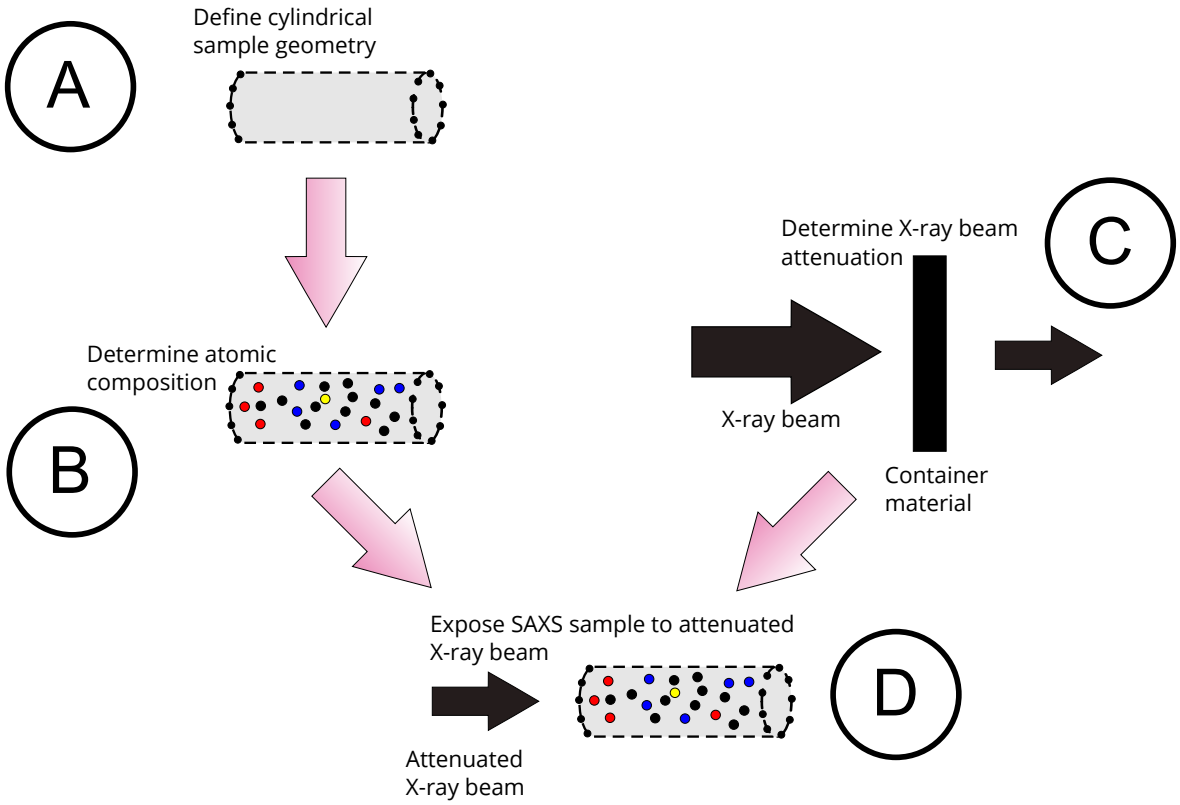


Figure 6.5: Flow diagram showing how the extensions described in this section are combined to enable dose calculation. A: Define the cylindrical geometry of the SAXS sample (section 6.2.2). B: Determine the atomic composition (section 6.2.3). C: Calculate the attenuation of the beam due to the sample container (section 6.2.4). D: Expose the SAXS sample to the attenuated beam according to the defined wedge.

6.2.6 Model considerations

The model of the SAXS experiment in RADDOSE-3D makes many implicit assumptions. For instance, with regards to the capillary, the atomic composition is assumed to be uniform throughout, and the thickness to be constant around the entire sample volume. The advantage of these assumptions are that the attenuation by the capillary only needs to be calculated once, regardless of any movement or rotation of the capillary. This is valid for a cylindrical capillary since the thickness penetrated by the X-ray beam is the same regardless of any rotations or translations.

The sample itself is assumed static, moving as a rigid body when rotated or translated, and completely filling the capillary. This greatly reduces the computational cost when compared to the possibility of modelling the exact fluid dynamics. The assumption that the capillary volume is completely filled is generally valid since this is usually the case during an experiment. The static assumption is not always valid, especially when the sample is flowed through the capillary. Hopkins and Thorne calculated that for typical experimental param-

eters (capillary diameters: 1.5-2 mm, flow velocities: 0.5-15 mm s^{-1}) the Reynolds numbers* are in the range 1-25 implying a laminar flow regime (**?**). In this case the velocity profile is expected to exhibit the quadratic Poiseuille flow profile, which arises from the axial symmetry and no-slip boundary assumptions (the velocity at the centre of the tube moves the fastest while the velocity at the boundary is equal to zero provided the capillary is also stationary). Furthermore, Hopkins and Thorne also calculated that the residence times of the sample in the beam were too short for any appreciable radial diffusive mixing, hence the flow profile results in radius-dependent residence times in the X-ray beam. Therefore the static assumption here will give misleading dose values if calculated for experiments where the sample is flowed through the beam.

Hopkins and Thorne additionally discuss diffusive turnover as a phenomenon that will affect the dose calculation (**?**). Molecules have the ability to diffuse into and out of the illuminated volume, with the additional complexity that a non-uniform beam profile will cause differential diffusion across the beam. In the current work, no account was taken for molecular diffusion. Extrapolation of the plot from (**?**) (reproduced in Figure 6.6) shows that diffusion is likely to be negligible for the experiment performed here (section 6.3) where the beam size was 600 $\times$ 600 μm^2 and the total exposure time was 120 seconds.

The SAXS sample can also be manipulated in the same way as a crystal in RADDOSE-3D. Hence helical scanning, translations and rotations of the sample can all be performed. Figure 6.7 is the final dose state of a glucose isomerase sample in an experiment where the sample was rotated 180° for a total of 200 seconds in a 700 $\times$ 700 μm^2 top-hat profile beam with a flux of 1.51 $\times$ 10^{13} ph/s and an incident photon energy of 12.1 keV.

*The Reynolds number is defined as the ratio of inertial forces to viscous forces and is a quantity used in fluid mechanics to determine properties of fluid flow (**?**).

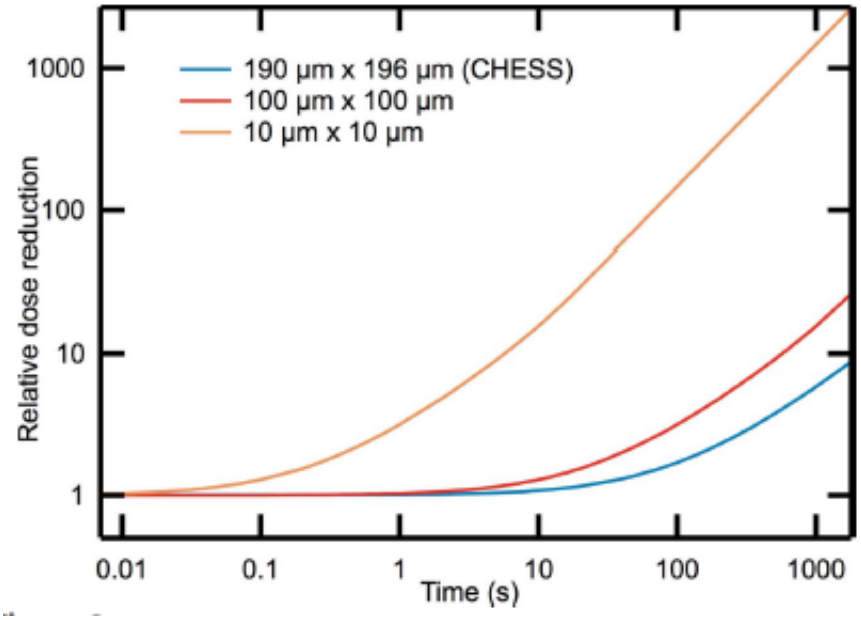


Figure 6.6: Reduction in the dose absorbed by the sample due to diffusive exchange of lysozyme against time for three beam sizes. Extrapolation of these curves for a beam size of 600 $\times$ 600 μm^2 along with 120 second exposure times suggests that the dose reduction due to diffusive turnover is negligible for the experiment reported in this chapter. Figure reproduced from (**?**).

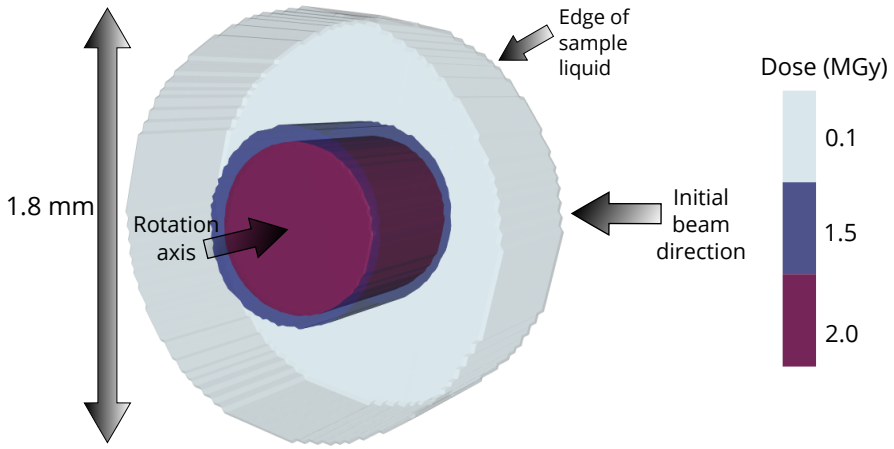


Figure 6.7: Final dose state of a glucose isomerase sample in an experiment where the sample was rotated 180° in a 700 μm $\times$ 700 μm top-hat profile beam with a beam flux of 1.51 $\times$ 10^{13} ph/s and an incident energy of 12.1 keV for a total exposure time of 200 seconds. The capillary was treated as completely transparent to the X-rays so there was no attenuation. The dose state is only calculated and shown for the sample, not the capillary. The contours correspond to dose iso-surfaces: light blue = 0.1 MGy, dark blue = 1.5 MGy and purple = 2 MGy.

6.3 Experimental methods

This section describes the experimental details for two experiments comparing the efficacy of various radioprotectant compounds. The first of these experiments was carried out in October 2014 whereas the second was carried out in December 2015, both at the ESRF, in Grenoble.

6.3.1 Sample preparation

Crystalline glucose isomerase (GI) used in both experiments was purchased in tetrameric form (1552 residues, 172 kDa) from Hampton Research. GI was chosen because it is a stable, soluble globular protein that has well defined SAXS behaviour and is sufficiently large to scatter well at modest concentrations.

Experiment 1 (Expt 1)

Radioprotectants and GI were dissolved separately in buffer (100 mM HEPES and 10 mM $MgCl_2$ at pH 7.0) to give stock solutions of both at twice the desired radioprotectant and protein concentrations respectively. The protein and scavenger stocks were then mixed thoroughly by pipetting in a 1:1 ratio to create the samples. The radioprotectant stock was also mixed in a 1:1 ratio with the buffer by the automatic liquid handling on the beamline to create the buffer sample. Both samples were prepared at the beamline immediately before the diffraction experiment in order to minimise protein aggregation before data collection.

The final concentrations of each component in solution were as follows: 0.54 mg/ml GI, 5 mM soluble radioprotectant, 5% v/v glycerol or ethylene glycol. Each sample was prepared in triplicate to allow repeats. The following radioprotectants were investigated: sodium ascorbate, sucrose, sodium nitrate, trehalose, ethylene glycol, (2,2,6,6-tetramethylpiperidin-1-yl)oxyl (TEMPO), glycerol and glycerol + nitrate.

Experiment 2 (Expt 2)

GI was dissolved and dialysed for 24 hours at 277 K with the same buffer components as those used in the first experiment. The final GI concentration, 1 mg/ml, used for all data collection runs was determined using the extinction coefficient given by 45,600 $M^{-1}cm^{-1}$ at 280 nm absorbance. Eight solution additives were tested for their radiation damage protection capabilities: dithiothreitol (DTT), ethylene glycol, glycerol, sodium ascorbate, sodium nitrate, sucrose, (2,2,6,6-tetramethylpiperidin-1-yl)oxyl (TEMPO) and trehalose. The additives were added to the buffer solutions at four different concentrations: 10 mM, 5 mM, 2 mM and 1 mM, except glycerol and ethylene glycol which were both prepared at 10% v/v, 5% v/v, 2% v/v and 1% v/v immediately prior to data collection. These additives were also prepared to the same final concentration in the solution containing both the buffer and protein.

6.3.2 SAXS data collection

Experiment 1

Data collection was carried out by Dr. Ed Lowe at the ESRF on beamline BM29. The X-ray photon energy was 12.5 keV (wavelength of 0.9919 Å), with a flux of 4.84×10^{11} photons per second at 100% transmission. The size of the beam was $700 \times 700\,\mu m^2$, however the beam profile was not experimentally measured. The profile was assumed Gaussian for the purposes of dose calculation (section 6.3.3). Using the EMBL sample loading robot, 15 μl of sample was loaded into a 1.8 mm diameter quartz capillary tube with a wall thickness of 0.03 mm and data were collected at 293 K with flow mode turned off. Frames were collected at 0.5 second intervals on a Pilatus 1M detector for durations of either 30 seconds (60 frames) or 60 seconds (120 frames).

Experiment 2

Data collection was performed at the ESRF by the author on beamline BM29 in collaboration with Adam Round and Martha Brennich. The photon energy used throughout was 12.5 keV and the photon flux was estimated from diode readings which were recorded for

every frame using the conversion formula

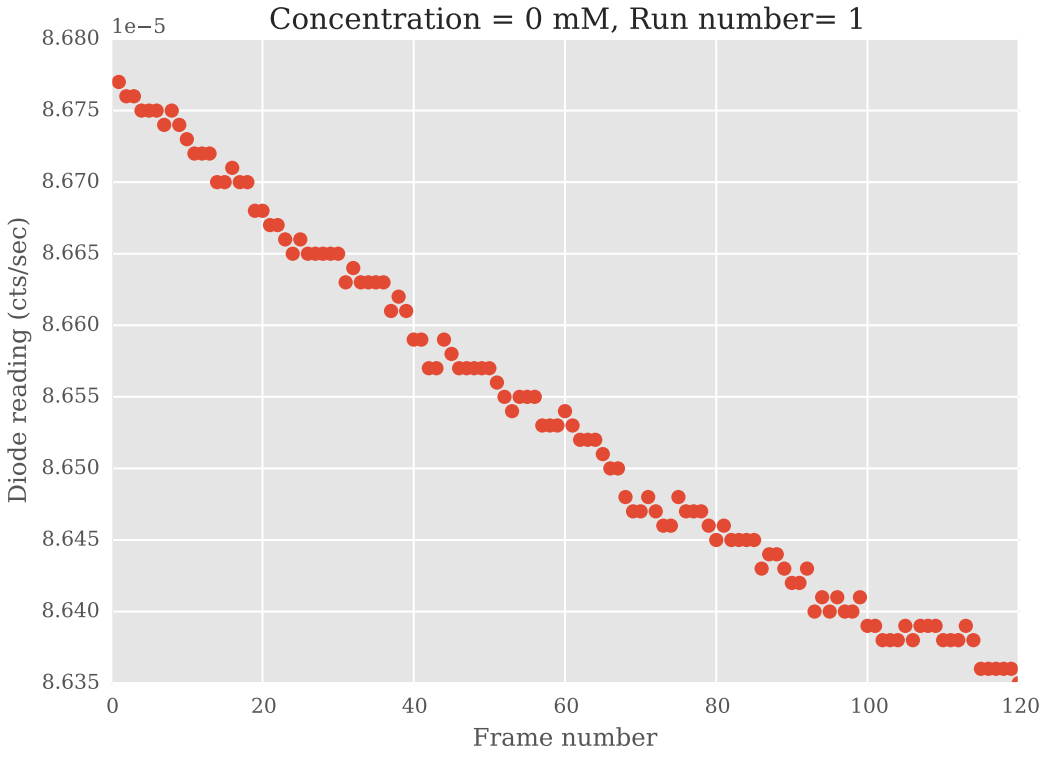
$$\text{flux} = 5.72293 + 2.72295 \times 10^{15} \times d_r, \tag{6.3.1}$$

where d_r is the diode reading. The flux obtained using this formula was calibrated with an OSD1-0 photodiode purchased from Optoelectronics, which was a 500 μm thick silicon diode with a 1 mm^2 active area as per Owen *et al.* (2009) . The flux was calculated for each frame because the diode readings constantly changed between frames (Figure 6.8), however the overall change in diode reading was only 0.54%. Despite the small percentage change, account was still taken for this effect in the analysis. The full 2D X-ray beam profile was determined as outlined in section **??**. The resulting beam is shown in Figure 6.9 as a greyscale image.

