



***RIDL*: a tool to investigate radiation-induced density loss**

Charles Simon Bury and Elspeth Frances Garman

J. Appl. Cryst. (2018). **51**, 952–962



IUCr Journals
CRYSTALLOGRAPHY JOURNALS ONLINE

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site or institutional repository provided that this cover page is retained. Republication of this article or its storage in electronic databases other than as specified above is not permitted without prior permission in writing from the IUCr.

For further information see <http://journals.iucr.org/services/authorrights.html>



RIDL: a tool to investigate radiation-induced density loss

Charles Simon Bury* and Elspeth Frances Garman*

Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK. *Correspondence e-mail: csbury@me.com, elspeth.garman@bioch.ox.ac.uk

Received 20 December 2017

Accepted 3 April 2018

Edited by A. R. Pearson, Universität Hamburg, Germany

Keywords: specific radiation damage; electron density loss; Fourier difference maps; macromolecular crystallography; *RIDL*.

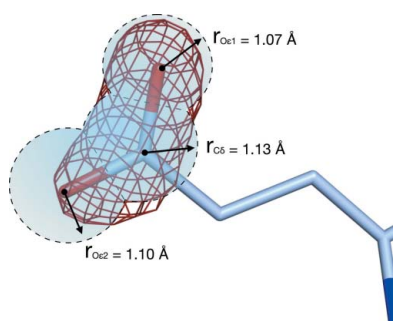
Supporting information: this article has supporting information at journals.iucr.org/j

An automated tool, *RIDL* (*Radiation-Induced Density Loss*), has been developed that enables user-independent detection and quantification of radiation-induced site-specific changes to macromolecular structures as a function of absorbed dose. *RIDL* has been designed to extract suitable per-atom descriptors of radiation damage, based on changes detectable in $F_{\text{obs},n} - F_{\text{obs},1}$ Fourier difference maps between successive dose data sets. Subjective bias, which frequently plagues the interpretation of true damage signal *versus* noise, is thus eliminated. Metrics derived from *RIDL* have already proved beneficial for damage analysis on a range of protein and nucleic acid systems in the radiation damage literature. However, the tool is also sufficiently generalized for improving the rigour with which biologically relevant enzymatic changes can be probed and tracked during time-resolved crystallographic experiments.

1. Introduction

With the wide use of existing third generation synchrotrons, and the current advent of higher flux density fourth generation sources, radiation damage is predicted to remain a major hindrance to the success of the macromolecular X-ray crystallographic (MX) pipeline. Even before any observable decay in the recorded diffraction intensities, at absorbed doses of just several MGy ($1 \text{ MGy} = 10^6 \text{ J kg}^{-1}$) radiation-induced changes to atom oxidation states, bond scission and conformational changes have been reported to rapidly occur for a range of model protein crystals at 100 K (Burmeister, 2000; De la Mora *et al.*, 2011; Nanao *et al.*, 2005; Ravelli & McSweeney, 2000). When undetected, such site-specific damage events inherently lead to modelling inaccuracies; this is of particular importance to structural biologists since active site regions appear to be distinctly radiation sensitive (Dubnovitsky *et al.*, 2005; Clavel *et al.*, 2016; Berglund *et al.*, 2016). In its most severe form, unaccounted damage can ultimately lead to the derivation of false enzymatic pathways (Matsui *et al.*, 2002).

Site-specific damage has traditionally been studied by collecting multiple complete, non-overlapping MX diffraction data sets from a single crystal, to observe how the electron density evolves as a function of dose (Helliwell, 1988). $F_{\text{obs},n} - F_{\text{obs},1}$ Fourier difference maps (where $n > 1$ is the index of a higher dose data set) allow changes in electron density with dose to be directly visualized around individual atoms. In contrast to the monitoring of atomic *B* factors or xyz coordinates with increasing dose, $F_{\text{obs},n} - F_{\text{obs},1}$ maps are far less tied to the quality of the refined model: a consideration which becomes particularly important at higher absorbed doses, when the reduced quality of diffraction data can lead to difficulties in the model building and refinement process.



$F_{\text{obs},n} - F_{\text{obs},1}$ maps have proved an invaluable tool for establishing a reproducible ordering of such specific damage events for non-metalloprotein crystals at 100 K (Weik *et al.*, 2000; Ravelli & McSweeney, 2000): (a) disulfide bond elongation and cleavage, (b) glutamate and aspartate decarboxylation, and (c) methionine methylthiol group loss. Cleavage of the aromatic OH group from tyrosine has also been suggested to occur in proteins (Burmeister, 2000), but the existence of such events has recently been refuted (Bury *et al.*, 2017). It is important to note that other protein moieties also undergo radiation-induced changes at 100 K [e.g. selenomethionine (Leiros *et al.*, 2006), lysine (Juers & Weik, 2011) and histidine (Fioravanti *et al.*, 2007), the latter two of which underwent conformational change as opposed to bond rupture], and despite the reported radiation insensitivity of nucleic acids in crystals at 100 K, difference density consistent with phosphodiester bond rupture has also been detected at high doses (≥ 20 MGy) (Bury *et al.*, 2015). Consequently, the entire coordinate model must be visually inspected for indications of damage with subjective judgement made as to its significance. For larger, more complex structures, however, this process quickly becomes intractable. This issue is compounded by the fact that the maps are inherently noisy and this worsens because of the effects of global damage with increasing dose.

An interest in data collection above 100 K, particularly at room temperature (RT) (Owen *et al.*, 2012; Keedy *et al.*, 2015), has also re-emerged in recent years, because of the potential loss of functionally relevant protein conformational states as a result of the cryo-cooling process (Fraser *et al.*, 2011). This has been facilitated by advances in detector technology, particularly regarding reductions in readout time (Owen *et al.*, 2014). An abundance of radical species originating from the irradiated solvent channels are known to be mobile above cryo-temperatures, and the dominant damage pathways are predicted to be quite distinct from those at 100 K (Juers & Weik, 2011). Difference maps have now enabled the location of disulphide cleavage events in RT data collection (Southworth-Davies *et al.*, 2007; Kmetko *et al.*, 2011); however, because of the fast deterioration of diffraction data quality with dose, a comprehensive characterization of the radiation chemistry in protein crystals above 100 K is still lacking.

Despite their utility, a challenge still remains in how to best extract damage information from $F_{\text{obs},n} - F_{\text{obs},1}$ maps. In practice, the difference map is stored as a finite grid of difference density values (referred to as voxels) expanding over a pre-determined volume (typically the unit cell or asymmetric unit), with a mean map value of $\sim 0 \text{ e } \text{\AA}^{-3}$ and standard deviation σ . To date, a standard practice in the radiation damage field has been to quote the σ height (the number of standard deviations that a peak value is above/below the map mean) of $F_{\text{obs},n} - F_{\text{obs},1}$ map peaks and holes in close proximity to protein residues as a metric for damage to each residue (e.g. Borek *et al.*, 2007; Fioravanti *et al.*, 2007). Unfortunately, peak σ heights cannot be rigorously compared between different MX radiation damage studies or even for different dose data sets collected from a single crystal, since

the value of σ is entirely data set specific. It is also non-trivial to suitably assign a σ threshold above which to screen for damage sites without the risk of neglecting low-lying peaks that represent true damage events. Moreover, detection of map peaks also provides no quantitative information on electron density change for the majority of atoms that are not adjacent to well defined difference peaks.

The program *RIDL* (*Radiation-Induced Density Loss*) has been developed to automate the process of creating and interpreting $F_{\text{obs},n} - F_{\text{obs},1}$ maps generated from an MX radiation damage series, and to provide a tool for analysing site-specific radiation damage that is free from user bias. *RIDL* samples the local region of the $F_{\text{obs},n} - F_{\text{obs},1}$ map surrounding each refined atom in a macromolecular coordinate file, and defines purposely designed per-atom summary metrics to describe the radiation-induced electron density change to individual atoms. Such an approach thus enables the relative measurement of radiation-induced density changes for every refined atom in a structure, as opposed to only atoms in close proximity to well formed and user-interpreted $F_{\text{obs},n} - F_{\text{obs},1}$ map peaks above a predetermined σ threshold. The subjective bias of the researcher is thus eliminated.