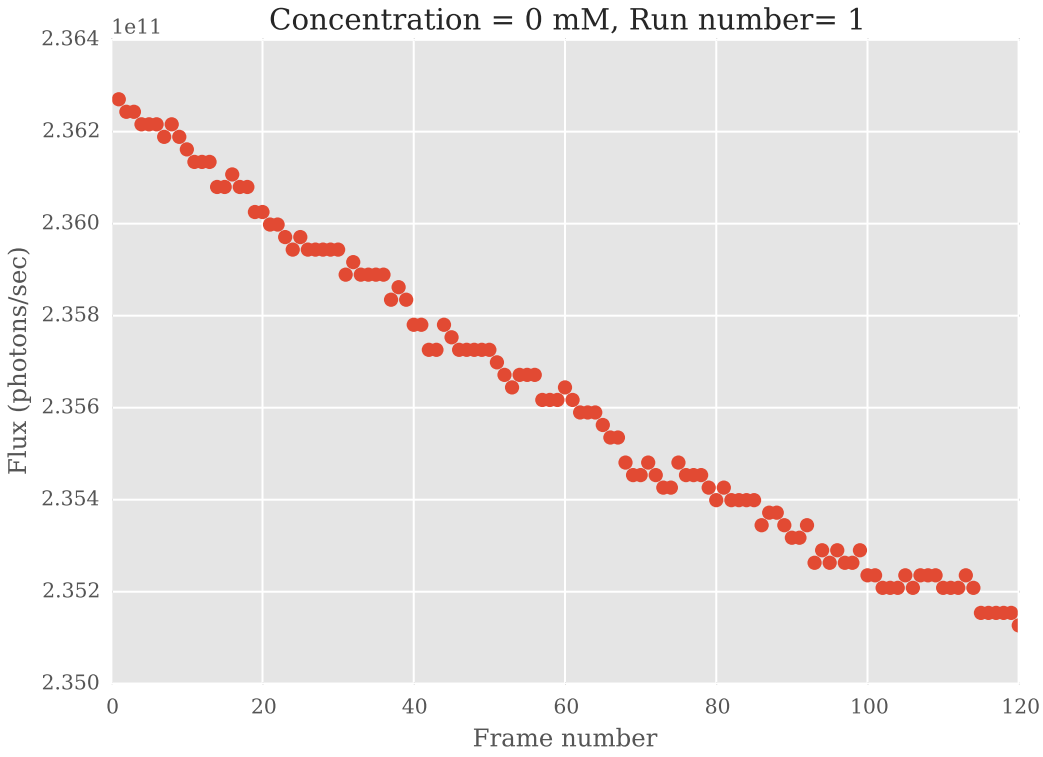
The data were recorded using a Pilatus 1M detector from Dectris. 15 μl of each sample was loaded into a 1.8 mm external diameter quartz capillary (1.7 mm internal diameter, thus the wall thickness is 50 μm) held at room temperature using the automated robotic sample changer (**?**). For each additive, data were collected at each concentration (given in section 6.3.1), and each of these runs was repeated 3 times. The exposure time for each frame was 1 second, and a total of 120 frames were collected for a single run with the sample kept static. For each radioprotectant concentration a single dataset was collected with only the buffer (no protein) so that a suitable buffer correction (subtraction) could be applied during data analysis.

6.3.3 Dose calculations

Dose calculations for both experiments were performed using RADDOSE-3D with the modifications described in section 6.2 for modelling SAXS experiments. An example input file for Expt 1 can be seen in Figure 6.10. The beam profile was assumed Gaussian with full width at half maximum of 110 $\mu m \times 200\,\mu m$. For Expt 2, the beam image shown in Figure 6.9 was read directly into RADDOSE-3D and the corresponding flux values recorded for each frame were used to ensure accurate dose estimates. The atomic composition is the same as that defined in Figure 6.10, although the 5 mM concentration of sodium is only defined for the radioprotectants that contain sodium (i.e. sodium ascorbate and sodium nitrate). DTT con-



(a)



(b)

Figure 6.8: Diode readings and flux estimates during the first SAXS repeat for the GI sample with no radioprotectant added (hence concentration = 0 mM). (a) Diode readings for each frame in the experiment. It can clearly be seen that the diode readings decrease throughout the experiment, which is due to the decay of the electron storage ring current. However, the total change during the course of the dataset is only 0.54%. (b) Flux estimates for each frame in the same experiment as (a). As a result of the decreasing diode readings the corresponding flux decreases by 0.54% too.

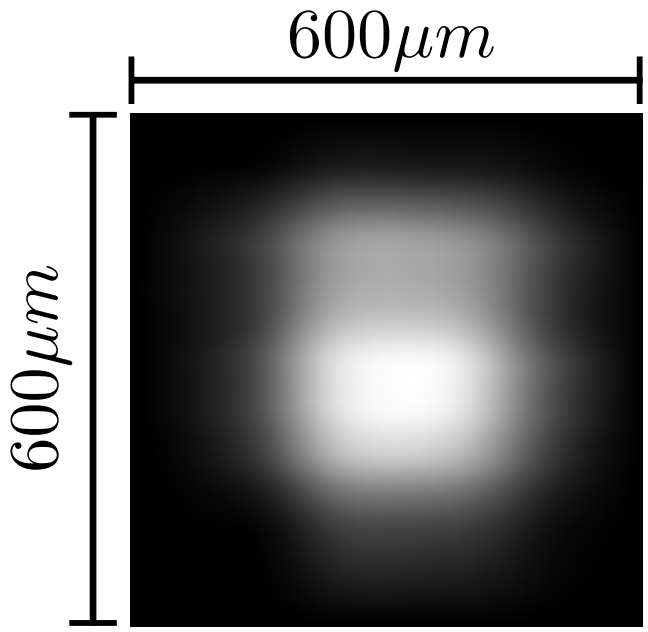


Figure 6.9: A 2D reconstruction of the beam used in the second SAXS experiment (Expt 2) shown as a greyscale image.

```
##############################################################################
#                                Crystal Block                               #
##############################################################################
Crystal
Type Cylinder
Dimensions 1740 1000                  #Diameter and height of sample cylinder
PixelsPerMicron 0.005
AbsCoefCalc SAXS
NumResidues 1552
ProteinHeavyAtoms S 32
ProteinConc 0.54                      #Protein concentration
ContainerMaterialType Elemental
MaterialElements Si 1 O 2             #Elemental composition of quartz
ContainerThickness 50
ContainerDensity 2.648
SolventHeavyConc S 100 Mg 10 Cl 10 Na 5 #Heavy atoms in solvent

##############################################################################
#                              Beam Block                                    #
##############################################################################
Beam
Type Gaussian
Flux 4.84e+11
Energy 12.5
FWHM 110 200
Collimation Rectangular 700 700

##############################################################################
#                              Wedge Block 1                                 #
##############################################################################
Wedge 0 0
ExposureTime 1

##############################################################################
#                              Wedge Block 2                                 #
##############################################################################
Wedge 0 0
ExposureTime 1
```

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

tains a sulphur atom for which account must be taken also in the solvent concentration. The elemental composition of the quartz capillary was entered into RADDOSE-3D as SiO_2. The capillary thickness was given as 50 μm and a density of 2.648 g/cm^3 was used.

6.4 1D scatter curve similarity analysis

To increase the signal to noise ratio when processing data in SAXS, it is necessary to merge 1D intensity curves from several frames. If the scattering for different frames is from the same molecule in the sample, then the frames will be similar i.e. the frames will overlap. However as radiation damage progresses during the experiment, the molecules begin to aggregate, which causes the intensity curve to increase at low q angles and is speculated to decrease at high q angles (**?**). On the other hand, fragmentation and molecule repulsion due to protein charging, cause a decrease in scattering at low q angles and an increase at high q angles. Figure 6.11 shows the scattering profile of several SAXS frames collected from the same glucose isomerase sample. As the dose increases, there is a clear decrease in the scattered intensity at low angles suggesting that the sample is undergoing fragmentation. This was also observed with the glucose isomerase samples used in the study conducted by Hopkins and Thorne (**?**). Therefore it is necessary to determine the similarity between any two pairs of frames to determine which frames to merge together. Due to the fact that a new method of assessing the similarity was published between performing experiment 1 and 2 (**?**), the analysis performed on the two sets of data and presented here are different. This was necessary because the new version of DATCMP, the software program used to perform the similarity analysis (**?**), does not incorporate some of its old functionality utilised in Expt 1, and the online manual describing the new features has not yet been updated.

6.4.1 Data analysis - experiment 1

Buffer subtraction was first carried out in ScÅtter (Rambo, R. at DLS, Didcot, UK). DATCROP, a utility program for cropping SAXS data, distributed as part of the ATSAS program suite (**?**), was then used to remove data points from the very low angles around the beam stop and also from the larger angles where the signal-to-noise ratio drops considerably: this was carried out by visual inspection. DATCMP (a program also distributed in the ATSAS suite) was then used to carry out a Scheffe *post hoc* analysis on the resulting 1D scattering data (frames) for each experimental run. This analysis compares the similarity of the first frame to each of the subsequent frames, since the first frame is produced at the lowest dose and so it is assumed to be of the best quality. The result of the Scheffe *post hoc* analysis is a 'fi-

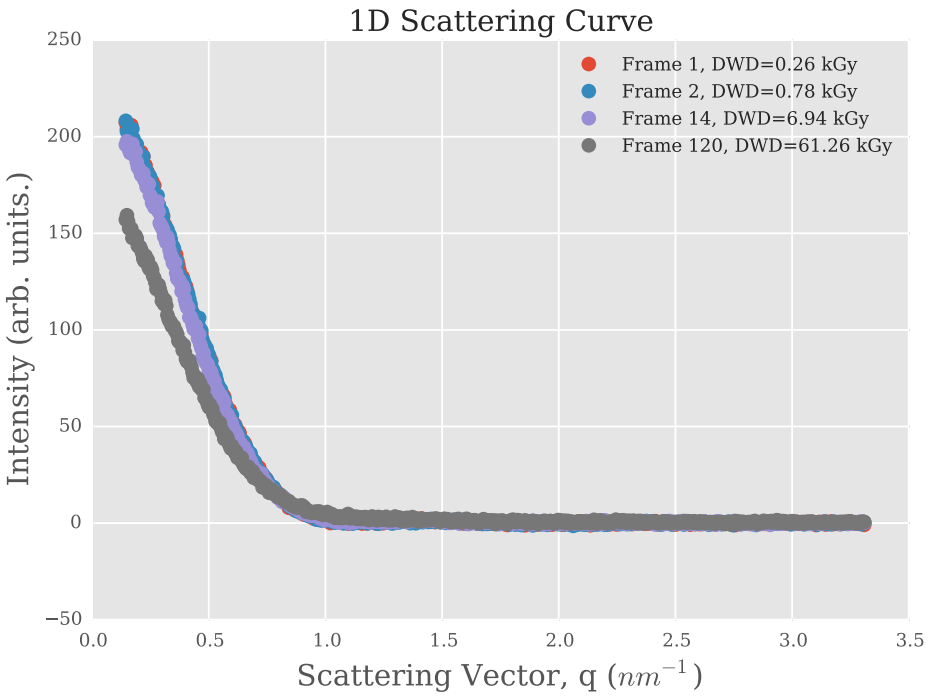


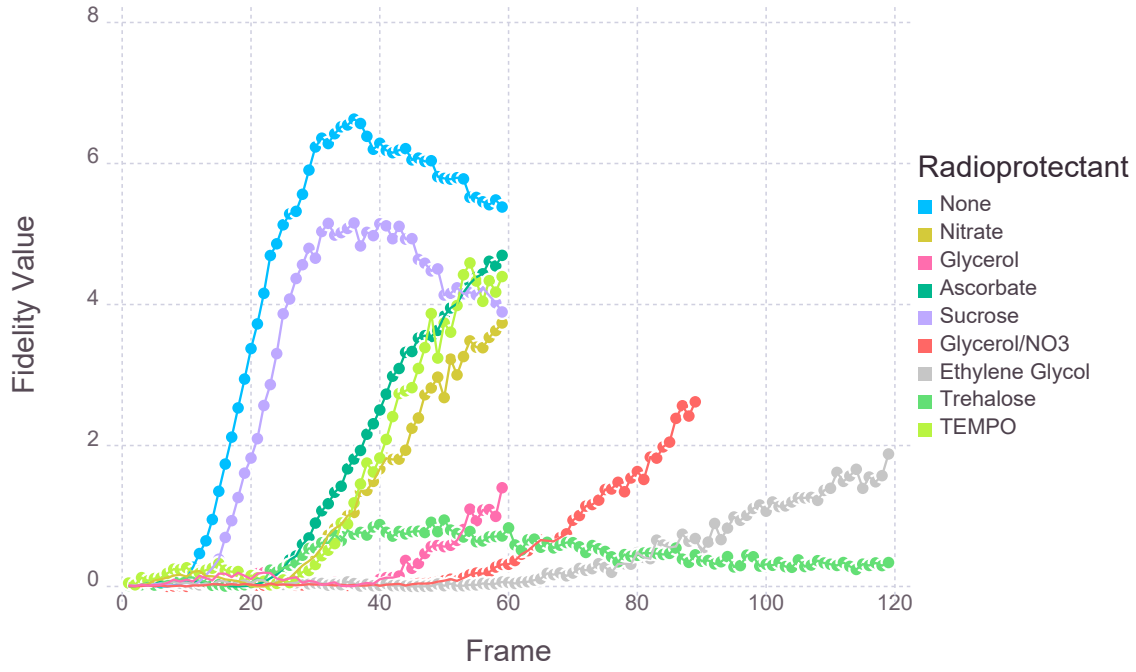
Figure 6.11: 1D scattering curves from the first run of the GI sample with no radioprotectant compounds added. Frames 1 and 2 visually seem to overlap well and the merging analysis determines these frames to be similar. Frame 14 is the first frame found to be dissimilar to frame 1. However, by visual inspection, the dissimilarity is not obvious. Frame 120 was the last frame collected in this run. The dissimilarity of frame 120 from the others is obvious. This suggests that the molecules in the sample have undergone significant radiation induced changes during the experiment.

delity value', which is a value given for each frame to describe its similarity to the first frame. An identical frame is given a fidelity value of 0. Increasing fidelity values correspond to increasing dissimilarity of a particular frame from the first frame. A plot of the fidelity values against time and diffraction weighted dose (DWD) is shown in Figure 6.11. These data show the changing similarity of frames throughout the experiment, but they do not explicitly provide a quantitative metric to compare the efficacies of the various radioprotectants. Further analysis of these data was required to achieve this.

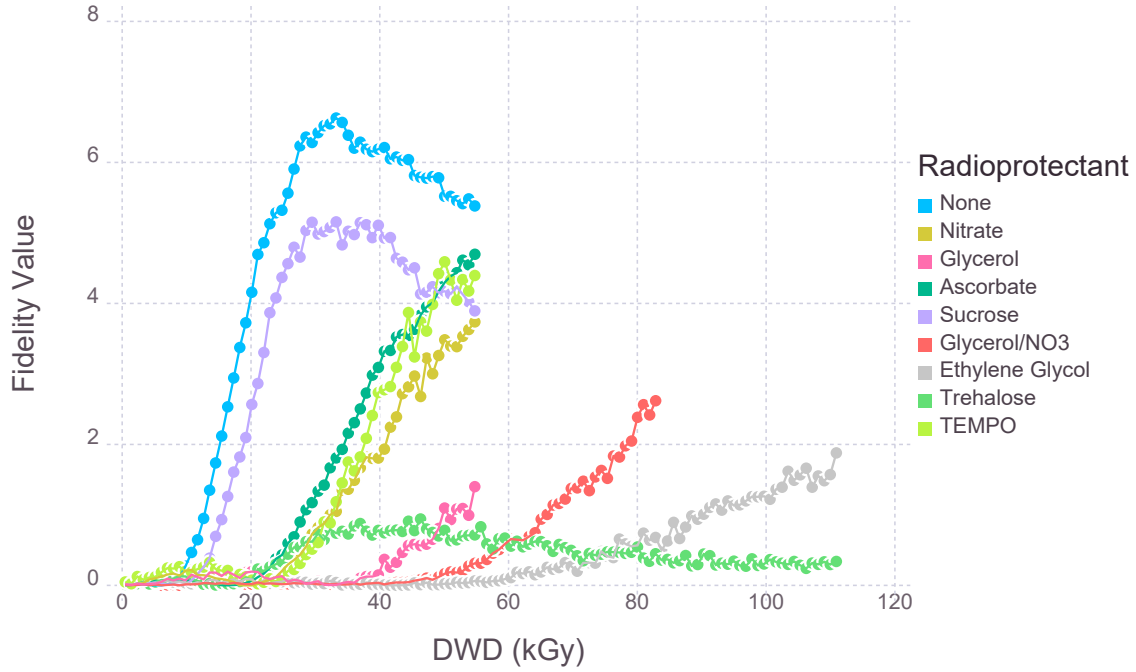
The curves of fidelity values against both time and DWD visually resemble a logistic relationship (they look like S-shaped curves). Thus 4-parameter logistic (4PL) functions were fitted to the data:

$$F(x) = \frac{a-d}{(1+(x/c)^b)^2} + d, \tag{6.4.1}$$

where F is the fidelity value, x is the x-axis coordinate (it could either be time or DWD), and a, b, c and d are the 4 parameters to be determined. Each of the parameters has graphical interpretation in relation to the logistic curve. a and d represent the minimum and maximum values of the curve respectively, b represents the steepness of the slope and c is the location



(a)



(b)

Figure 6.11: (a) Fidelity value as a function of time (b) Fidelity value as a function of dose. Fidelity values start at 0 for the first frame and increase as the frames become more dissimilar.

of the point of inflection. The 4PL curve was used instead of the commonly used 3 parameter logistic function (also known as the Hill function) because it allows much more flexibility for the steepness of the curve, which was likely to change with different radioprotectants.

To determine which frames should be merged it is necessary to decide on the threshold similarity criterion. Two criteria were used:

1. the DWD absorbed to reach an arbitrary value, and
2. the DWD absorbed to reach maximum curvature of the fidelity curve.

The first of the two criteria is straightforward to assess. A fidelity value is chosen and then the fitted logistic function can be rearranged to find the DWD value for which the chosen fidelity value is reached.

The second of the criteria is slightly more involved. For frames to be similar, their fidelity values have to be close to zero. Therefore the early frames should be close in value. However when frames start to become dissimilar, the fidelity value increases. In terms of the fitted logistic curve, this corresponds to an increasing gradient as the fidelity values increase. The required value is the total absorbed dose at the point when the increase in the gradient of the fidelity values reaches a maximum. To perform this analysis it was necessary to calculate the curvature, κ, of the logistic curves, which is given by:

$$\kappa(x) = \frac{\frac{\mathrm{d}^2 F}{\mathrm{d}x^2}}{\left(1 + \frac{\mathrm{d}F}{\mathrm{d}x}\right)^{3/2}}. \tag{6.4.2}$$

This equation was calculated symbolically using the symbolic math toolbox in MATLAB. The maximum of this function was found using the MATLAB function `fminbnd` to find the minimum value of $-\kappa(x)$ i.e. the problem was converted from trying to find the maximum into a problem in which the same solution is found by finding the minimum of the negative of the function. This is a common procedure for optimisation problems.

An implicit assumption with the methods described is that damage is entirely progressive and subsequent frames beyond the threshold are all significantly dissimilar. Mathematically this assumption manifests itself as logistic functions representing the fidelity as a monotonically increasing function of the dose.