RIDL has been written so that it is sufficiently generalized to be used for further site-specific radiation damage studies, but is also designed to be applicable to the growing field of time-resolved crystallography, in which subtle changes in electron density must frequently be tracked across successive time-stamped data sets.

2. Description of program

2.1. The pipeline flow and modular structure

RIDL is written in Python (compatible with Python 2.x and 3.x), to allow cross-platform compatibility and so that it can be run directly from the command line together with a pre-defined plain-text input file. At runtime, *RIDL* performs three discrete tasks sequentially: firstly, *RIDL* utilizes a range of widely used *CCP4* suite programs (Winn *et al.*, 2011) to generate $F_{\text{obs},n} - F_{\text{obs},1}$ maps and to determine a local region of interest around each refined atom within the structure (the search radius), for each data set $n > 1$. *RIDL* then proceeds to calculate per-atom summary metric values for each higher dose data set ($n > 1$) sequentially. Finally, *RIDL* outputs an HTML-format summary file containing concise damage analysis for the input data. The logical flow for the complete pipeline is provided in Fig. 1. The following sections will discuss the construction of this pipeline.

As described below, the use of *RIDL* has been demonstrated for the example case of a crystal of a glycoside hydrolase family 7 (GH7) cellobiohydrolase from *Daphnia pulex*. An MX damage series (2 Å) has been used, consisting of 11 increasing dose data sets between 1.11 and 22.7 MGy, that has recently been deposited in the Protein Data Bank (PDB) (Bury *et al.*, 2017). The reported dose values are diffraction-weighted doses (Zeldin, Brockhauser *et al.*, 2013) calculated using *RADDOSE-3D* (Zeldin, Gerstel *et al.*, 2013),

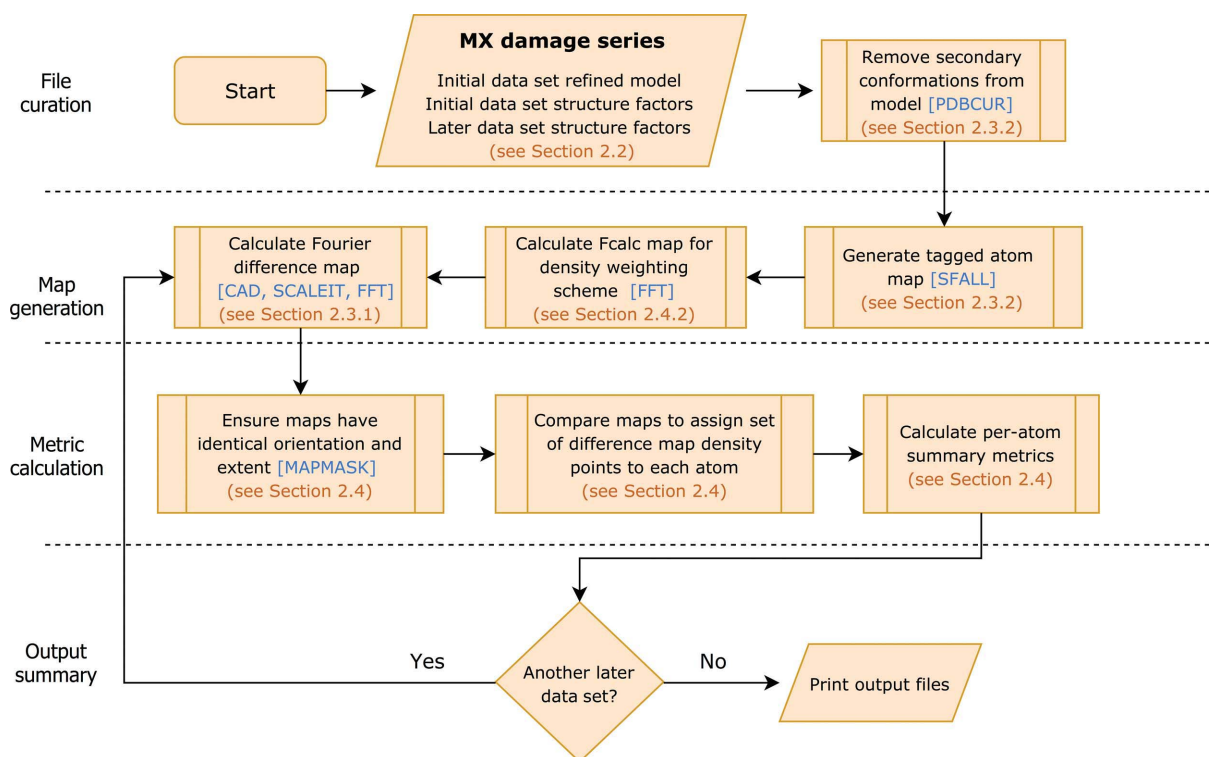


Figure 1

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

full details of which were provided in the supporting information of the article by Bury *et al.* (2017). This protein was chosen for the purpose of the current demonstration primarily because of its large size: 96.8 kDa per asymmetric unit, containing 2 monomers, each of 445 amino acids. As such, this protein represents a case where the characterization of specific damage through manual inspection of $F_{\text{obs},n} - F_{\text{obs},1}$ maps at multiple doses would have been challenging. Each asymmetric unit contains a large test set of established radiation-sensitive residues (Cys, 36; Glu, 28; Asp, 74; Met, 22) which is ideal for verifying the development of damage detection tools. Coordinate models and MTZ-format structure factor amplitude files have been retrieved directly from the PDB for accession codes 5mcc, 5mcd, 5mce, 5mcf, 5mch, 5mci, 5mcj, 5mck, 5mcl, 5mcm and 5mcn. The files can also be downloaded directly from <https://doi.org/10.5281/zenodo.1043864>.

2.2. Data preparation

RIDL can be used to analyse MX damage series composed of multiple data sets, $n = 1, \dots, N$, collected from either single crystals or merged multi-crystal data sets. Here, an MX damage series is defined by a single PDB-format coordinate file corresponding to a refined model for the initial data set, in addition to a series of $N > 1$ MTZ-format merged and scaled structure factor files (*e.g.* output from *CTRUNCATE*), which correspond to disjoint wedges of diffraction data collected on the crystal at increasing doses. The information for a full MX damage series consisting of a low dose data set and either single or multiple higher dose data sets must be specified

within a single plain-text input file to be parsed by *RIDL*. An example input file is provided in Fig. 2.

2.3. Map generation

2.3.1. Difference map generation. $F_{\text{obs},n} - F_{\text{obs},1}$ maps are generated for each data set $n > 1$ using a widely accepted

```

dir ./RIDL-example-output/

INITIALDATASET
name1 GH7-1
mtz1 ./GH7-1.mtz
mtzlabels1 FP_1
pdb1 ./GH7-1.pdb
RfreeFlag1 FreeR_flag
dose1 1.11

LATERDATASET
name2 GH7-2, GH-3
mtz2 ./GH7-2.mtz, ./GH7-3.mtz
mtzlabels2 FP_2, FP_3
dose2 3.27, 5.43

PHASEDATASET
name3 GH7-1
mtz3 ./GH7-1.mtz
phaseLabel PHIC
FcalcLabel FC
  
```

Figure 2

Example plain-text input file for the *RIDL* pipeline for the case of the GH7 damage series. For simplicity, $F_{\text{obs},n} - F_{\text{obs},1}$ maps will only be generated here for $n = 2$ and 3. This input file can be run with the GH7 data located at <https://doi.org/10.5281/zenodo.1043864>.

protocol in the radiation damage community (Southworth-Davies *et al.*, 2007). For each higher dose data set $n > 1$ of an MX damage series, the program *CAD* combines the observed structure factor amplitudes $|F_{\text{obs}}|$ and corresponding standard deviation σF_{obs} columns from the merged MTZ-format files for the initial and higher dose data sets. The high dose $|F_{\text{obs},n}|$ values are then scaled to the low dose $|F_{\text{obs},1}|$ reference set with *SCALEIT*, with a default anisotropic temperature factor scaling function of the form

$$C \exp[-(h^2 B_{11} + k^2 B_{22} + l^2 B_{33} + 2hkB_{12} + 2hlB_{13} + 2klB_{23})] \quad (1)$$

being applied to a higher dose $|F_{\text{obs},n}|$ value for Miller index hkl . Here, C and B_{ij} [$i, j \in \{1, 2, 3\}$] are the scale and the anisotropic temperature factor values as determined in *SCALEIT* through a least-squares fitting approach to all the structure factors by a modification of the method of Fox & Holmes (1966). The scaling step is performed sequentially for all higher dose data sets of the damage series, such that all data sets are scaled to an identical reference set. It is noted that alternative scaling protocols are available in *SCALEIT* (namely isotropic B factor scaling and single scale factor scaling); these were also trialled on the GH7 data, and for these sample data the site-specific damage events generally persist as reported through manual inspection of $F_{\text{obs},n} - F_{\text{obs},1}$ maps. The user can optionally override the default setting in *RIDL* in order to select these alternative scaling protocols if desired. However, caution must be taken since metrics derived from *RIDL* may differ depending on the scaling protocol used, and thus it is recommended to retain the default anisotropic scaling.