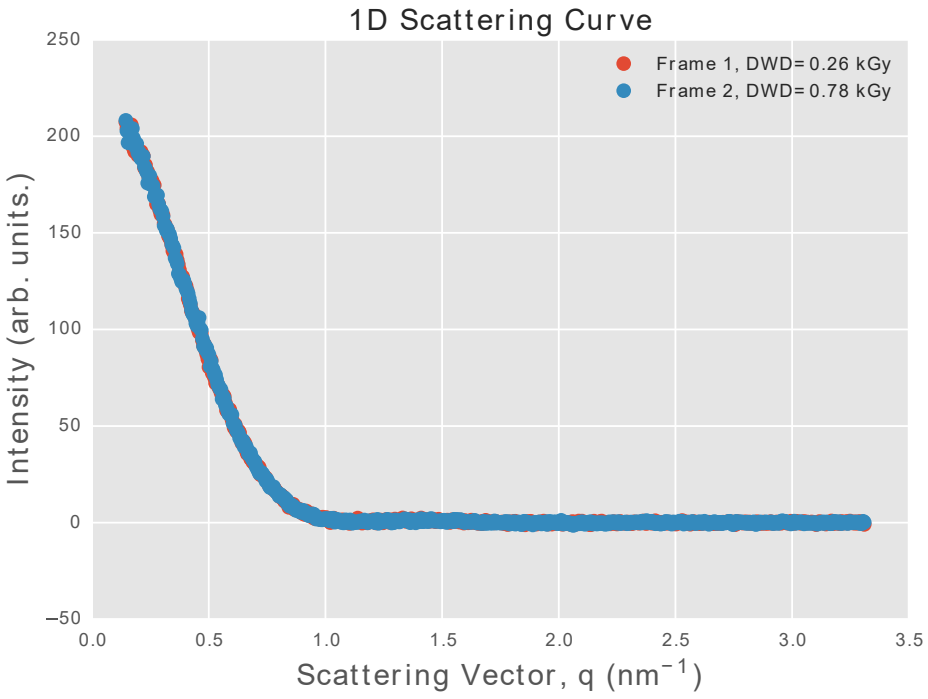
6.4.2 Data analysis - experiment 2

In a similar manner to the analysis performed on the results from Expt 1, buffer subtraction and cropping were performed before 1D curve similarity analysis. However these steps were performed with custom written Python scripts. This was done for two reasons: the first was that the pipeline could then be completely scripted (ScÅtter, which was used for buffer subtraction for data processing for Expt 1, is a GUI based program written in Java), and secondly, DATCROP subtracts data from files, so the scripts that were written would require many file handling operations, which are relatively time consuming computationally. Therefore writing the scripts to perform these simple operations (subtraction and cropping) was much more time efficient and allowed for more flexibility.

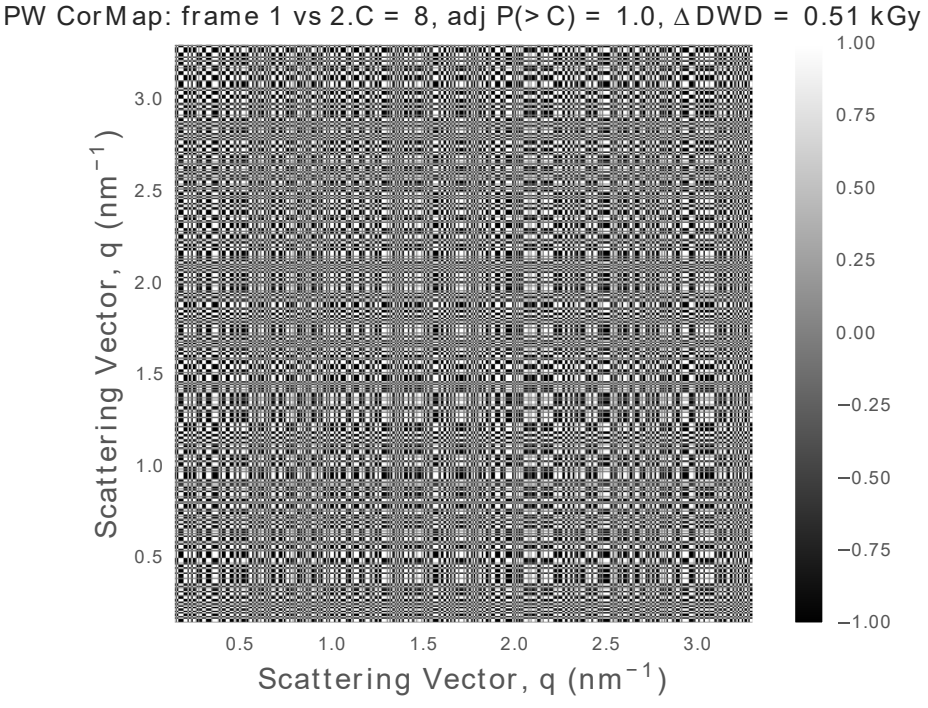
DATCMP was again used for the similarity analysis. However, as mentioned in section 6.4, the method used in the current version has been changed by the authors of this software since the Expt 1 analysis. The new method is called the Correlation Map (CorMap) test (**?**). Unlike the algorithm used to generate the fidelity values, the CorMap test does not require any estimates of the experimental errors to test similarity.

A full explanation of the method can be found in Franke *et al.* (2015), but the main ideas are presented here and tested for their suitability for application to our data. Any two frames can be compared using a pairwise correlation, which involves calculating the difference between the two intensity curves. If the two frames are identical up to the noise level, then the difference between them is a random number. Given that the intensity values are assumed to come from a normal distribution (**?**), the difference between them is normally distributed with a mean of zero. Due to the symmetry of the normal distribution, the probability of the difference of the data being either positive or negative is 0.5. The pairwise CorMaps between selected frames from the experiment are shown in Figure 6.12. Notice that for similar frames 1 and 2 (Figure 6.12a), the pairwise CorMap resembles a randomised lattice (Figure 6.12b) as expected for identical data. Conversely, for dissimilar frames 1 and 120 (Figure 6.12c), the pairwise CorMap shows large regions of white and black patches (Figure 6.12d), suggesting that the frames are systematically different. DATCMP performs quantitative analysis that formalises the similarity conclusions that can be drawn from these CorMaps.

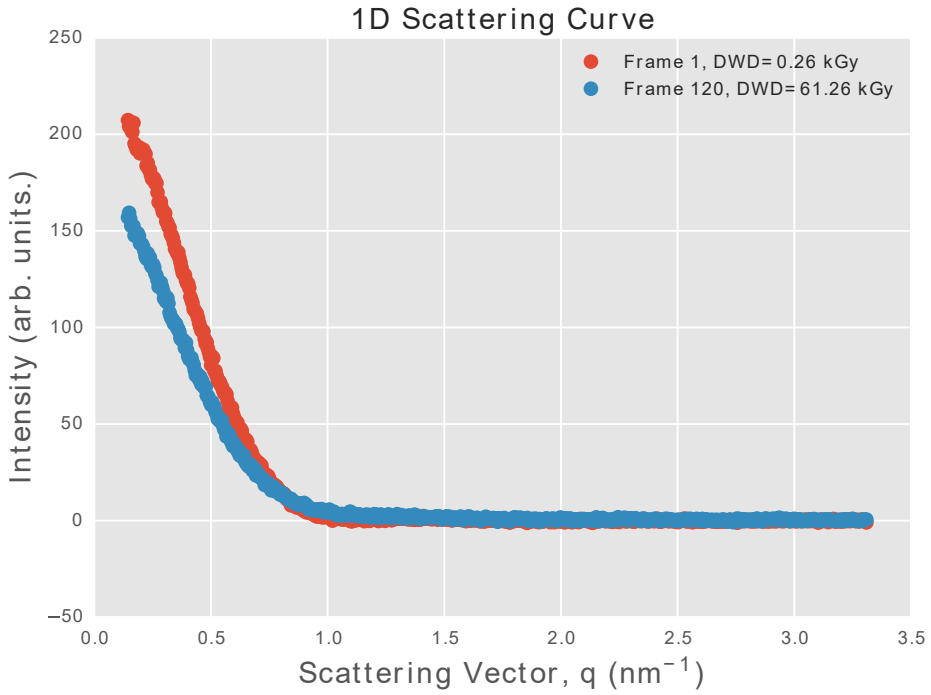
The problem can be thought of in an identical manner to a coin toss experiment with the


1D Scattering Curve
Frame 1, DWD= 0.26 kGy
Frame 2, DWD= 0.78 kGy
Intensity (arb. units.)
Scattering Vector, q (nm^{-1})

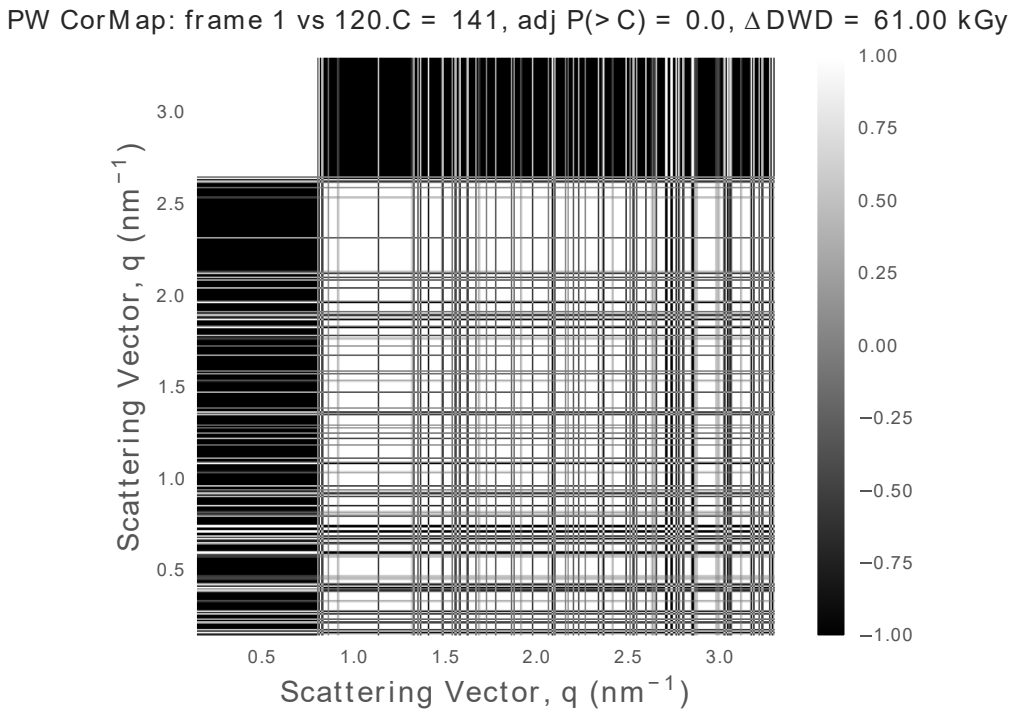
(a)


PW CorMap: frame 1 vs 2.C = 8, adj P(> C) = 1.0, ΔDWD = 0.51 kGy
Scattering Vector, q (nm^{-1})
Scattering Vector, q (nm^{-1})

(b)



(c)



(d)

Figure 6.12: Similarity comparison with selected frames from the first experimental repeat with no radioprotectant added. (a) 1D scatter curves for frames 1 and 2. These two curves overlap well and are classed as similar (b) Pairwise CorMap between frames 1 and 2. The ostensibly randomised lattice pattern suggests that the 1D curves are similar. (c) 1D scatter curves for frames 1 and 120. It is clear that these frames do not overlap. (d) Pairwise CorMap between frames 1 and 120. The dissimilarity between the two frames is represented by the large black and white regions.

following two conditions:

1. the probability of the difference of any two data points being either positive or negative is 0.5,
2. the result of the difference of any two data points with another two are assumed independent,

In the coin toss experiment one can ask: 'what is the probability of observing more than n consecutive heads (or tails)?'. The Schilling distribution (**?**) calculates this probability. For the SAXS experiment this is the same as asking for the probability of observing a patch of white (+1) or black (-1) larger than the longest patch of white or black that was actually observed in the pairwise CorMap. More formally, we ask for the probability (P value) of obtaining an edge length larger than C (denoted $P(> C)$) within an n-by-n correlation matrix. This is calculated from the Schilling distribution with parameters n and C. If the p value is smaller than a threshold value α ($\alpha \leq 0.01$ is recommended (**?**)) then the two frames can be considered dissimilar.

The CorMap test is implemented in the current distribution of DATCMP, although the software does not include visualisation tools. Since multiple pairwise tests have to be made from several comparisons, the Bonferroni correction[†] is applied to the $P(> C)$ values resulting in $P_{adj}(> C)$ values. The raw output from the program is essentially a list of $P(> C)$ and $P_{adj}(> C)$ values for all possible pairwise comparisons. The set of $P_{adj}(> C)$ values that result from comparing the first frame to all subsequent frames is shown in Figure 6.13.

Applying this methodology to Expt 2 results

As can be seen in Figure 6.13, the first frame which is calculated to be dissimilar to frame 1 is frame 55 (first orange circle) which has a corresponding dose of 14.73 kGy. Therefore the threshold for radiation damage onset to be significant could be set at that frame. However the frames immediately after 55 do not have $P_{adj}(> C) < 0.01$ (green or blue circles) suggesting that these frames are not significantly dissimilar to frame 1. Thus frame 55 may

[†]If multiple hypotheses are tested then the likelihood of observing a rare event (and thus the likelihood of incorrectly rejecting the null hypothesis) is increased. The Bonferroni correction divides the overall statistical significance level, α, by the number of tests, n, so that the hypotheses are tested individually at a significance level of α/n.

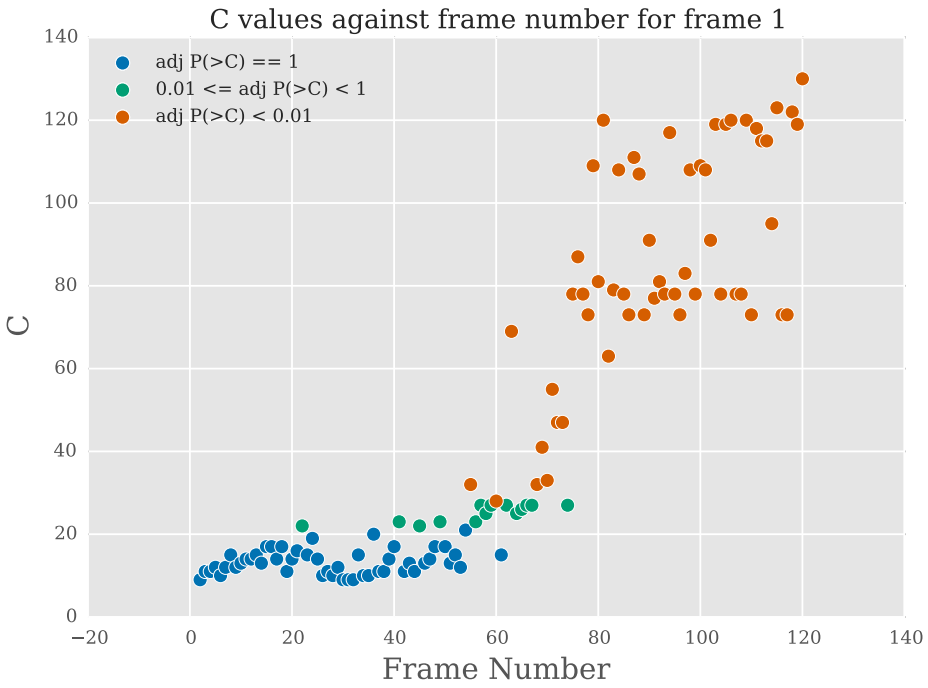
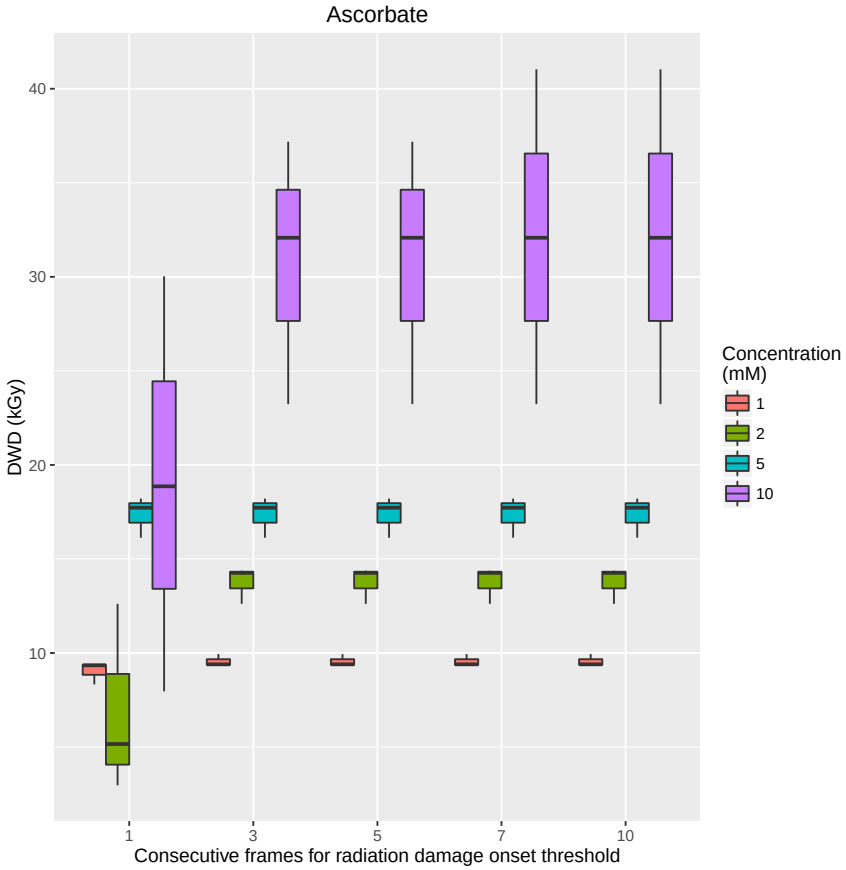