FFT (Ten Eyck, 1973) is then used to generate each $F_{\text{obs},n} - F_{\text{obs},1}$ map over the extent of the whole unit-cell dimensions corresponding to the input PDB-format coordinate model for the first data set. A model phase set φ_{calc} must be supplied by the user through the *RIDL* input file. It is recommended to use an identical fixed phase set for all data sets in the damage series, such that the experimentally recorded $|F_{\text{obs},n}|$ remain the only variable between increasing dose data sets. However, differing phase sets for each data set can be specified in the *RIDL* input file if required by the user. By default, the calculated phase set corresponding to the refined model for data set 1, $\varphi_{\text{calc},1}$, should be specified.

It is noted that other accepted protocols already exist to generate Fourier difference map coefficients, including the Bayesian approach of ‘ q weighting’, which improves the signal-to-noise ratio in difference maps (Ursby & Bourgeois, 1997), and the recent use of rank scaling of Fourier syntheses prior to direct comparison between maps (Urzhumtsev *et al.*, 2014). In order to maximize the adaptability of the *RIDL* pipeline, the user can optionally override the map generation step outlined in detail above, and instead directly supply map coefficients (input as MTZ-format columns) corresponding to their own custom-built maps.

Although no attempt has been made within the *RIDL* pipeline to perform any zero dose extrapolation of observed

intensity values, for a damage series containing enough data points, extrapolation models have been presented in the literature (e.g. Diederichs *et al.*, 2003) that could generate a zero dose ‘undamaged’ $|F_{\text{obs},0}|$ set. The $|F_{\text{obs},0}|$ set could then replace the $|F_{\text{obs},1}|$ set in the input to *RIDL*, thus generating a series of $F_{\text{obs},n} - F_{\text{obs},0}$ maps. Exploring such extrapolation possibilities is left to the discretion of the user and will not be discussed further here.

2.3.2. Assignment of per-atom search radius. A local region of space to be attributed to each atom is determined using *SFALL*, in the form of a tagged atom map generated over the unit-cell dimensions corresponding to the input coordinate model (identical to that of each $F_{\text{obs},n} - F_{\text{obs},1}$ map). The atom map is an integer field where every point in space is associated with the unique index of the nearest atom, and as discussed in §2.4, it is employed here as a template to indicate which voxels in each $F_{\text{obs},n} - F_{\text{obs},1}$ map should be assigned to each atom. To determine which atom-map voxel is associated with which atom, *SFALL* assumes that electron density is distributed as a Gaussian function, centred at each atomic position, that is dependent on both element identity and isotropic B factor, in order to calculate a non-overlapping region of space for each atom that factors in the relative contribution of electron density to that region by each local atom.

Prior to running *SFALL*, *RIDL* first calls *PDBCUR* at runtime to remove any hydrogen atoms, anisotropic B factors and secondary conformations from the input coordinate model. Any anisotropic B factors are replaced by equivalent isotropic B factors, as required for *SFALL* to run. By default, for an atom with isotropic B factor $B > 0 \text{ \AA}^2$, *SFALL* assigns a maximum search radius of

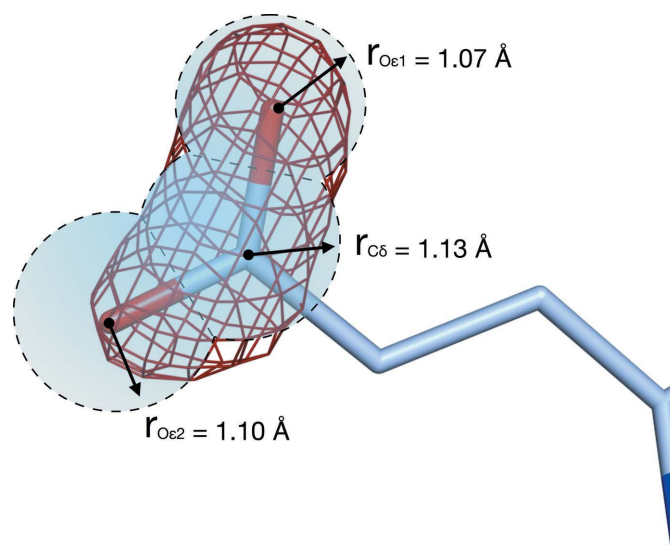


Figure 3
Per-atom search radius (dashed lines) for three adjacent atoms in an example glutamate side group (Glu150, chain A) in the GH7 protein. An $F_{\text{obs},3} - F_{\text{obs},1}$ difference map (contoured at -3σ in red) provides evidence for decarboxylation of the group. Search radii r_X illustrate the maximum allowed search radius defined by equation (2) for each selected atom ($X = C_\delta, O_{\epsilon 1}$ and $O_{\epsilon 2}$). In *RIDL*, for each atom, the $F_{\text{obs},n} - F_{\text{obs},1}$ map voxels, V_{atom} , within the assigned search radius are located and used to derive per-atom damage metrics as defined in §2.4.

$$r_{\max} = \alpha \frac{(B + 25)^{1/2}}{2\pi} \quad (2)$$

around the atomic centre (where $\alpha > 0$ is a user-defined scalar of default value 2.5 in *SFALL*, and the offset of 25 has units of \AA^2), provided this region does not overlap with that of any adjacent atoms. Fig. 3 provides an illustration of how such non-overlapping regions are defined. For the purposes of specific damage detection in *RIDL*, $\alpha = 1$ has been fixed as a default, such that only density changes in close proximity to atomic centres are considered (such that the search radius is $1 \leq r_{\max} \leq 2 \text{\AA}$ for the example case where the atomic B factor lies between 10 and 100\AA^2). The default α value can be overridden by the user if required. It has been found that the ordering of high electron density loss sites (C_α -normalized $D_{\text{loss}} > 5$, see §2.4.3) detected by *RIDL* is well conserved across differing α between 0.5 and 3 (the upper bound permitted by *SFALL*). However, α should be fixed at a constant value, to ensure that the calculation of summary metrics is comparable between different structures or data collection sessions.

The program *MAPMAN* (Kleywegt & Jones, 1996) can also be used to extract electron density values at atomic positions, by interpolating, averaging or integrating over grid points surrounding an atom; however, it does not output a set of raw voxel values assigned to each atom and is therefore unsuitable for use in *RIDL*. *MAPMAN* also only permits a single search radius for all atoms in a coordinate model, which does not account for the spread of electron density as parametrized by individual atomic B factors. It is noted that in the paper by Tickle (2012) the calculations of search radii used by different programs (*MAPMAN* and *SFALL*) were reviewed and a more sophisticated search radius r_{\max} that depends on atomic B factor, atom type and map resolution was implemented in *EDSTATS*. For a given atom, the *EDSTATS* r_{\max} is computed as the radius at which the radius integral of the calculated electron density derived from the model, $\rho_{\text{calc}}(r)$, attains a value that is 95% of the theoretical maximum at an infinite radius (Tickle, 2012). In the current *RIDL* pipeline, such a radius determination scheme has not been explicitly implemented. Instead, *RIDL* provides the option to construct damage metrics that weight the importance of each $F_{\text{obs},n} - F_{\text{obs},1}$ map voxel by the relative $\rho_{\text{calc}}(r)$ density at that voxel, in order to bypass the requirement for a strictly defined search radius (see §2.4.2).

2.4. Calculation of per-atom damage metrics

The tagged atom map and each $F_{\text{obs},n} - F_{\text{obs},1}$ map (for $n = 2, \dots, N$) are cropped to the crystal asymmetric unit corresponding to the refined coordinate model using *MAPMASK*, such that the resulting maps have exactly the same grid sampling dimensions and orientation in the asymmetric unit. The tagged atom map indicates the exact set of voxel indices, V_{atom} , in the $F_{\text{obs},n} - F_{\text{obs},1}$ map that reside in the defined search radius of an atom. The following sections describe several candidate metrics, and discuss their applic-

Table 1
Metric definitions output by *RIDL*.

$\rho_{\Delta}(v)$ and $\rho_{\text{calc}}(v)$ denote the density value at point v in an $F_{\text{obs},n} - F_{\text{obs},1}$ map and a map calculated directly from F_{calc} , respectively. V_{atom} is the set of distinct points in the $F_{\text{obs},n} - F_{\text{obs},1}$ map that are assigned to a specific atom. V_{atom}^- is the subset of V_{atom} for which the corresponding value in the $F_{\text{obs},n} - F_{\text{obs},1}$ map is negative. C_α represents the set of atoms constituting the C_α backbone of a protein that are present within the refined coordinate model. Only the C_α -normalized form of the $D_{\text{neg}}(\text{atom})$ metric is included; however, the normalization schemes are identical for $D_{\text{loss}}(\text{atom})$ and $D_{\text{neg}}^p(\text{atom})$.

Metric	Formula
$D_{\text{loss}}(\text{atom})$	$\max_{v \in V_{\text{atom}}} [-\rho_{\Delta}(v)]$
$D_{\text{loss}}^p(\text{atom})$	$[\max_{v \in V_{\text{atom}}} -\rho_{\Delta}(v)\rho_{\text{calc}}(v)] / [\max_{v \in V_{\text{atom}}} \rho_{\text{calc}}(v)]$
$D_{\text{neg}}(\text{atom})$	$[\sum_{v \in V_{\text{atom}}^-} -\rho_{\Delta}(v)\rho_{\text{calc}}(v)] / [\sum_{v \in V_{\text{atom}}^-} \rho_{\text{calc}}(v)]$
C_α -normalized $D_{\text{neg}}(\text{atom})$	$[D_{\text{neg}}(\text{atom}) - \langle D_{\text{neg}}(a) \rangle_{a \in C_\alpha}] / [\sigma_{a \in C_\alpha} [D_{\text{neg}}(a)]]$

ability and limitations. Table 1 provides a summary of such metrics.

2.4.1. D_{loss} : a simple indicator for damage. In MX experiments at 100 K, cleavage of chemical bonds and disordering of atoms typically leaves a signature of electron density loss in $F_{\text{obs},n} - F_{\text{obs},1}$ maps. The $D_{\text{loss}}(\text{atom})$ metric (in units of e \AA^{-3}) is defined as

$$D_{\text{loss}}(\text{atom}) = \max_{v \in V_{\text{atom}}} [-\rho_{\Delta}(v)], \quad (3)$$

where $\rho_{\Delta}(v)$ is the $F_{\text{obs},n} - F_{\text{obs},1}$ map value at voxel $v \in V_{\text{atom}}$. This metric is a simple yet clear indicator of sites of significant electron density loss near atoms and provides a worst case scenario for the electron density lost by an atom. A major appeal of $D_{\text{loss}}(\text{atom})$ is that it shares a direct link with manual inspection of $F_{\text{obs},n} - F_{\text{obs},1}$ map peaks (e.g. Fioravanti *et al.*, 2007; Borek *et al.*, 2007), since for a negative $F_{\text{obs},n} - F_{\text{obs},1}$ map peak located in the vicinity of an atom, $D_{\text{loss}}(\text{atom})$ exactly equals the height of the peak in e \AA^{-3} . The $D_{\text{loss}}(\text{atom})$ metric has previously been employed to characterize the differential radiation damage susceptibility of protein and nucleic acid in a 202 kDa protein–RNA complex (Bury *et al.*, 2016), and in a systematic study (Bury *et al.*, 2017) on radiation-induced changes to tyrosine over a survey of 18 independent MX damage series (all conducted at 100 K) that were previously deposited in the PDB. Both these studies would have been difficult without the use of the $D_{\text{loss}}(\text{atom})$ metric to efficiently probe for damage events.

The false negative rate associated with $D_{\text{loss}}(\text{atom})$ is negligible by design, making $D_{\text{loss}}(\text{atom})$ a useful indicator to screen for potential damage sites that should be flagged for the attention of the user. The $D_{\text{loss}}(\text{atom})$ metric is, however, limited by its inherent susceptibility to false positives, where map noise in the vicinity of an atom may be incorrectly interpreted as damage to the atom. To mitigate this issue, a variant of $D_{\text{loss}}(\text{atom})$ is presented in §2.4.2.

2.4.2. Density-weighting scheme. Although the $D_{\text{loss}}(\text{atom})$ metric provides a useful indicator of potential damage sites, it does not provide a suitable description of the overall magnitude of electron density lost at an atomic site. It is non-trivial

to define such metrics, since metrics that require any averaging or integration of voxel values inside a specified search radius are particularly sensitive to the search radius used, because of the inclusion/exclusion of voxels as the search radius is varied. A possible solution is to apply a voxel-weighting scheme, in which each voxel $v \in V_{\text{atom}}$ is weighted by how much of an atom's electron density contributed to that voxel. This is achieved in *RIDL* by generating a ρ_{calc} map with *FFT*, calculated using the structure amplitudes $|F_{\text{calc}}|$ and phases φ_{calc} derived from the input coordinate model. The ρ_{calc} map is calculated over the same sample cell dimensions and grid orientation as each $F_{\text{obs},n} - F_{\text{obs},1}$ map above. The $D_{\text{neg}}(\text{atom})$ metric is defined as

$$D_{\text{neg}}(\text{atom}) = \frac{\sum_{v \in V_{\text{atom}}^-} -\rho_{\Delta}(v)\rho_{\text{calc}}(v)}{\sum_{v \in V_{\text{atom}}^-} \rho_{\text{calc}}(v)}, \quad (4)$$

and is a weighted average over all voxels $V_{\text{atom}}^- \subseteq V_{\text{atom}}$ in the vicinity of an atom that attain $\rho_{\Delta}(v) < 0$. Only voxels attaining negative $\rho_{\Delta}(v)$ values have been considered as contributing to a metric for density loss, and it is noted that a full weighted average over V_{atom} can obscure the presence of density loss if a significant number of local voxels with $\rho_{\Delta}(v) \geq 0$ are also present. The $D_{\text{neg}}(\text{atom})$ and $D_{\text{loss}}(\text{atom})$ metrics are reasonably correlated, as shown in Figs. 4(a) and 4(b). However,

deviations are expected for the highest ranked damage sites, since whereas $D_{\text{neg}}(\text{atom})$ incorporates information from all voxels in the vicinity of an atom, $D_{\text{loss}}(\text{atom})$ is only affected by the most negative voxel value.

Similarly, a density-weighted version of $D_{\text{loss}}(\text{atom})$ can also be defined as

$$D_{\text{loss}}^{\rho}(\text{atom}) = \frac{\max_{v \in V_{\text{atom}}} -\rho_{\Delta}(v)\rho_{\text{calc}}(v)}{\max_{v \in V_{\text{atom}}} \rho_{\text{calc}}(v)}, \quad (5)$$

in order to suppress noisy background in the vicinity of each atom, and this is proposed as a more suitable metric than $D_{\text{loss}}(\text{atom})$ for the detection of damage sites. Figs. 4(c) and 4(d) illustrate that, whereas the density-weighting scheme most significantly affects atoms attaining low $D_{\text{loss}}(\text{atom})$ values (undamaged atoms), where the $D_{\text{loss}}^{\rho}(\text{atom})$ values are lower, the ordering of top damage sites is conserved. At the most pronounced protein damage sites (e.g. disulphide bonds), the negative $F_{\text{obs},n} - F_{\text{obs},1}$ peak is anticipated to align directly with each Cys-S_γ centre, and density weighting will have a negligible effect. As the arrow in Fig. 4(d) indicates, instances do exist of atomic sites where a difference peak is poorly aligned with the atomic centre.