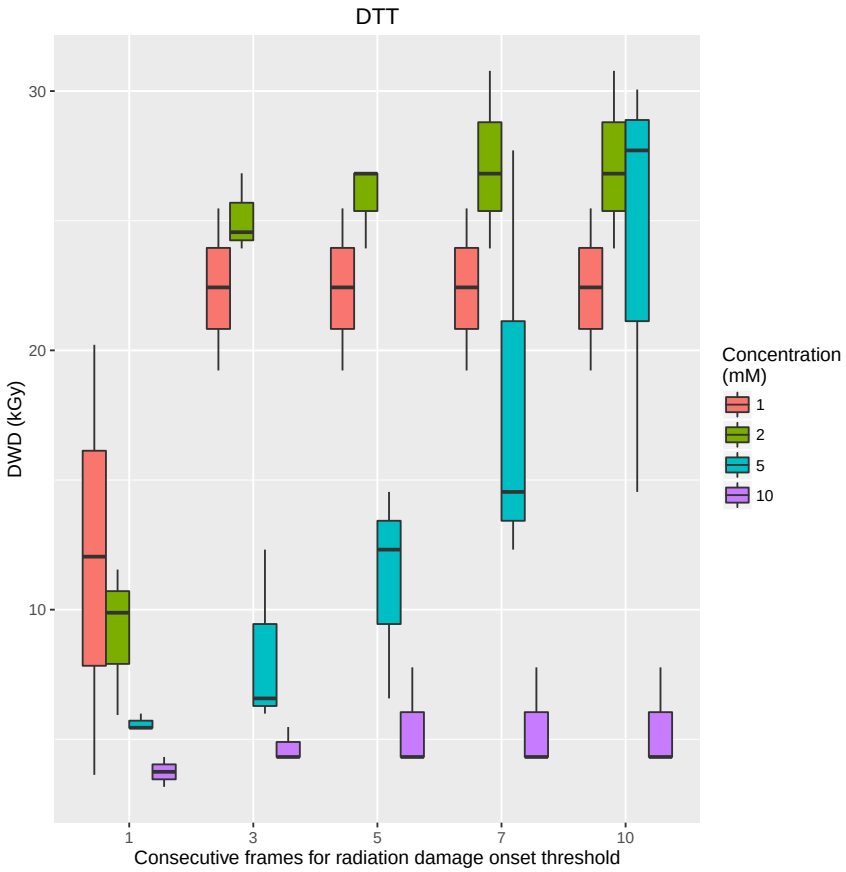
Figure 6.13: Longest observed edge length C against the frame number for pairwise comparisons with frame 1. For similar frames to frame 1, the pairwise CorMaps are more like randomised lattices (Figure 6.12b) and hence C is fairly small. Therefore the chance of observing a longer edge length than C is high ($P_{adj}(> C) = 1$ - blue circles). As frames start becoming more dissimilar, the C values increase and the $P_{adj}(> C)$ values fall. These circles are coloured green. When frames are very dissimilar, C becomes very large (Figure 6.12d) and $P_{adj}(> C) < 0.01$. The circles representing the comparison with these frames are coloured orange. The first dissimilar frame (coloured orange - frame 55) may not necessarily be the point at which frames should stop being merged according to this analysis, because there are circles after frame 55 that are not orange in colour.

be a noisy outlier and radiation damage may not necessarily be significant at that frame. A more robust check may be to find the first dissimilar frame for which m consecutive frames are dissimilar. This possibility was thus investigated and tested for m = 1, 3, 5, 7 and 10 to determine the value $m = m_0$ for which $m > m_0$ did not significantly change the dose at which frames were determined to be dissimilar. Figure 6.14 shows the result of the test for all of the radioprotectant experiments (note that the dose is used as the threshold for radiation damage instead of frame number). It can be seen that for most radio protectant compounds if m = 1, the spread of the apparent radiation damage onset is generally larger than for m = 3, 5, 7, 10. For m = 3, 5, 7, 10 the values corresponding to the onset of significant radiation damage are practically identical (except for DTT, but this is dealt with in section 6.5.2). Thus for the subsequent radiation damage analysis, the comparisons were performed with m = 3.

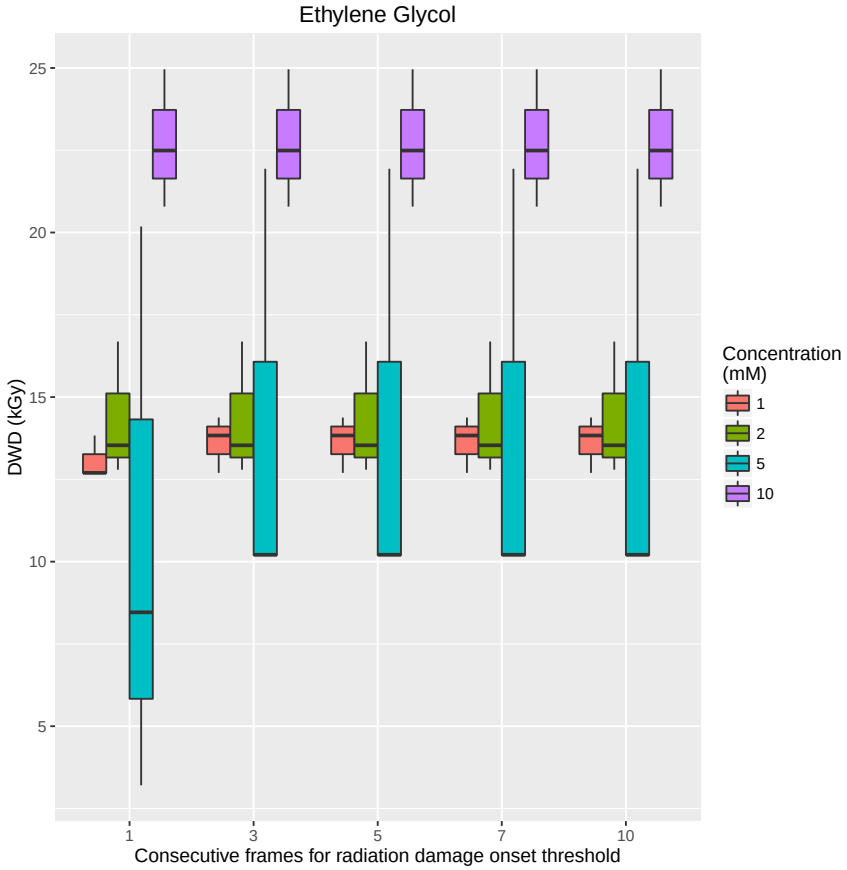
The other parameter that may affect the conclusions of this method for radiation damage analysis is the threshold level α. Therefore three different thresholds were set, α = 0.01, 0.05 and 0.1. Figure 6.15 shows the result of this test with all of the radioprotectant data. It can be seen that the median values are very similar for the various α values. Given this fact and that α = 0.01 is the recommended (and sufficiently strict) threshold (**?**), α = 0.01 was chosen as the value used for all radioprotectant compounds in the subsequent radiation damage analysis. The advantage of using α = 0.01 is that frames have to be very dissimilar before the $P_{adj}(> C)$ value falls below that value. This means that it is less likely that frames are discarded when they actually are similar (in statistical speak this means there is less chance of a type I error).


Ascorbate
DWD (kGy)
40
30
20
10
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

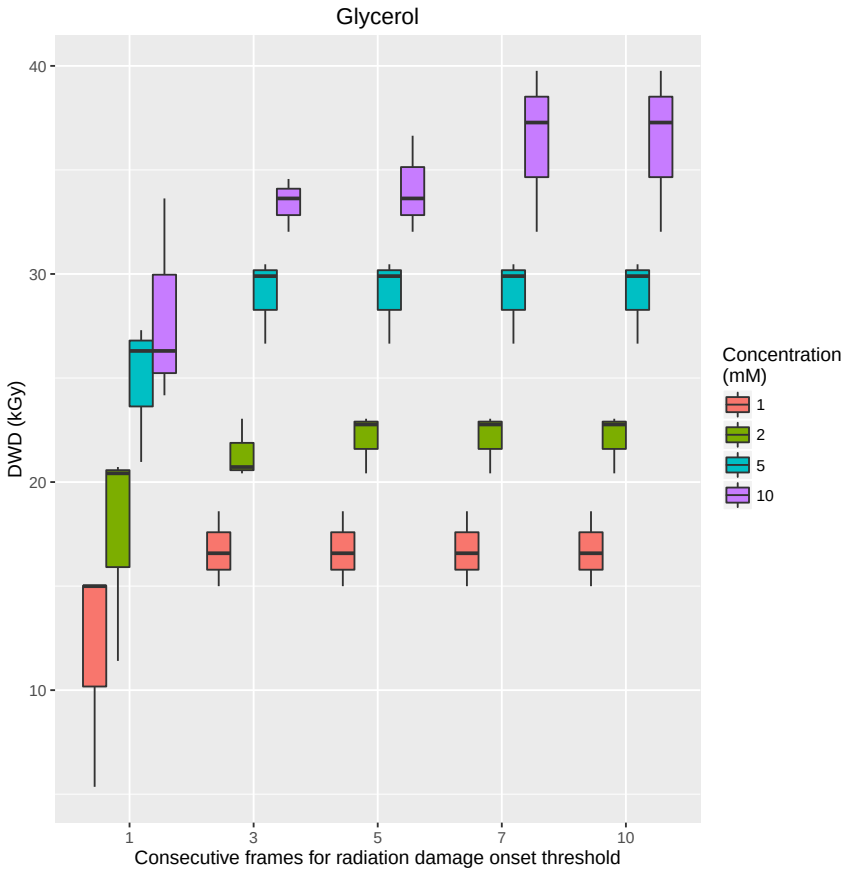
(a)


DTT
DWD (kGy)
30
20
10
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

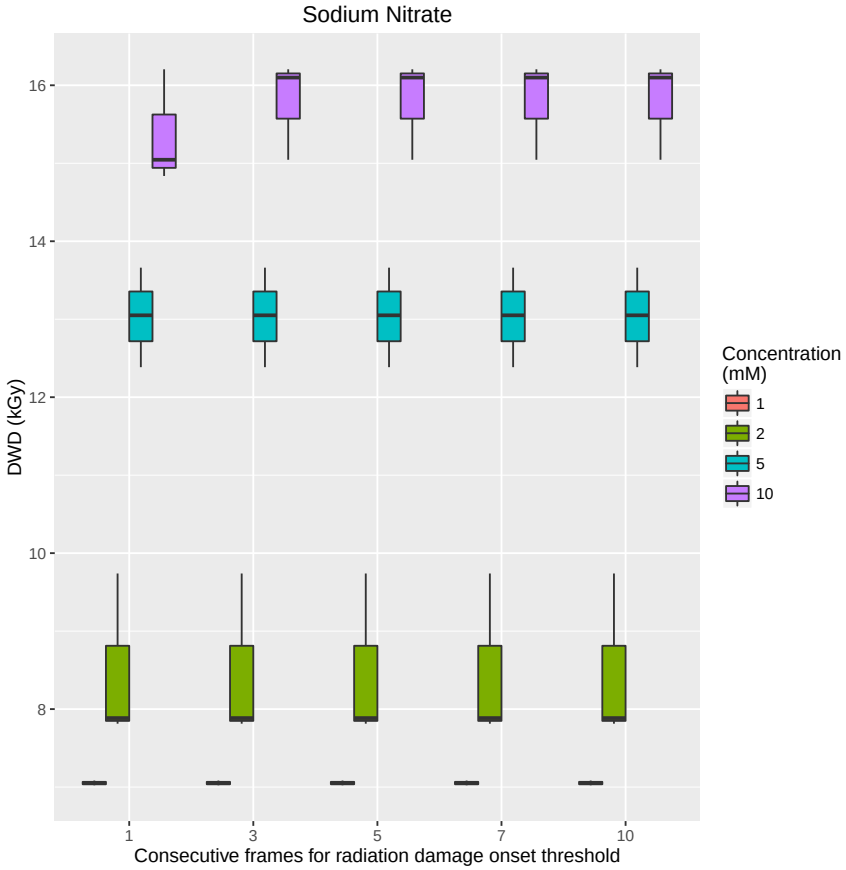
(b)


Ethylene Glycol
DWD (kGy)
25
20
15
10
5
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

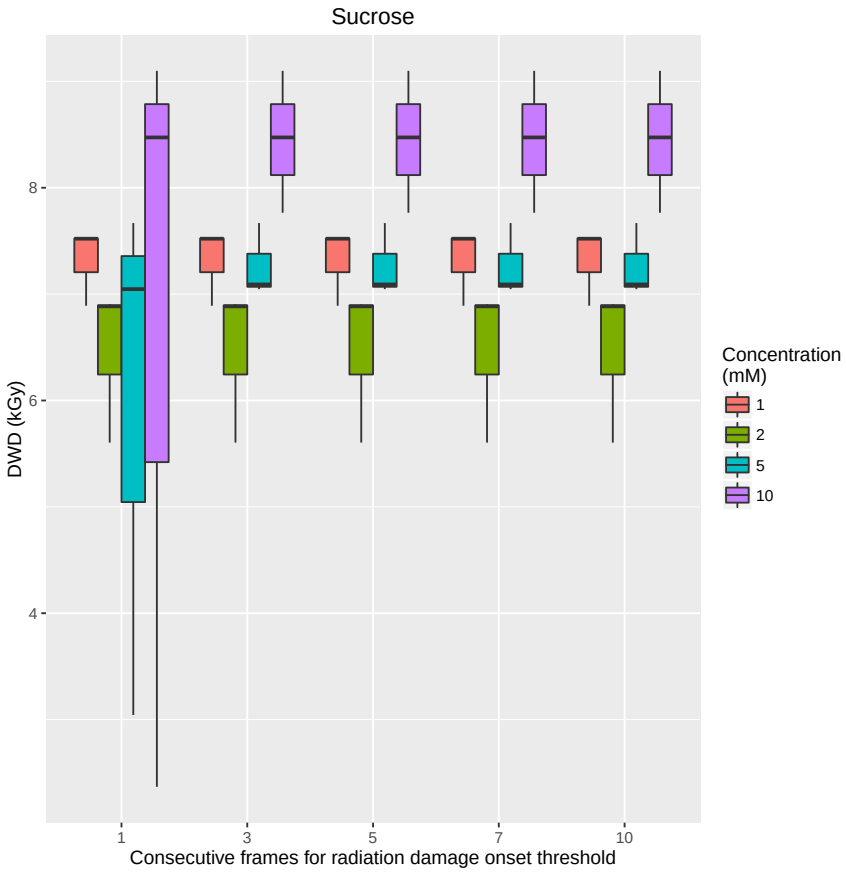
(c)


Glycerol
DWD (kGy)
40
30
20
10
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

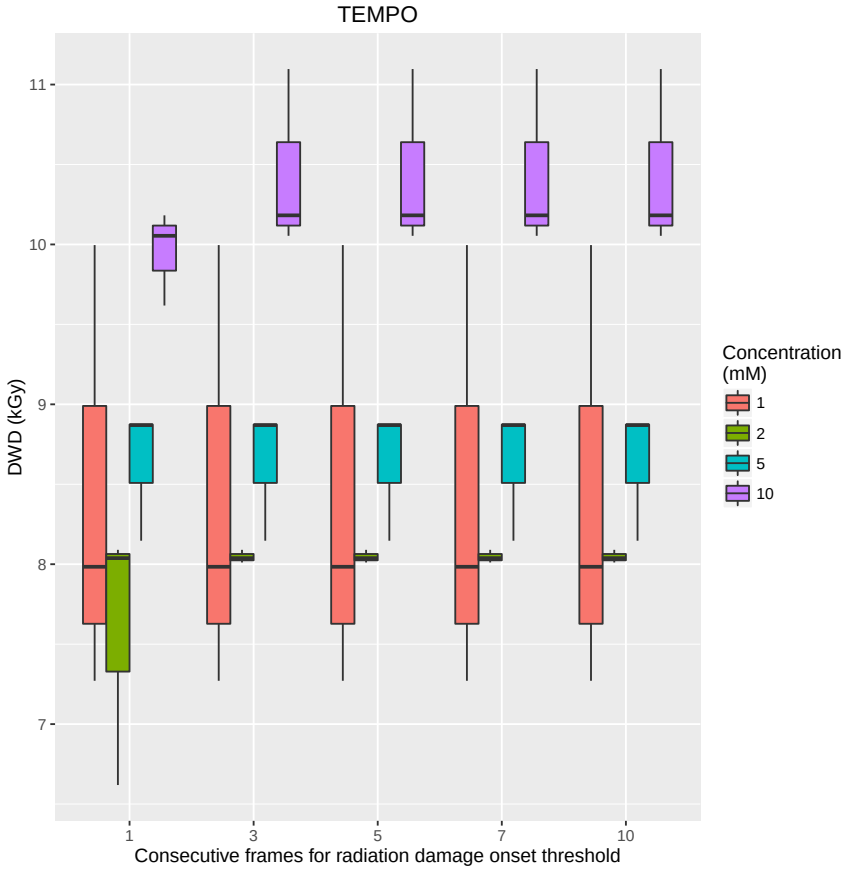
(d)


Sodium Nitrate
DWD (kGy)
16
14
12
10
8
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

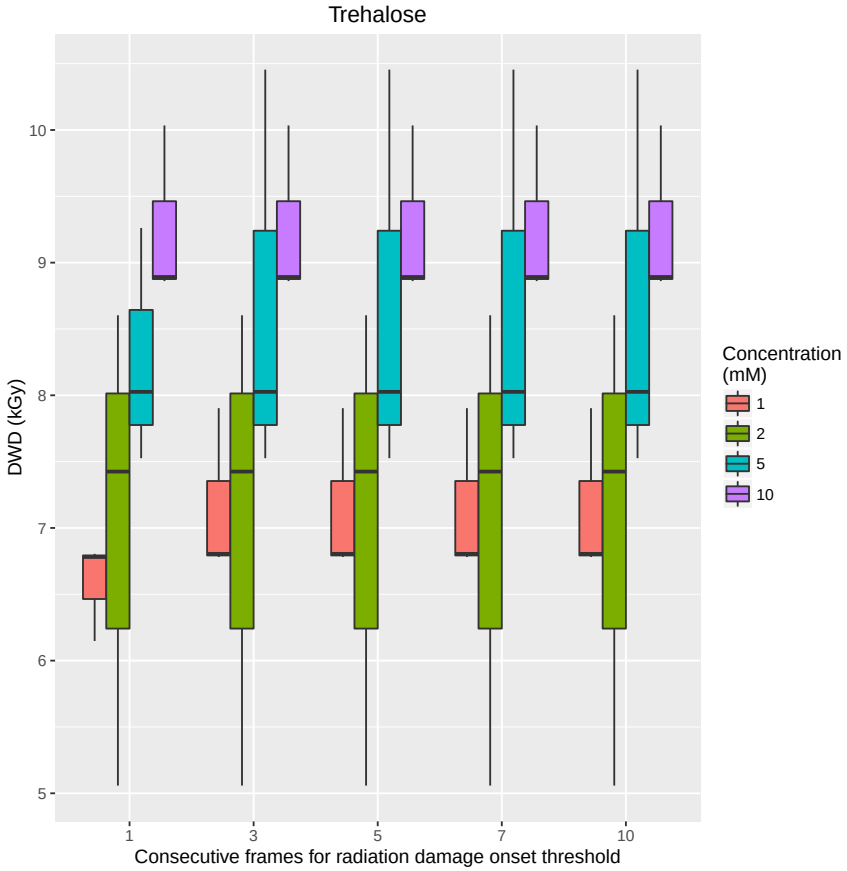
(e)


Sucrose
DWD (kGy)
8
6
4
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

(f)

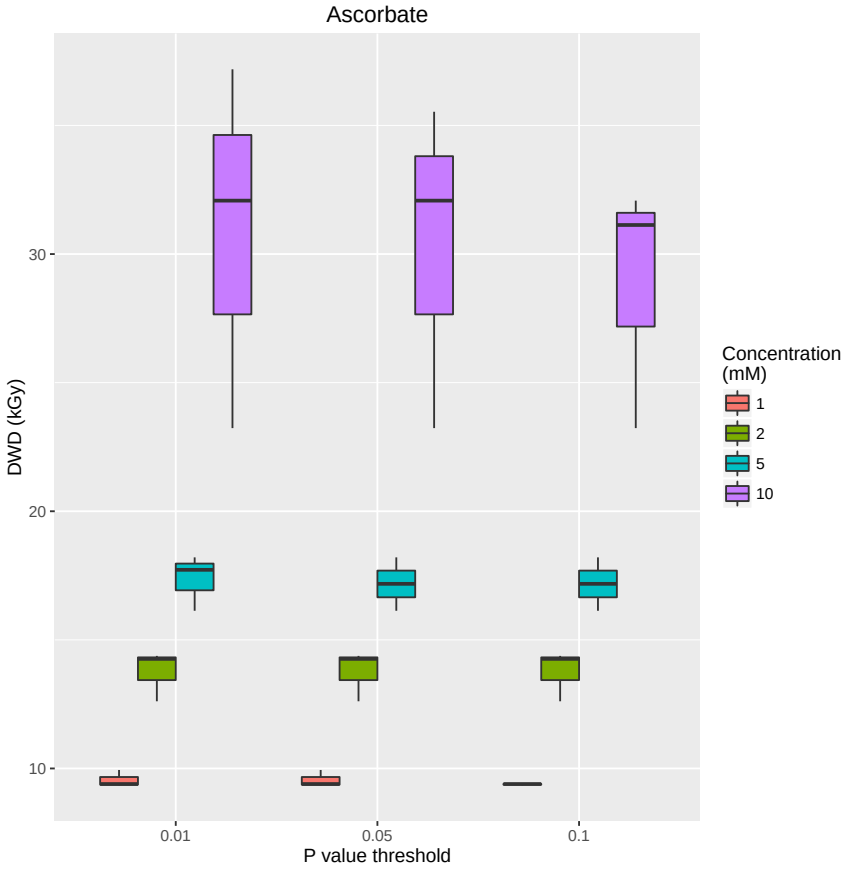


(g)

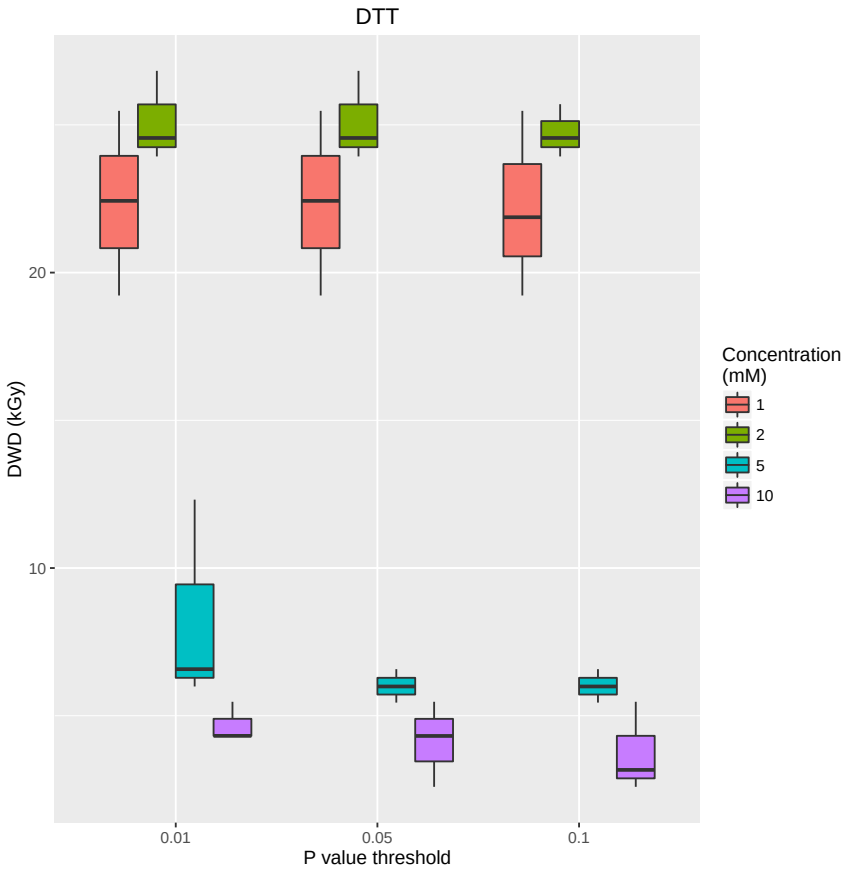


(h)

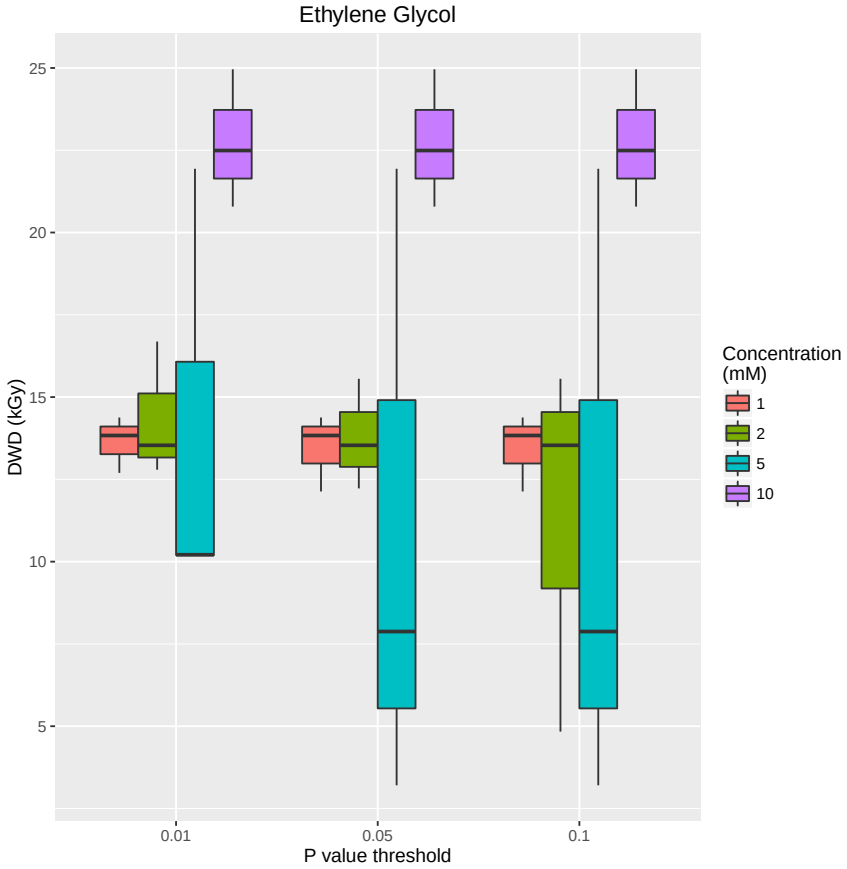
Figure 6.14: Dose at which significant radiation damage is determined to have occurred for different values of m, the number of consecutive dissimilar frames, for the 8 radioprotectants tested in Expt 2.


Ascorbate
DWD (kGy)
30
20
10
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

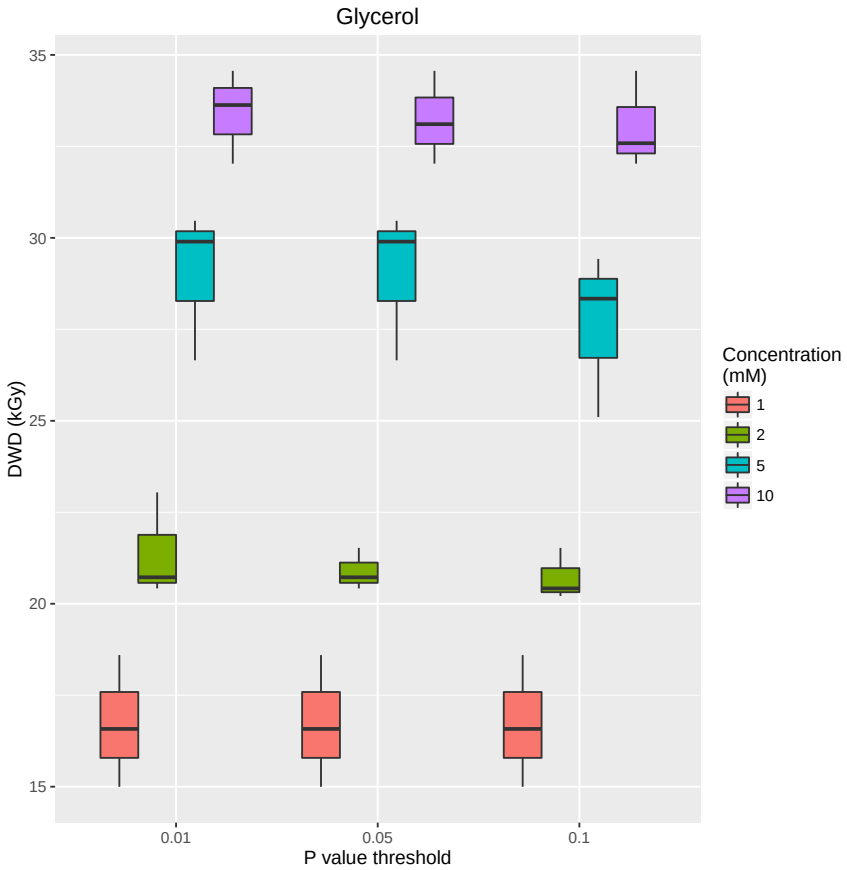
(a)


DTT
DWD (kGy)
20
10
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

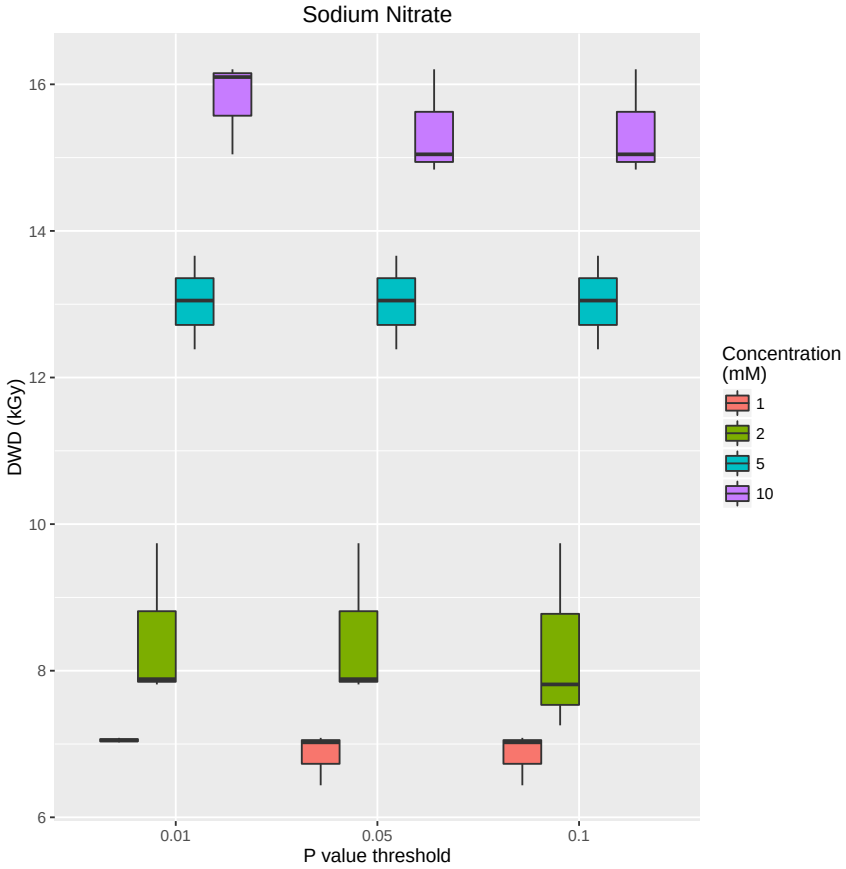
(b)


Ethylene Glycol
DWD (kGy)
25
20
15
10
5
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

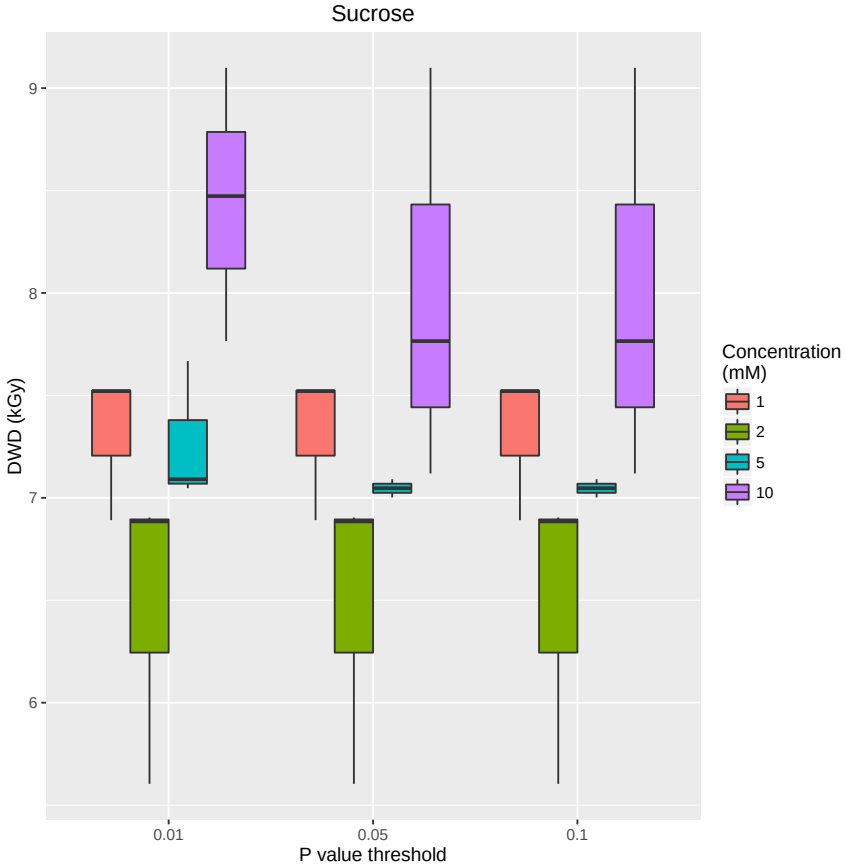
(c)


Glycerol
DWD (kGy)
35
30
25
20
15
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

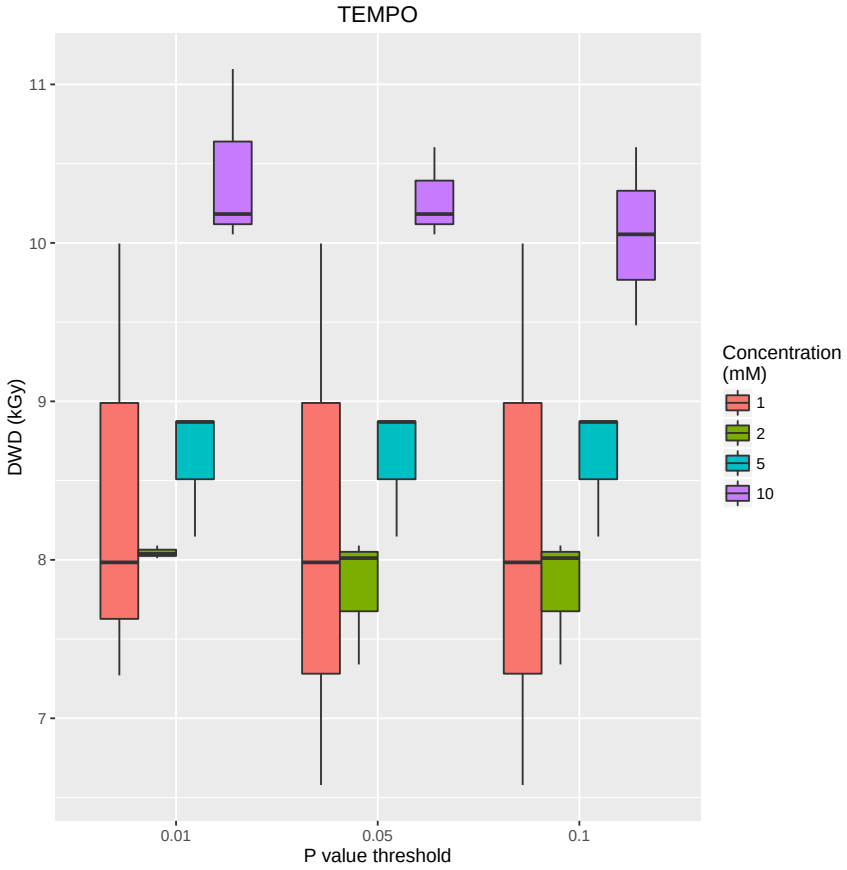
(d)