2.4.3. C_{α} -normalization scheme. It cannot be assumed that the metrics proposed above are directly comparable between independent MX damage series, particularly if collected on different proteins, at different doses and up to different high resolution cut-offs. Furthermore, as Fig. 5(a) demonstrates, the modal value for the distribution of $D_{\text{neg}}(\text{atom})$ values attained for the GH7 test protein shifts to larger values with increasing dose, such that even for radiation-insensitive atoms the damage metrics are observed to increase with dose. This is due to the inevitable increase in $F_{\text{obs},n} - F_{\text{obs},1}$ map noise with increasing dose as a result of the diminishing diffraction data quality with dose. Ultimately, the raw $D_{\text{neg}}(\text{atom})$ [or indeed $D_{\text{loss}}(\text{atom})$] value of an atom alone is insufficient to determine whether an atom has experienced significant electron density loss, without reference to the rest of the structure.

A normalization scheme can be constructed in which metric values are compared relative to the reference set of C_{α} atoms making up the protein backbone. The C_{α} atom set was chosen since it is widely considered not to be radiation sensitive at the MGy doses typically reported for MX data collections at 100 K, and no observation of widespread explicit damage to the protein backbone at RT has been published to date. Note that for non-protein-containing crystals this normalization step is not achievable, and currently no suitable scheme is proposed for such cases.

For the case of the $D_{\text{neg}}(\text{atom})$ metric, the scheme is as follows:

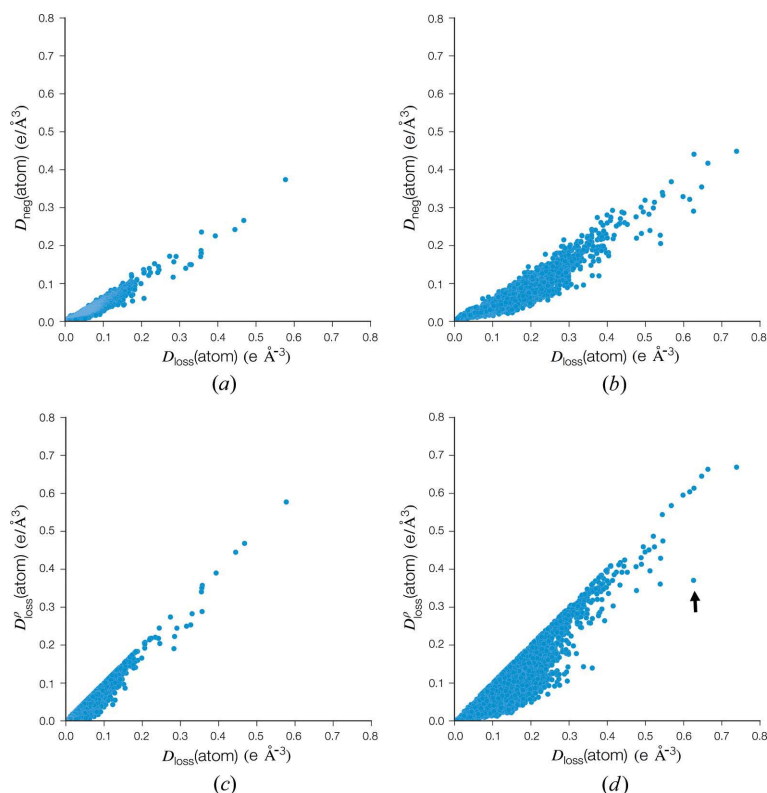


Figure 4
For the GH7 protein example, $D_{\text{loss}}(\text{atom})$ metric values have been compared for all atoms present in the coordinate model against (a), (b) $D_{\text{neg}}(\text{atom})$ and (c), (d) $D_{\text{loss}}^{\rho}(\text{atom})$ at 5.67 MGy (a), (c) and 22.7 MGy (b), (d) relative to the first data set at 1.11 MGy. In (d) the arrow indicates an example atomic site at which a difference map peak is not centrally aligned with the atom, such that $D_{\text{loss}}^{\rho}(\text{atom})$ and $D_{\text{loss}}(\text{atom})$ differ significantly.

$$C_{\alpha}\text{-normalized } D_{\text{neg}}(\text{atom}) = \frac{D_{\text{neg}}(\text{atom}) - \langle D_{\text{neg}}(a) \rangle_{a \in C_{\alpha}}}{\sigma_{a \in C_{\alpha}}[D_{\text{neg}}(a)]}, \quad (6)$$

where $\langle D_{\text{neg}}(a) \rangle_{a \in C_{\alpha}}$ and $\sigma_{a \in C_{\alpha}}[D_{\text{neg}}(a)]$ represent the average and standard deviation of $D_{\text{neg}}(\text{atom})$ attained by the set of C_{α} atoms. For the other metrics defined above [e.g. $D_{\text{loss}}(\text{atom})$], the normalization scheme can be similarly defined: see Bury *et al.* (2017) for an example in which the C_{α} -normalized $D_{\text{loss}}(\text{atom})$ metric has been exploited to compare independent proteins.

This normalization scheme preserves the ordering of damage to atoms as given by the non-normalized $D_{\text{neg}}(\text{atom})$ metric; however, it has now become a unitless metric.

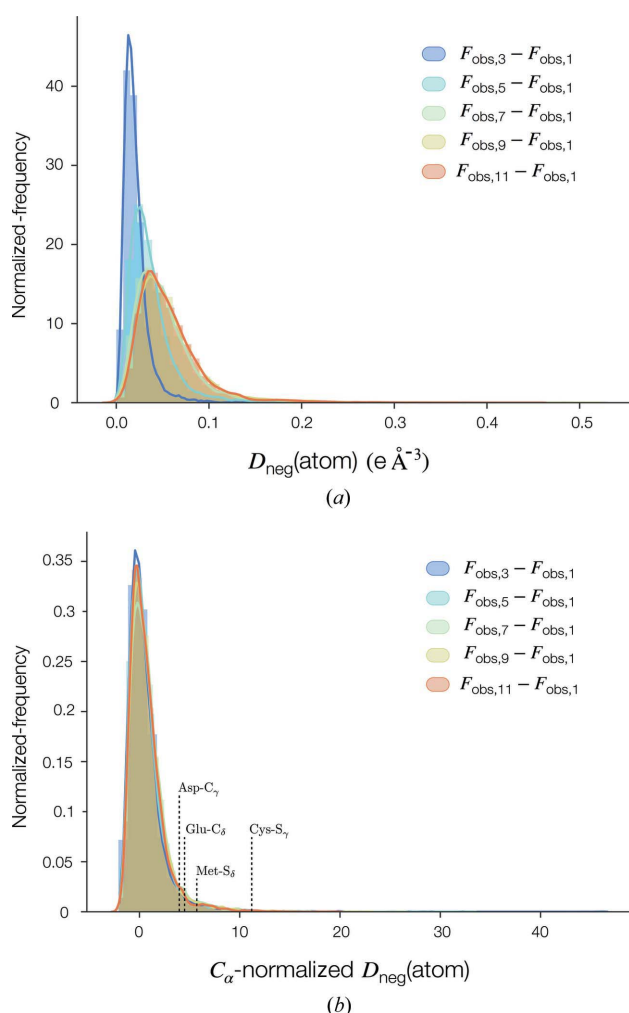


Figure 5

For the GH7 protein damage series, the distribution of (a) $D_{\text{neg}}(\text{atom})$ and (b) C_{α} -normalized $D_{\text{neg}}(\text{atom})$ metrics attained by all (7300) refined atoms in the coordinate model, shown for each higher dose data set in the series. Each distribution has been presented as a histogram with a kernel density estimation (KDE) overlaid, *i.e.* as a non-parametric estimate of the probability density function underlying each distribution, as an efficient way to visualize multiple sets of data on a single set of axes. KDEs have been generated using the function *kdeplot* within the *Seaborn* plotting library in Python. In (b), the location of the mean metric value for known radiation-susceptible atoms has been displayed, having been first averaged over all atoms of each type, and also averaged across the five displayed data sets.