Sodium Nitrate
DWD (kGy)
16
14
12
10
8
6
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

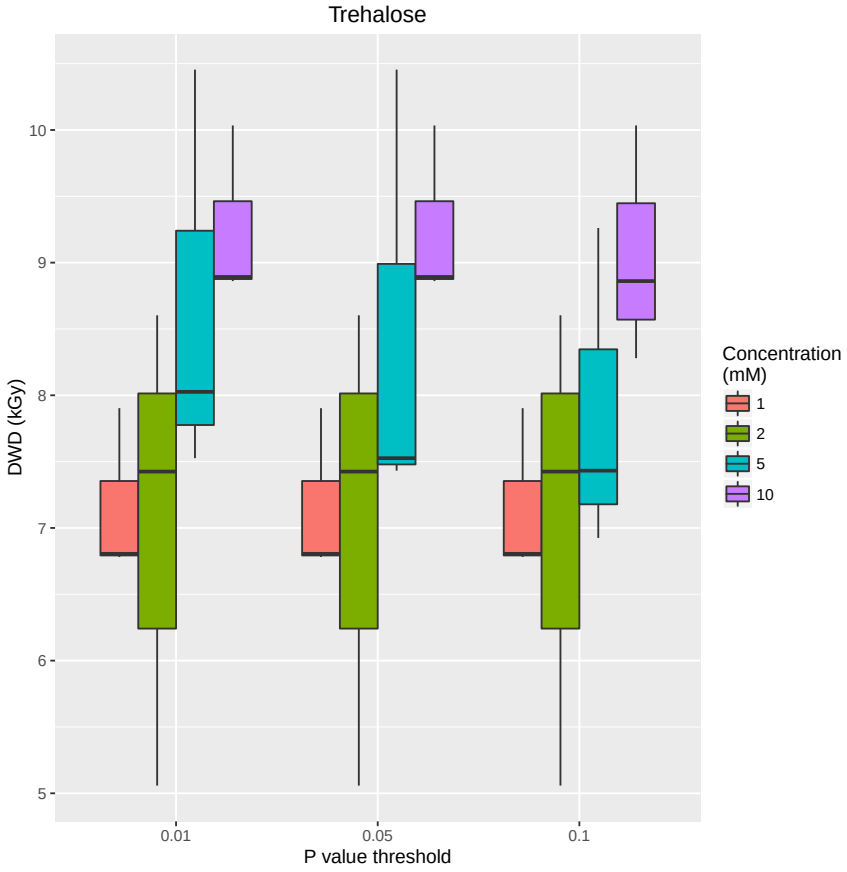
(e)


Sucrose
DWD (kGy)
9
8
7
6
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

(f)



(g)



(h)

Figure 6.15: Dose at which significant radiation damage is determined to have occurred for different values of α, the threshold probability value to determine frame similarity, for the 8 radioprotectants tested in Expt 2.

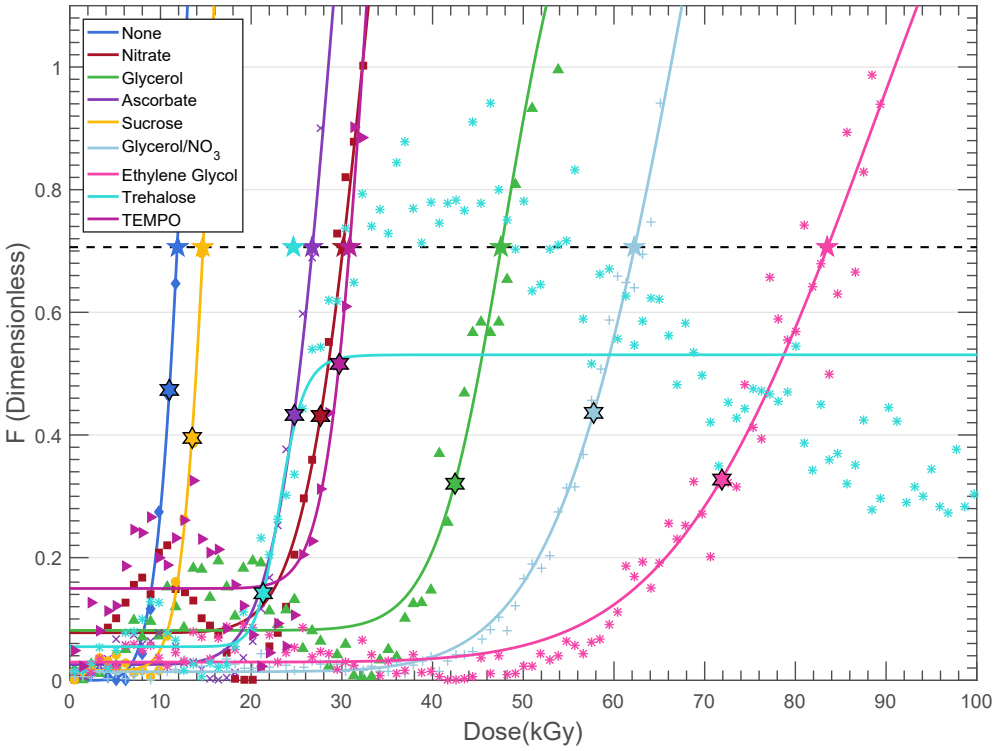


Figure 6.16: Fidelity values against dose for each radioprotectant along with their corresponding fitted logistic curves for the data collected in Expt 1. The five-pointed stars represent the points where the fitted logistic curves reach a fidelity value of 0.71. The six-pointed stars are the maximum curvature points of the fitted curves.

6.5 Results

6.5.1 Experiment 1

Figure 6.16 shows the results of the logistic curve fits to the fidelity data plotted in Figure 6.11. The five-pointed stars are the points where the fitted logistic curves reach a fidelity value of 0.71. This was an arbitrary value chosen because all samples reached that value. The six-pointed stars are the maximum curvature points of the fitted logistic curves. To compare the efficacies of each radioprotectant compound, the dose at which these points occur are plotted for each scavenger in Figure 6.17.

Figure 6.17a shows the dose taken to reach a fidelity value of 0.71 for each scavenger. Ethylene glycol is the most effective radioprotectant at 5 mM concentration, followed by the glycerol/sodium nitrate mixture and then glycerol alone. It is clear from these results that adding a radioprotectant allows useable data to be collected for longer because the sample is less sensitive to irradiation.

Similar conclusions are obtained from Figure 6.17b where the dose at which maximum curvature is reached is plotted for each radioprotectant. The results obtained using this metric give an identical relative efficacy ordering to the order seen when using the other metric. i.e. ethylene glycol is most effective, followed by the glycerol/sodium nitrate mixture etc. The main difference between the metrics is that the doses at which the thresholds are determined using the maximum curvature method are consistently lower than those found using the high value threshold.

6.5.2 Experiment 2

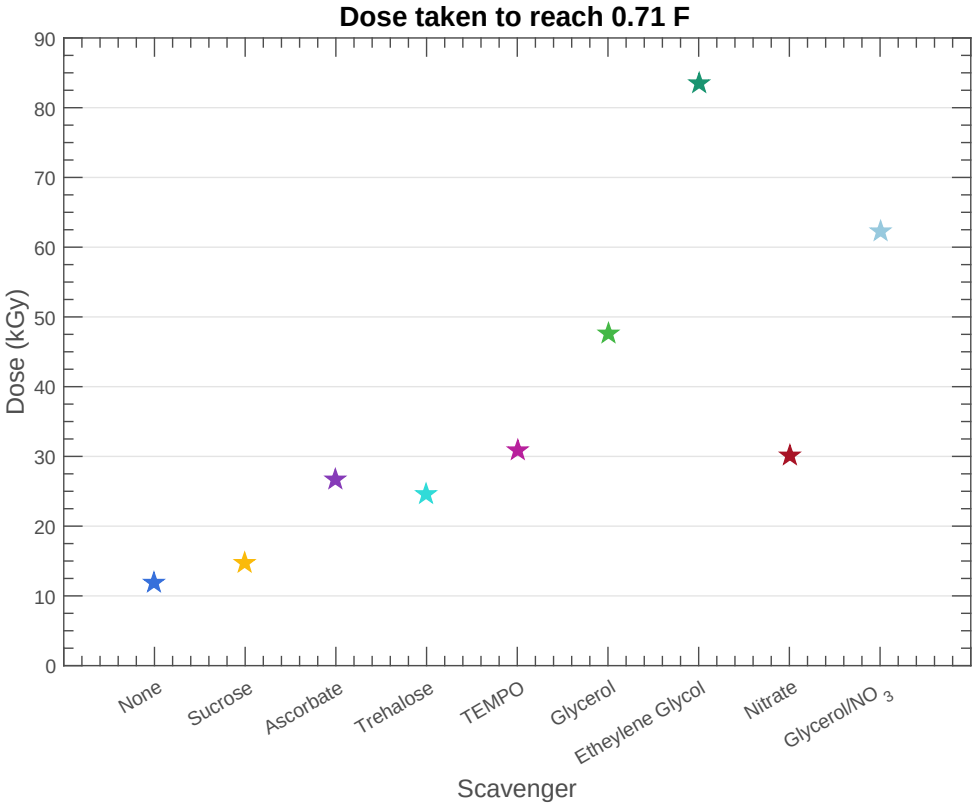
Comparing radiation damage onset metrics

In section 6.4.2 a metric for assessing the frame at which radiation damage had become significant was presented based on the CorMap test. Explicitly, this metric was defined as the point at which three consecutive frames were defined as dissimilar (m = 3) resulting from the CorMap test with threshold α = 0.01. This metric will be referred to as the *CMD* (CorMap Derived) metric.

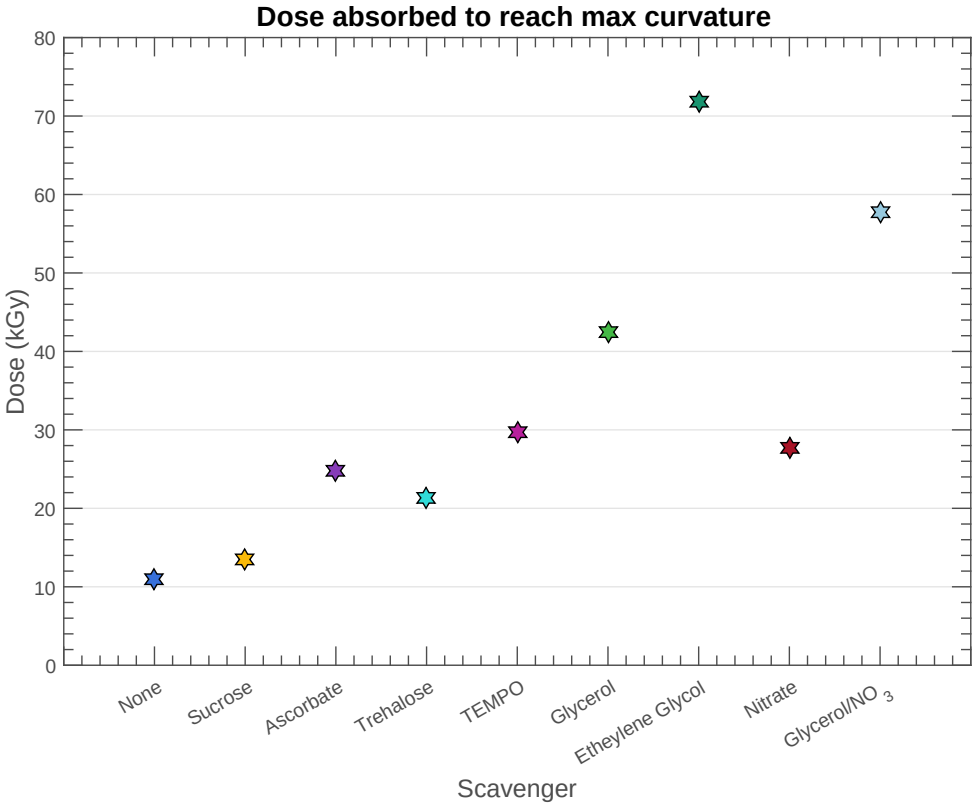
In addition to the *CMD* metric, an automatic data analysis pipeline at beamline BM29 is integrated into the beamline control system, *BsxCuBE*. This pipeline additionally performs analysis of frames and hence gives merging thresholds calculated for radiation damage onset using a metric henceforth denoted the *BsxCuBE* metric.

The results from analysis with both metrics were calculated/recorded and figure 6.18 shows how the two metrics compare for each compound at all concentrations. Generally the two metrics agree on the order of the efficacy of the various radioprotectant compounds, but the *BsxCuBE* metric always suggests that more frames can be merged then the *CMD* metric. Given that the *CMD* metric was developed to avoid individual dissimilar frames prematurely being flagged as the point of significant radiation damage onset, this result is quite surprising. It suggests that the *BsxCuBE* metric employs a very different method to assess the similarity of frames than the *CMD* metric.

The biggest discrepancy between the two metrics is the result for DTT. The *BsxCuBE* metric predicts a much higher dose tolerance than the *CMD* metric, especially for the 5 mM and

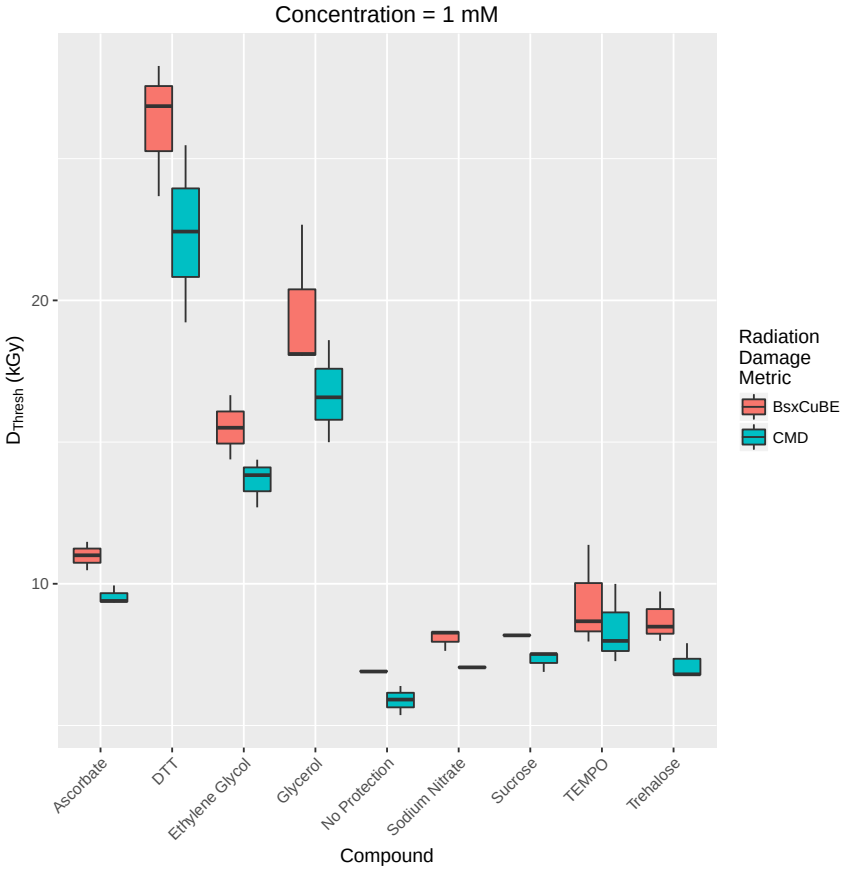


(a) Expt 1: Dose taken to reach a fidelity value of 0.71 for each scavenger.