Equation (6) is strictly a Z-score, with the properties that $\langle C_{\alpha}\text{-normalized } D_{\text{neg}}(a) \rangle_{a \in C_{\alpha}} = 0$ and $\sigma_{a \in C_{\alpha}}(C_{\alpha}\text{-normalized } D_{\text{neg}}) = 1$ must hold at any dose. As such, equation (6) is valid under the assumption that both the mean and standard deviation of $D_{\text{neg}}(\text{atom})$ values over the set of C_{α} atoms are dose independent. Fig. 5(b) demonstrates the effect of applying equation (6) to the GH7 test protein; the metric values exhibit a skewed unimodal distribution with mode ~ 0 across the tested dose range. The C_{α} -normalized $D_{\text{neg}}(\text{atom})$ metric thus quantifies the magnitude of electron density loss at an atomic site, whilst accounting for the presence of global radiation damage within the system. Interestingly, the C_{α} -normalized distributions overlap well across the range of investigated doses for the GH7 test protein, indicating that an increase in dose from 5.43 to 22.7 MGy does not yield a significant increase in radiation damage signal in each $F_{\text{obs},n} - F_{\text{obs},1}$ map.

The distribution shapes in Fig. 5(b) indicate the overall specific damage to the protein at each dose. For a perfectly radiation-insensitive sample, for which the spread of positive and negative $F_{\text{obs},n} - F_{\text{obs},1}$ map density should be uniformly distributed across the asymmetric unit, the distribution of metric values should theoretically be symmetric about 0 (however they should not be assumed to follow a Gaussian distribution). In practice, even in the absence of site-specific damage events, for any sample the side chains are generally anticipated to be more flexible than the protein backbone, and thus the resulting metric distribution will not be symmetric about 0. It is noted that skewness provides a measure of asymmetry about the distribution *mean*; this mean cannot be assumed to be zero (and will take larger positive values for a sample for which a significant fraction of atoms are radiation damaged). Skewness is therefore a suboptimal measure of overall radiation sensitivity of a sample at a given dose. We propose that the asymmetry about 0 can be directly evaluated as $\text{ASYM} = \Sigma_{+}/\Sigma_{-}$, where Σ_{+} and Σ_{-} are the sums of C_{α} -normalized $D_{\text{neg}}(\text{atom})$ metric values over all atoms for which the metric takes a positive or negative value, respectively. As illustrated in Table 2 for the GH7 test protein, the ASYM score increases initially with dose, indicating that the sample is exhibiting greater radiation damage. However, the ASYM score values eventually diminish at high doses, as $F_{\text{obs},n} - F_{\text{obs},1}$ map background noise also increases with dose as an inevitable manifestation of global damage.

2.4.4. The effect of the data resolution. It has been widely reported that diffraction resolution, and thus the achievable resolution of the resulting electron density map, diminishes as a function of dose. It is essential to be able to predict the effect of resolution, in order to be able to compare *RIDL* metric values between different resolution data sets, whether in a single MX damage series or between independent crystals. For the GH7 protein damage series, *RIDL* was run repeatedly, with the maximum permitted resolution of each $F_{\text{obs},n} - F_{\text{obs},1}$ map truncated using *FFT* in 0.5 Å increments between 2 and 4 Å. As illustrated in Figs. 6(a) and 6(b), the $D_{\text{neg}}(\text{atom})$ metric appears to be highly dependent on $F_{\text{obs},n} - F_{\text{obs},1}$ map resolution, with a greater reduction in map resolution leading to

Table 2

Statistical characterization of the overall distributions of C_α -normalized $D_{\text{neg}}(\text{atom})$ metric values, as a function of increasing dose for the GH7 test protein.

Results are calculated over metric values for all refined atoms present in the input refined coordinate model. The ASYM score is calculated as $\text{ASYM} = \Sigma_+ / \Sigma_-$, where Σ_+ and Σ_- are the sums of C_α -normalized $D_{\text{neg}}(\text{atom})$ metric values over all atoms for which the metric takes a positive or negative value, respectively.

	Fourier difference map used to derive metrics				
	$F_{\text{obs},3} - F_{\text{obs},1}$	$F_{\text{obs},5} - F_{\text{obs},1}$	$F_{\text{obs},7} - F_{\text{obs},1}$	$F_{\text{obs},9} - F_{\text{obs},1}$	$F_{\text{obs},11} - F_{\text{obs},1}$
Asymmetry score (ASYM)	3.05	3.55	4.53	4.13	3.69
Skewness	5.43	3.72	2.92	2.72	2.69
Mean (standard deviation in brackets)	0.61 (2.08)	0.69 (1.97)	0.84 (1.95)	0.77 (1.86)	0.69 (1.79)

lower metric values. As shown in Figs. 6(c) and 6(d), C_α normalization can compensate for the explicit dependence of $D_{\text{neg}}(\text{atom})$ on resolution at a moderate truncation of up to 3.5 Å. However, it is noted that resolution dependence eventually manifests again at larger truncation extents. Similar

behaviour is also exhibited for the C_α -normalized form of the $D_{\text{loss}}(\text{atom})$ metric (data not shown). This result implies that, if metric values are to be directly compared between independently collected damage series, C_α normalization using equation (6) must first be performed.

2.4.5. The effect of phase error.

The reliability of each $F_{\text{obs},n} - F_{\text{obs},1}$ map is ultimately dictated by the quality of the phase information derived from the refined coordinate model. In order to determine the sensitivity of the metrics output from *RIDL* to phase quality, phases were degraded by perturbing all xyz atomic coordinates in the lowest dose GH7 protein coordinate model, using a custom written Python script. A maximum displacement factor $\varepsilon \geq 0$ Å was introduced to perturb each atom, such that the coordinates of an atom became

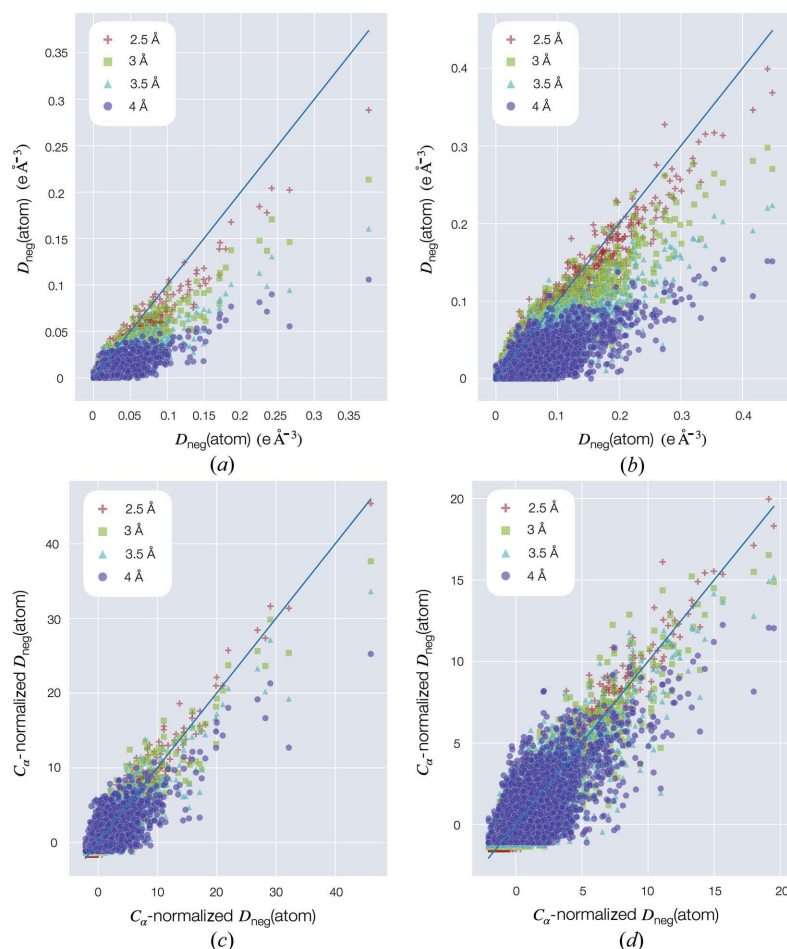
$$x_{\text{pert}} = x + U[-\varepsilon, \varepsilon], \quad (7)$$

$$y_{\text{pert}} = y + U[-\varepsilon, \varepsilon], \quad (8)$$

$$z_{\text{pert}} = z + U[-\varepsilon, \varepsilon], \quad (9)$$

where $U[-\varepsilon, \varepsilon]$ is a uniformly distributed random variable between $-\varepsilon$ and ε (in units of Å). This process was repeated successively for increasing ε between 0 and 1 Å. At $\varepsilon = 1$ Å, a mean difference in phase of 55.5° from the starting phases was calculated over all hkl Miller indices. For each perturbation level ε , the resulting coordinate model was used to calculate a new phase set, $\varphi_{\text{calc,pert}}$, using *SFALL*.

RIDL was then run using the degraded phases, but with all other input parameters held constant. Here, it was assumed that the $\varphi_{\text{calc,pert}}$ phase set were supplied to *RIDL* independently from the refined coordinate model, such that the placement of atoms (and consequently per-atom search radii) was unaffected. Fig. 7 indicates that, for the GH7 test study, *RIDL* was relatively robust to the effects of phase degradation, with the ordering of the detected top damaged atom sites [with high $D_{\text{neg}}(\text{atom})$ values] conserved up to a mean phase error of $\leq 41.6^\circ$. At 55.5° , and particularly at 22.7 MGy, the introduced phase error becomes too substantial for *RIDL* to correctly identify damage

**Figure 6**

The dependence of the $D_{\text{neg}}(\text{atom})$ metric on the maximum $F_{\text{obs},n} - F_{\text{obs},1}$ map resolution, for the example case of the GH7 protein. Non-normalized and C_α -normalized metric values have been derived from the (a), (c) $F_{\text{obs},3} - F_{\text{obs},1}$ map and (b), (d) $F_{\text{obs},11} - F_{\text{obs},1}$ map, with the higher dose data set corresponding in each case to 5.67 and 22.7 MGy, respectively. The maximum map resolution has been varied in 0.5 Å increments between 2 and 4 Å. Each x axis indicates the $D_{\text{neg}}(\text{atom})$ metric values across all atoms at 2 Å. The corresponding y axes indicate metric values at each tested lower resolution cut-off. Deviations of points from the line $y = x$ (in blue) illustrate the resolution dependence of a metric. Data for other doses for the GH7 damage series (not shown) exhibited similar behaviour.

sites, as measured by deviations of data points from the line $y = x$ (in blue in Fig. 7).

2.4.6. Application to time-resolved experiments. Although *RIDL* has been written predominantly for radiation damage analysis, the objective metrics it outputs to track electron density changes between successive diffraction data sets are directly applicable to the growing field of time-resolved (TR) MX studies. Such studies are limited by user subjectivity, especially since the functional changes may only occur in a minor fraction of unit cells, are often dominated by stochastic map noise and also occur in the inevitable presence of radiation damage. It is proposed that the alternative normalization scheme

$$X\text{-normalized } D_{\text{neg}}(\text{atom}) = \frac{D_{\text{neg}}(\text{atom}) - \langle D_{\text{neg}}(a) \rangle_{a \in X}}{\sigma_{a \in X}[D_{\text{neg}}(a)]}, \quad (10)$$

where X denotes a set of radiation-sensitive atoms (e.g. Glu C_{δ} , $O_{\epsilon 1}$ and $O_{\epsilon 2}$, and Asp C_{γ} , $O_{\delta 1}$ and $O_{\delta 2}$ atoms), could provide a method to distinguish whether an electron density change to an amino acid group between successive data sets should be assigned as a functionally relevant change with confidence, or whether significant site-specific radiation damage has dominated. This principle holds either when full discrete data sets are collected by rotation crystallography (exploiting fast pixel detectors), or when multi-crystal merging is performed to produce discrete data sets [using either multiple pulse lengths at an X-ray free electron laser (XFEL) or fixed-target samples at synchrotrons]; the only strict condition is that multiple complete data sets can be reas-

sembled corresponding to different dose/exposure-time points (or pulse lengths at the XFEL). The user can explicitly specify the normalization set X in the *RIDL* input file using any atoms present within the coordinate model.

2.5. The program output

For an MX damage series consisting of $N \geq 2$ disjoint increasing dose diffraction data sets collected on a crystal, a run of *RIDL* produces a series of $(N - 1) F_{\text{obs},n} - F_{\text{obs},1}$ MAP-format files generated over the crystal asymmetric unit. These map files can be immediately visualized in *COOT* (Emsley *et al.*, 2010) to inspect radiation damage sites detected by *RIDL*. Comma-separated value (CSV)-format files are also output by *RIDL* that contain non-normalized metric values for each data set, and additionally C_{α} -normalized values for protein-containing crystals. This format is readily parsable and the user can easily extract these values for further independent analysis. An auto-generated HTML-format summary report is also output by *RIDL*. It has been designed to provide a concise yet informative graphical and statistical feedback of radiation damage within an input structure, in order to efficiently advise a non-expert user whether the input data have suffered significant site-specific radiation damage.

3. Discussion of metrics

The $D_{\text{loss}}(\text{atom})$ metric, as well as its density-weighted counterpart $D_{\text{loss}}^{\rho}(\text{atom})$, have been identified as suitable indicators of sites of significant radiation-induced electron density loss in macromolecular crystals, with the $D_{\text{loss}}(\text{atom})$ metric already proving useful in damage analysis across a multitude of systems (Bury *et al.*, 2016, 2017). Often, in addition to identifying sites of damage, the investigator wishes to probe exactly how much electron density is lost from that site with increasing dose. To this end, $D_{\text{neg}}(\text{atom})$ has also been proposed here, as a ρ_{calc} map weighted average of $F_{\text{obs},n} - F_{\text{obs},1}$ negative map density in the vicinity of each atom. It has further been demonstrated that metric normalization to the set of C_{α} backbone atoms can reduce the data resolution dependence of the metric, thus providing a metric, the C_{α} -normalized $D_{\text{neg}}(\text{atom})$, that can be compared directly between independently conducted damage experiments.

Several metrics have been proposed here and a question remains as to which single metric is optimal for use in radiation damage analysis. To establish this, radiation damage was simulated at known positions in the GH7 test protein coordinate model, and the ability of each metric to correctly identify each known damage site was compared using the following protocol. A custom Python script was written to modify the refined GH7 protein coordinate model, to reduce the occupancy of a randomly assigned 50% of the Glu

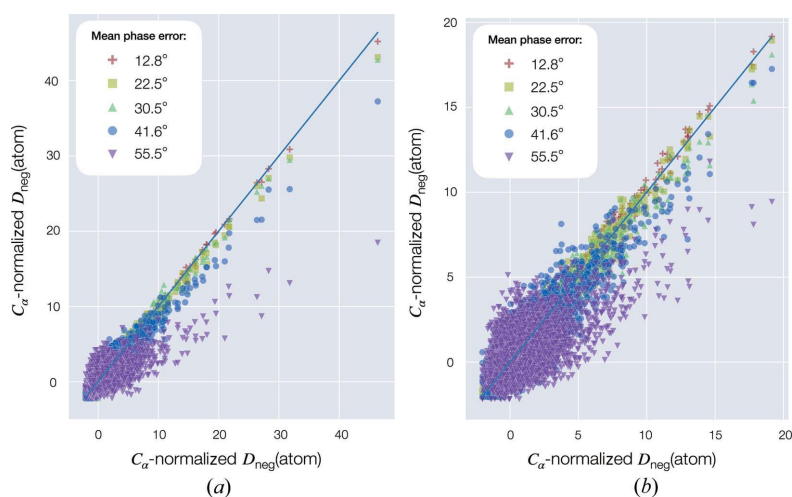


Figure 7

The dependence of the C_{α} -normalized $D_{\text{neg}}(\text{atom})$ metric on the $F_{\text{obs},n} - F_{\text{obs},1}$ map phase quality for the example case of the GH7 protein, with the higher dose data set corresponding in each case to (a) 5.67 and (b) 22.7 MGy, respectively. Each x axis indicates the $D_{\text{neg}}(\text{atom})$ metric values across all atoms in the protein, using the phases corresponding to the original refined coordinate model. The corresponding y axes indicate the metric values with degraded phases with a mean difference in phase between the original and perturbed model of between 12.8 and 55.5°. Deviations of points from the line $y = x$ (in blue) indicate atoms for which the metric is sensitive to the quality of the φ_{calc} phase set. Other doses for the GH7 protein damage series (not shown) exhibited similar behaviour.