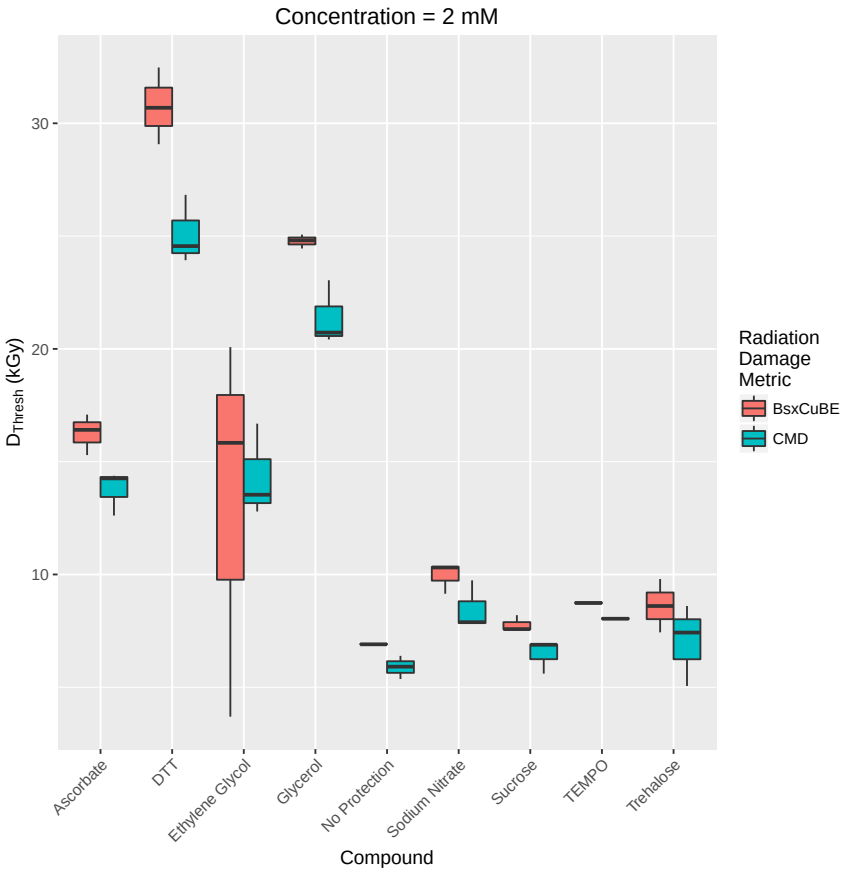


(b) Expt 1: Dose at which maximum curvature is reached

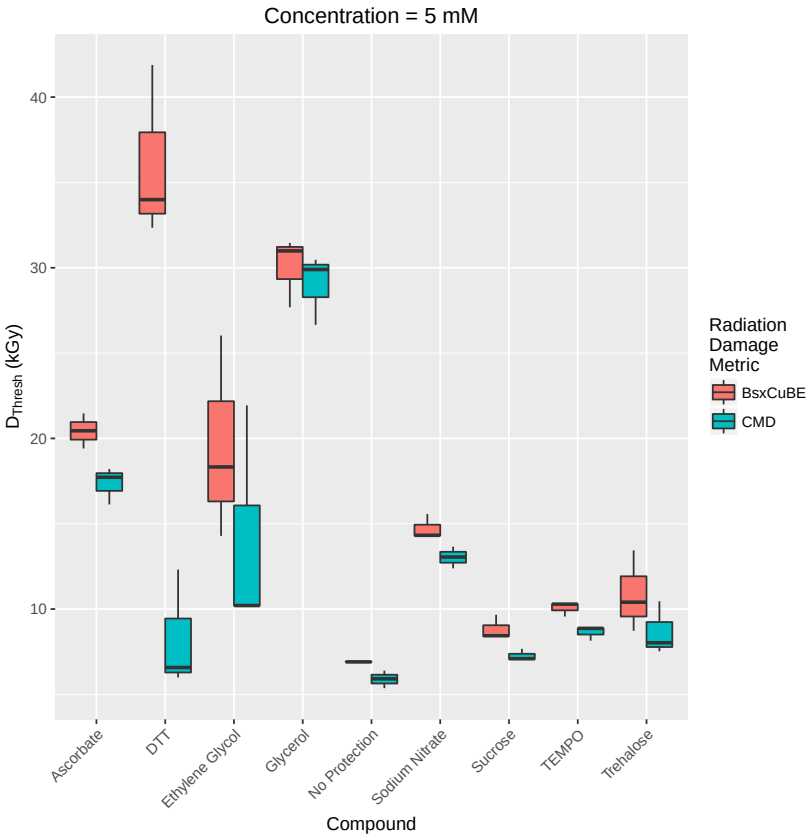
Figure 6.17


Concentration = 1 mM
D_{Thresh} (kGy)
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

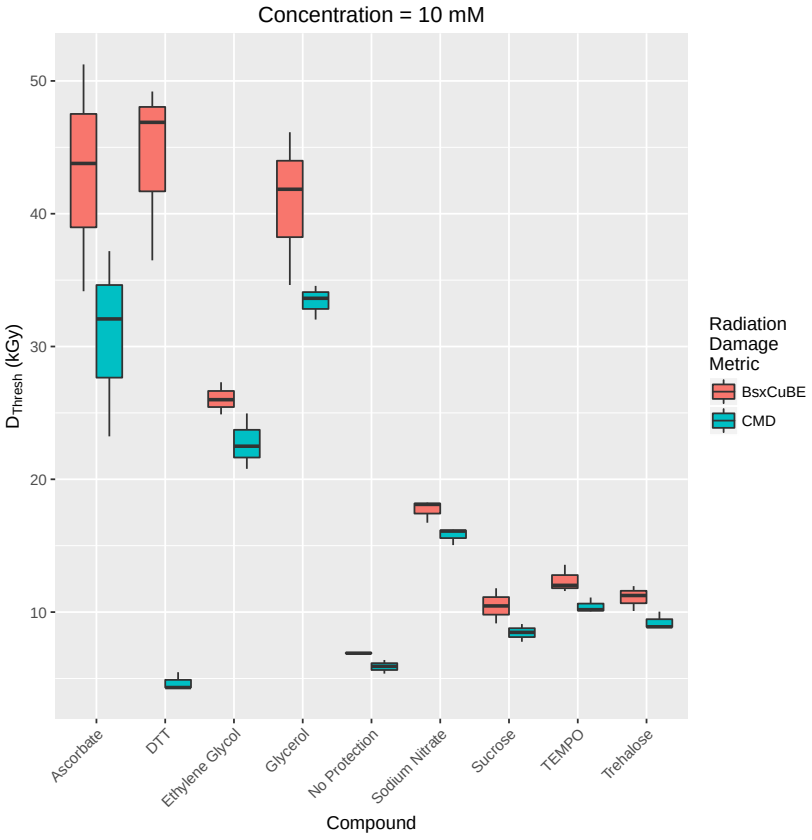
(a)


Concentration = 2 mM
D_{Thresh} (kGy)
30
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

(b)



(c)



(d)

Figure 6.18: Expt 2: Dose value at which radiation damage is considered significant for each concentration of radioprotectant used. Each box plot is created from the threshold dose values calculated for three different experimental runs of the same radioprotectant compound at a particular concentration. Pink boxes correspond to the *BsxCuBE* metric, blue boxes correspond to the *CMD* metric

10 mM concentrations. One reason thought to cause this discrepancy in the current work was the choice of m = 3 for the *CMD* metric. Figure 6.14b shows the results of the various values of m for DTT. Unlike the results for the other compounds (Figure 6.14), the chosen value of m could be the problem for 5 mM concentrations. However, this does not necessarily seem to be true for the 10 mM concentration. Thus it was necessary to re-evaluate some of the underlying assumptions of the analysis.

Pairwise comparisons with all frames

One assumption that is made in the merging analysis is that because the first frame suffers the least amount of dose, all subsequent frames should be compared to the first frame, the *reference frame*. If instead frames were compared to a different 'reference frame' would the threshold be different? The results shown in Figure 6.19 address this question, where the analysis was performed for the first repeat of GI sample with DTT added at 10 mM concentration. Along the y-axis are the frames which would be considered as the merging limit if the corresponding frame on the x-axis was chosen as the 'reference frame'. It can be seen that the radiation threshold value obtained using the *CMD* metric is highly dependent on the reference frame. This will greatly affect the conclusions that could be drawn from the analysis. From Figure 6.19, if the reference frame is 1, as is the case for the main analysis, then the frame at which radiation damage is considered significant is frame 8 (DWD = 4.32 kGy). However, if the reference frame was frame 7, then the threshold frame would be number 57 (DWD = 32.50 kGy). This value is closer to the value obtained from the *BsxCuBE* metric for that run (DWD = 46.88 kGy).

The fact that performing the pairwise comparisons with different reference frames gives varying results suggests that more information can be gained by analysing the results from all possible pairwise comparisons. Figure 6.20 is a heat map showing the results from all possible pairwise frame comparisons. The first row shows the results that were obtained using frame 1 as the reference frame. It shows the *CMD* metric highlighting frames becoming dissimilar very early on in the experiment (frame 8 out of 120). However, there is a large square region of similar frames in the top left of the map suggesting that the best results may be obtained by merging the frames from the larger region. This region notably does not include the first frame. The structure of this map could suggest that there are fast

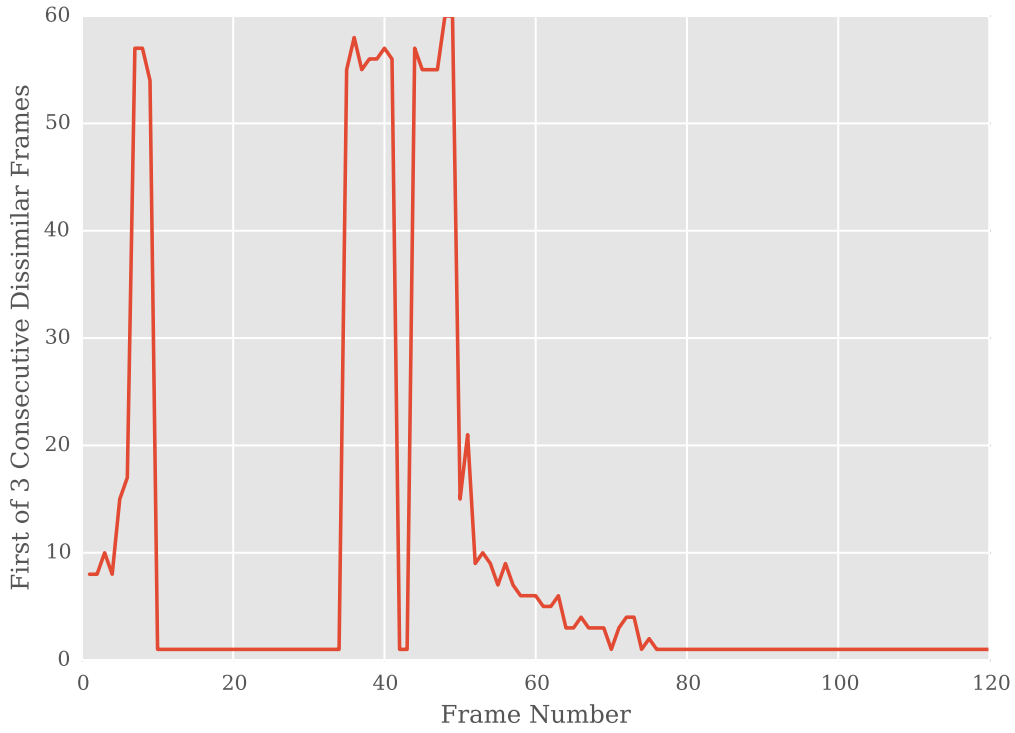


Figure 6.19: Expt 2: Radiation damage threshold frame against reference frame (m = 3) for the first repeat with DTT as the added radioprotectant at a concentration of 10 mM.

changes occurring in the sample early on in the experiment until a relatively stable molecular conformation is reached. Perhaps merging frames from these different regions may result in molecular envelopes resembling the molecules in different conformations. It may be the case that some of these states are functionally important. Another plausible explanation for these results could be that there were errors at the beginning of the measurements.

Using dose value and frame number as damage thresholds

Using the dose as the metric to evaluate the efficacy of the different radioprotectants means that the results are normalised for the energy absorption of the compounds. It is sensible to ask whether this normalisation changes the conclusions that would be drawn if the frame number alone was used as the threshold metric. Figure 6.21 shows the results for the 1 mM and the 10 mM concentrations of radioprotectants with dose value and frame number as the damage thresholds. For the more radiation tolerant samples (ascorbate, ethylene glycol, glycerol and even DTT) there is not much difference in using either the dose or the frame number. However, for the less efficient radioprotectants (sucrose, TEMPO and trehalose)

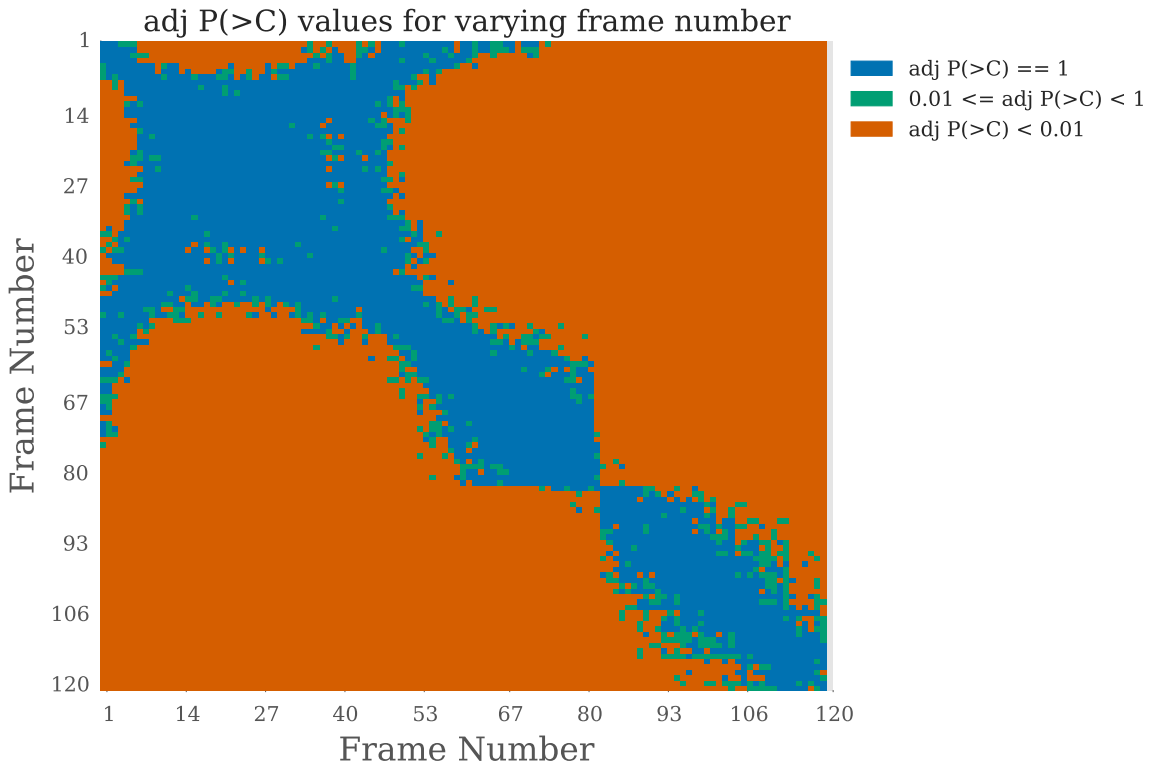
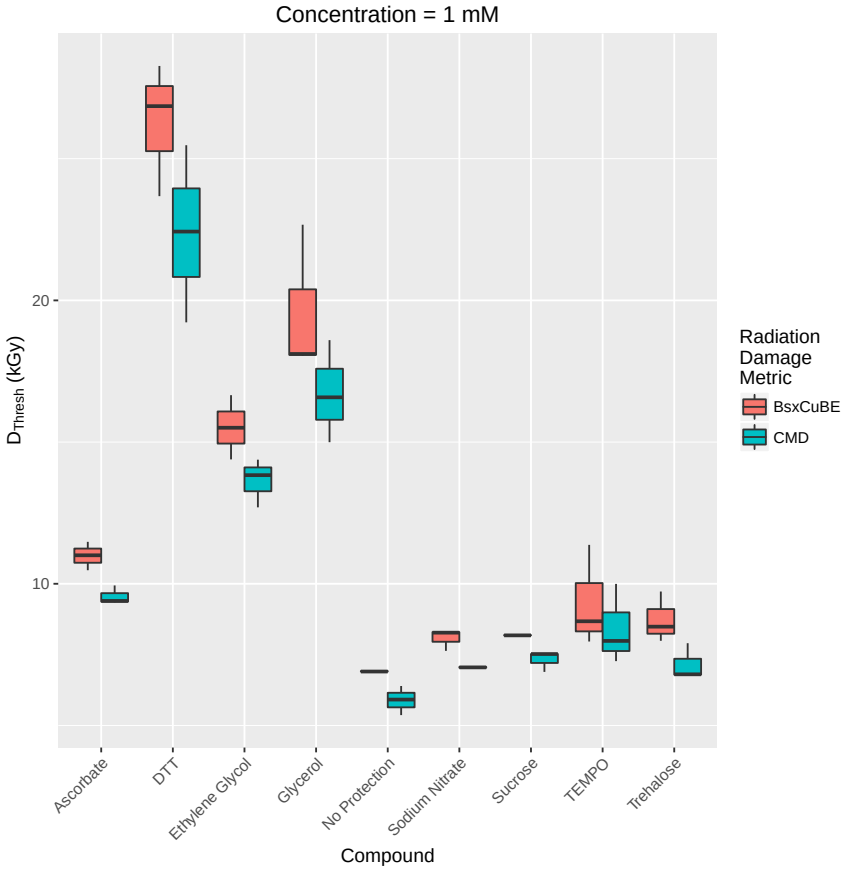


Figure 6.20: Expt 2: Heat map of all possible pairwise frame comparisons for the first repeat with 10 mM concentration DTT added. The y-axis represents the reference frame to which all other frames on the x-axis are compared. Blue - $P_{adj}(> C)$ = 1. Green - 0.01 $\leq P_{adj}(> C) <$ 1. Orange - $P_{adj}(> C) <$ 0.01.

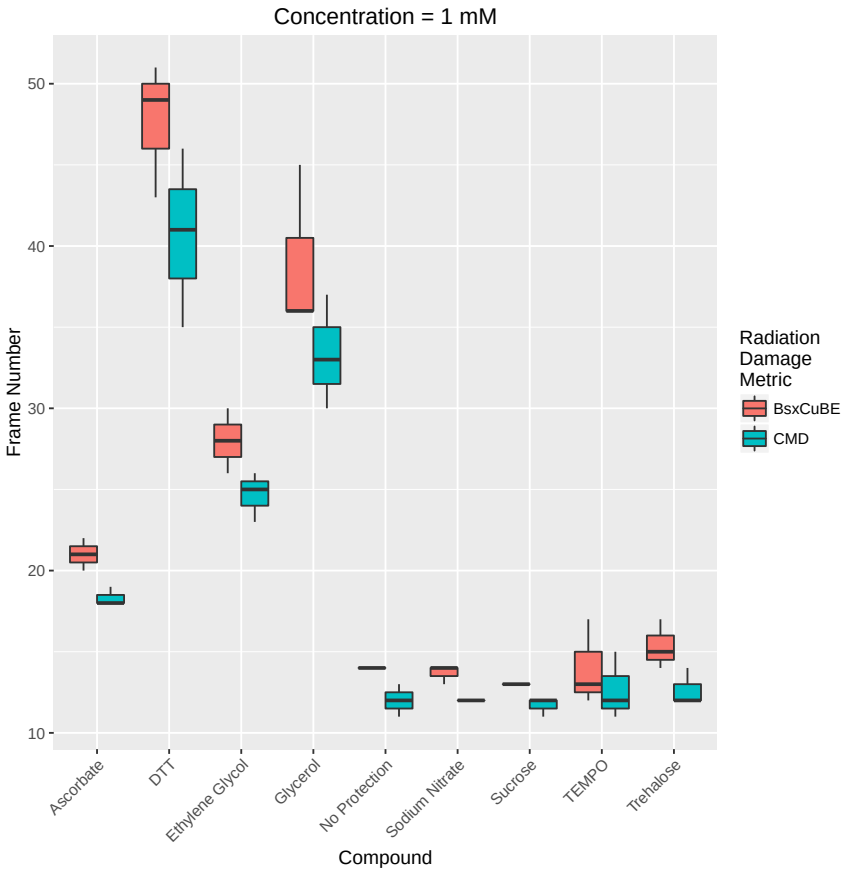
there is a significant difference. If the frame number is used (Figures 6.21b, 6.21d) the order of the relative efficacies of those radioprotectants is not too obvious. However, if the dose value is used instead (Figures 6.21a, 6.21c) then the order becomes much clearer, particularly at 1 mM concentration. The spread of the threshold values for those compounds is also smaller using dose instead of frame number for the 10 mM concentration.

Concentration dependence

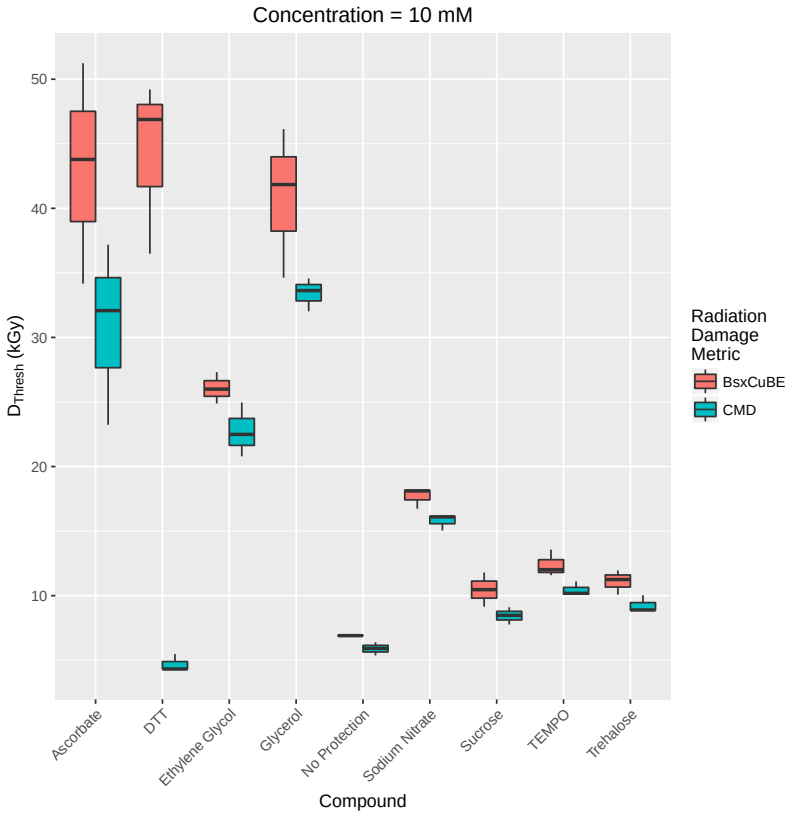
As a result of this work, a new metric, RD onset ratio, was developed to assess the change in radiation tolerance in the sample with the radioprotectant added compared to no protection. This metric is defined as the ratio of the median threshold value with added radioprotectant to the median threshold value with no protection using the *CMD* metric. Values below 1 correspond to a reduction in radiation tolerance whereas values above 1 show improved radiation tolerance. This metric was calculated for each compound at each concentration and the results are plotted in Figure 6.22. Significant concentration dependence can be ob-


Concentration = 1 mM
D_{Thresh} (kGy)
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

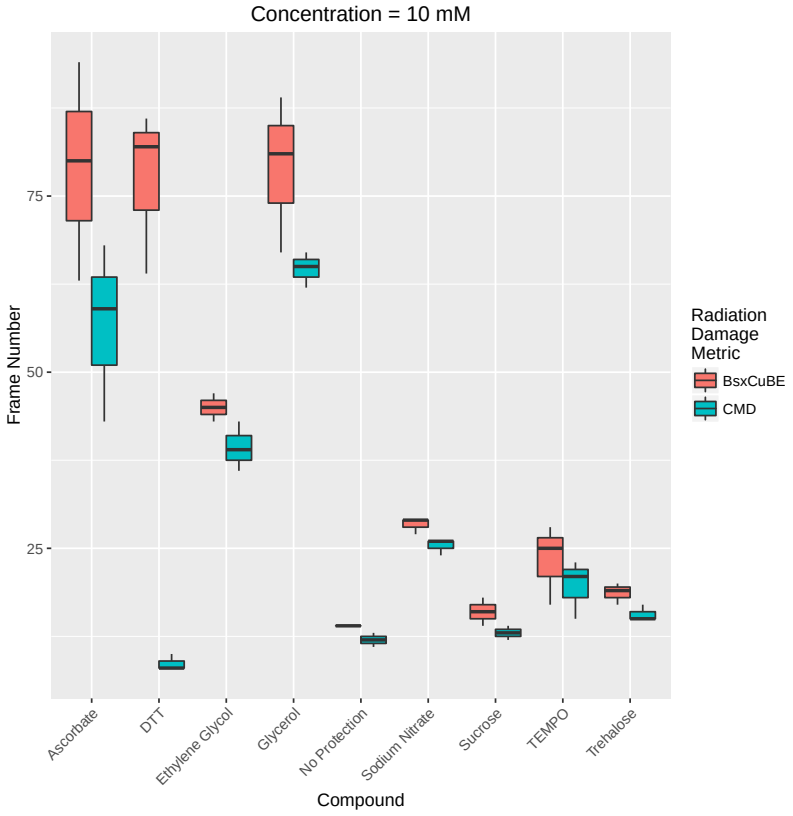
(a)


Concentration = 1 mM
Frame Number
50
40
30
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

(b)



(c)



(d)

Figure 6.21: Expt 2: Comparing dose and frame number as the metric by which to track radiation damage in SAXS for all SAXS repeats. Figures (a) and (c) use the dose value as the y-axis whereas (b) and (d) use the frame number. Each box plot is created from the threshold dose values for three different runs of the same radioprotectant compound. Pink boxes correspond to the *BsxCuBE* metric, blue boxes correspond to the *CMD* metric.

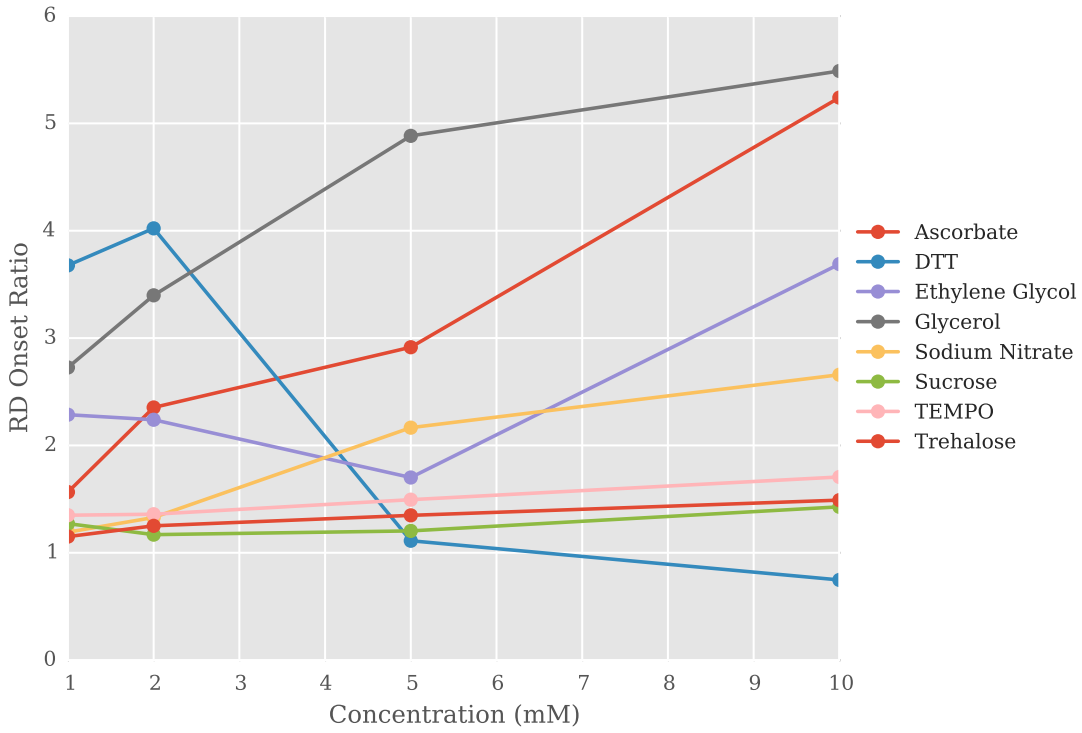


Figure 6.22: Expt 2: RD onset ratio against concentration for the 8 radioprotectants.

served for several radioprotectants. In particular, ascorbate, glycerol, and sodium nitrate all exhibit a strong positive concentration dependence i.e. the higher the concentration, the better the protection ability. DTT exhibits the opposite behaviour. At low concentrations DTT exhibits the highest ratio but this decreases at the higher concentrations. Sucrose, TEMPO and trehalose show a very small positive dependence but even at the highest concentration (10 mM) the RD onset ratio is less than 2. This suggests that these radioprotectants are not very efficient at increasing the dose tolerance of the sample. The RD onset ratio for ethylene glycol decreases as the concentration increases from 1 mM to 5 mM, but then there is a large increase at 10 mM. Thus the protection ability of ethylene glycol is not a simple monotonic function of the concentration.

Relative radioprotectant efficacy

The results shown in Figure 6.22 indicate that the most effective radioprotectant varies depending on the concentration of the compound used in the sample. The *BsxCuBE* and *CMD* metrics show agreement that at low concentrations (1 mM and 2 mM) DTT is the most effective radioprotectant. However, at higher concentrations these metrics disagree. At 5 mM

and 10 mM the *BsxCuBE* metric suggests that the most effective radioprotectant is still DTT. On the other hand, the *CMD* metric suggests that the most effective radioprotectant is glycerol. Furthermore, it also suggests that DTT is the least effective radioprotectant. It is important to take into account that using a different reference frame to perform pairwise comparisons can result in different conclusions with the *CMD* metric as shown in Figures 6.19 and 6.20.

The *CMD* metric result also disagrees with the results from Expt 1 where ethylene glycol was found to be the most effective radioprotectant at 5 mM. Furthermore, in Expt 1 ascorbate was the one of the least effective radioprotectants. It only outperformed sucrose and trehalose. Whereas in Expt 2, ascorbate was the second most effective radioprotectant when used at a 5 mM concentration. These results could be caused by the fact that the beam profile was not well characterised in the first experiment as opposed to the second.

An important difference between the results reported for Expt 1 and Expt 2 is that the dose thresholds for Expt 1 are consistently higher than for Expt 2. This is not surprising according to Figure 6.16 because all of the threshold values occur after the fidelity curves have already began to increase steeply away from the original similarity value. Whether the resulting statistics for the structural data can be improved beyond the threshold values given by the *CMD* metric remains to be seen, but this discrepancy may go some way to explaining why the *BsxCuBE* metric gives different results for the DTT data.

6.6 Discussion

RADDOSE-3D extensions to SAXS

The analysis of data from the two SAXS experiments has relied on extensions to RADDOSE-3D to allow it to simulate SAXS data collection. In particular the three major additions were:

1. implementation of a cylindrical sample geometry,
2. determination of the sample composition given a protein concentration,
3. attenuation of the X-ray beam due to a surrounding capillary.

The cylinder geometry (diameter and height) is now an explicit input option for the user to specify in a RADDOSE-3D job. RADDOSE-3D then converts the description of the cylinder provided by the user to a polyhedral geometry description, which is the base crystal implementation. In fact the current cuboid geometry implementation does exactly the same thing. Currently the only shape modelled by RADDOSE-3D that is not converted to a polyhedral representation for the simulation is the sphere.

Some of the RADDOSE-3D code was refactored so that the library of residue information (atomic composition, molecular weight) is now stored as a separate class to the atomic information. This refactoring reduced the complexity of the implementation for RADDOSE-3D to read a FASTA sequence file. The extension for RADDOSE-3D to read any other sequence file format should now be trivial.

RADDOSE-3D now implements a *container* class which defines the material that houses the sample. The attenuation of the beam by the container is considered for all experiments. For SAXS experiments this is usually a quartz capillary, whereas for MX it is assumed to be a container that is *transparent* to X-rays. However, the user can define any container for any experiment, SAXS or MX, giving greater flexibility. An immediate application is for *in situ* RT MX data collection.

As mentioned in section 6.2.1, RADDOSE-3D has been written in a very modular style so that extensions to existing functionality can be easily incorporated. At its core, RADDOSE-3D essentially deals with three main objects: a sample, a beam and a wedge. The work on

RADDOSE-3D presented here shows how the "crystal" description was extended to incorporate liquid SAXS samples, which ultimately depends on the input specification. In a similar manner the beam, which currently describes incident X-ray photons, could be extended to represent other beam types (e.g. an electron beam) if its properties are defined differently.

One of the current drawbacks of the extension of RADDOSE-3D to simulate SAXS experiments is that the implementation does not yet support sample flow. This is a common method used in SAXS experiments to mitigate radiation damage (**?**). This sample flow results in any given small volume of fluid passing across the X-ray beam as described in section 6.2.6. In RADDOSE-3D this situation reduces to how long a given volume of fluid is exposed to each region of the X-ray beam, which will be dependent on the radial position from the centre of the tube due to the Poiseuille flow velocity profile (**?**). In order to implement this method of data collection in RADDOSE-3D, a sufficiently small discretisation of time would be necessary so that the fluid volume does not completely 'pass' the entire beam in a single time step. Care would be needed to avoid the case where the time discretisation is so small that the computational time is unnecessarily long.

Additionally diffusive turnover, as described in section 6.2.6, could be incorporated into the dose calculations. However, beam sizes in SAXS experiments are typically quite large and exposure times are not long enough for the diffusive turnover to cause significant effect (500 μm × 250 μm with maximum exposure time 141 seconds (RT) (**?**), 220 μm × 190 μm with 60 seconds exposures (100 K) (**?**)). Therefore this would not be a high priority extension.

The formal tests for the SAXS extensions still need to be written before the code is officially released. These include testing that the cylindrical implementation is stable for different inputs and is aligned with the axial length perpendicular to the beam by default. It is also necessary to check that the absorption coefficient calculation works for different types of elemental and mixture compositions. However, the code is currently public on the RADDOSE-3D Github repository in the branch named "SAXS" `https://github.com/GarmanGroup/RADDOSE-3D`. Alpha versions of the SAXS extensions have already been distributed to Dr. Edward Snell (Hauptman-Woodward Institute, Buffalo) and Dr. Adam Round (ESRF, Grenoble) for user testing.

Radioprotectant analysis

Two experiments to compare the efficacy of different radioprotectants for SAXS samples have been carried out and analysed. Although methods already exist to assess the similarity of frames in a pairwise manner, few established methods deal with determination of a threshold for significant radiation damage onset. Establishing a threshold was necessary to compare the efficacy of various radioprotectants. Therefore metrics were developed to establish thresholds for merging analysis so radioprotectant efficacy could be compared using the data that were collected.

For experiment 1 the analysis performed by DATCMP gave a fidelity value which was a measure of the similarity between frame 1 and a subsequent frame. Overall these frames resembled logistic curves, and hence 4PLs were fitted to the data. Two metrics were then developed to give a merging threshold. The first metric was the total dose absorbed to reach an arbitrary fidelity value. This metric is fundamentally flawed because there is no universal value to choose and it would have to change for each comparison experiment. It is also a relative metric i.e. it only compares the efficacy of one radioprotectant with another, but it does not give any absolute stand-alone information about whether the frames at that fidelity value are damaged or not.

The second metric developed was the dose absorbed before reaching maximum curvature. This metric is absolute in the sense that the threshold represents the point at which the fidelity values begin to deviate maximally from the initial frame(s). A downside to the maximum curvature metric is that if it is being used to ultimately determine which frames to merge, then the value that is determined may be too strict. If the overall curvature of the function is fairly small then it is possible that merging some frames beyond the threshold may still improve data quality because they may not be significantly dissimilar.

Potentially a big problem with the fitting approach overall is that it relies on a parametric function to be fitted to data where there is no guarantee that the data obey that particular form.

More recently, the DATCMP similarity analysis was extended to implement the CorMap test which assesses the similarity of two frames. As this is the default test in DATCMP, it is likely that this is the analysis that the SAXS community will adopt. During the analysis of the

SAXS data, there was no open source package available to generate the plots related to the CorMap analysis, including the pairwise CorMap plots themselves (Figures 6.12b and 6.12d). Therefore a library was written to perform all analysis and also to produce the plots. This library is currently available on Github `https://github.com/JonnyCBB/AnalyseSAXSLib/blob/master/CorMapAnalysis.py` but is intended to be released as a distributable Python package.

On its own, the CorMap test performs pairwise comparisons of frames. This means that an individual frame can be detected as being dissimilar to another frame regardless of the similarity assessment of frames immediately before and after the frame in question. This dissimilarity of an individual frame therefore may not indicate a true systematic change in the molecules of the sample. It may just be an outlier. Therefore a more robust metric was developed for this work whereby three consecutive dissimilar frames were detected to establish a merging threshold. It was used to determine that there is a concentration dependence on the efficacy of individual radioprotectants. Some radioprotectants such as glycerol, ascorbate and sodium nitrate exhibited a positive concentration dependence, whereas DTT showed a negative concentration dependence. It was also found that glycerol, ascorbate and ethylene glycol were the most effective scavengers at high concentrations (10 mM). These compounds are all hydroxyl scavengers ($k = 1 - 5 \times 10^9\ M^{-1}\ s^{-1}$ (**?**)), a radical species which is known to be mobile at room temperature. On the other hand sodium nitrate, which is an electron scavenger ($k = 10^8 - 10^9\ M^{-1}\ s^{-1}$ (**??**)), generally does not protect as well. DTT was found to be the most effective radioprotectant at the lower concentrations of 1 mM and 2 mM, which were tested.

One aspect that was not considered in the analysis is how the radioprotectants affect the environment of the sample. For example, it is known that DTT reduces disulphide bonds and undergoes oxidation. Glycerol on the other hand, is known to increase the noise and hence reduce the observable signal obtained from the sample (**?**). These are considerations that the experimenter must also take into account when deciding on which radioprotectant compound to use.

The heatmap shown in Figure 6.20 displays some exciting possibilities for data analysis. Firstly, it shows that pairwise comparisons with the first frame alone may not necessarily be the best approach to get all of the information possible from SAXS data. Secondly it

shows distinct regions of frame similarity. These regions may correspond to distinct and functionally important molecular conformations that would otherwise not be noticed. This finding warrants further analysis to determine how physiologically significant these regions are.