and Asp carboxyl side groups to between 0.25 and 0.75 (whilst leaving the rest unaltered), in order to simulate decarboxylation to a subset of residues. The calculated structure amplitudes, $|F_{\text{calc}}|$, for both the original and modified coordinate models were generated using *SFALL*, and an $F_{\text{calc,modified}} - F_{\text{calc,original}}$ map was calculated in *RIDL* between the calculated structure amplitudes corresponding to the original model and the newly altered model. This map acted as a direct substitute for the $F_{\text{obs},n} - F_{\text{obs},1}$ map derived from experimental data, now with difference map peaks coincident with the modified Glu and Asp side groups at known locations. $F_{\text{obs},n} - F_{\text{obs},1}$ maps are typically noisy and increasingly so with rising dose. In order to simulate the presence of background map noise, Gaussian noise $N(0, \sigma^2)$ was added to the $|F_{\text{calc,modified}}|$ column using *SFTOOLS*. The noise variance σ^2 was increased in increments and the fraction of correctly identified simulated damage sites reported for each σ^2 .

The true positive rate (or recall) was compared for C_α -normalized forms of the $D_{\text{loss}}(\text{atom})$, $D_{\text{loss}}^\rho(\text{atom})$ and $D_{\text{neg}}(\text{atom})$ metrics, with an atom qualifying as a detected damage site only if the C_α -normalized metric exceeded 3 (so

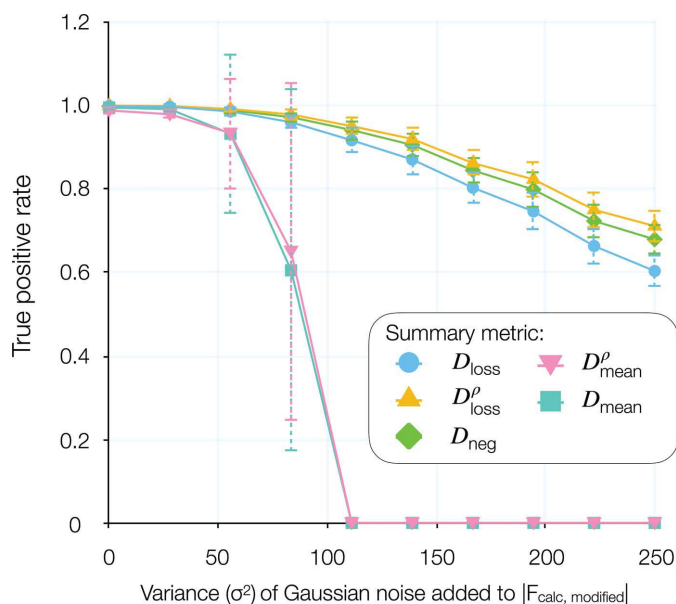


Figure 8

The true positive detection rate for simulated Glu and Asp damage sites using $F_{\text{calc,modified}} - F_{\text{calc,original}}$ maps. The true positive rate is reported for the C_α -normalized forms of the metrics output by *RIDL*, with an atom qualifying as a detected damage site only if the C_α -normalized metric exceeded 3. The x axis indicates the variance σ^2 of the Gaussian noise $N(0, \sigma^2)$ that is added to the $|F_{\text{calc,modified}}|$ structure amplitude column, in order to introduce noise into the $F_{\text{calc,modified}} - F_{\text{calc,original}}$ map. For each x axis value, the average true positive rate over 50 independently generated $F_{\text{calc,modified}} - F_{\text{calc,original}}$ maps is reported, with error bars representing standard deviations across the 50 repeats. A maximum tested σ^2 value of 250 is illustrated since the true positive rate for all reported metrics had decayed below 70% beyond this point. The data also provide a clear demonstration of why integration of $F_{\text{obs},n} - F_{\text{obs},1}$ map density around each atom is a poor indicator of radiation damage, with the $D_{\text{mean}}(\text{atom})$ metric, $D_{\text{mean}}(\text{atom}) = (1/|V_{\text{atom}}|) \sum_{v \in V_{\text{atom}}} -\rho_\Delta(v)$, as well as its ρ_{calc} -weighted counterpart, $D_{\text{mean}}^\rho(\text{atom})$, performing poorly compared to the $D_{\text{loss}}(\text{atom})$, $D_{\text{loss}}^\rho(\text{atom})$ and $D_{\text{neg}}(\text{atom})$ metrics.

there was a <1% chance that such a metric value would be attained by the C_α backbone, assuming the metric values for C_α atoms are normally distributed). Noticeably, as shown in Fig. 8, the true positive rate for carboxyl damage is consistently higher for the ρ_{calc} -weighted metrics, $D_{\text{loss}}^\rho(\text{atom})$ and $D_{\text{neg}}^\rho(\text{atom})$, than for $D_{\text{loss}}(\text{atom})$, and moreover the $D_{\text{neg}}(\text{atom})$ and $D_{\text{loss}}^\rho(\text{atom})$ metrics yield highly similar results. It is therefore suggested that a single metric, $D_{\text{neg}}(\text{atom})$, can be used as both an indicator and a quantifier of radiation damage to individual atoms in macromolecular structures.

It is crucial to note that the inclusion of a density-weighting scheme also increases the dependence of the metric on the quality of the coordinate model. Errors in the model, such as the poor positioning of an atom, the wrong element identity, or inaccuracies in either the B factor or occupancy of the atom, will propagate directly to errors in ρ_{calc} . As such, density-weighted metrics should only be trusted when the model is known to be satisfactorily refined. In cases where there is considerable uncertainty in the positioning of an atom, the $D_{\text{loss}}(\text{atom})$ metric still provides an efficient metric to screen for damage that is robust to model quality.

4. Conclusions

To date, limitations in $F_{\text{obs},n} - F_{\text{obs},1}$ map analysis have stemmed from the subjectivity of manual map inspection, and also the non-standardized protocols by which maps have been generated and analysed in various damage studies. The tool *RIDL* has now been developed to entirely remove the requirement for manual inspection of $F_{\text{obs},n} - F_{\text{obs},1}$ maps and to calculate metrics that have been specifically designed to characterize site-specific radiation damage to atoms. Furthermore, the accessible Python framework of *RIDL* permits easy adaptation to allow future researchers to develop additional useful metrics.

5. Availability

The source code and user guide for *RIDL* are openly available at <https://github.com/GarmanGroup/RIDL>. Test MX damage series data related to this article are located at <https://doi.org/10.5281/zenodo.1043864>. The custom Python script for simulating radiation damage (§3) is also available upon request to the authors.

Acknowledgements

We thank our testers (Eugenio De la Mora and Gianluca Santoni) for providing feedback on the performance of *RIDL*. We also thank Jonathan Brooks-Bartlett, Martin Weik and Pietro Roversi for insightful discussions on the development of *RIDL*.

Funding information

We gratefully acknowledge the UK Engineering and Physical Sciences Research Council (grant No. EP/G03706X/1 to

Charles Simon Bury) for studentship funding in the Systems Biology Programme of the University of Oxford Doctoral Training Centre (CSB).

References

- Borek, D., Ginell, S. L., Cymborowski, M., Minor, W. & Otwinowski, Z. (2007). *J. Synchrotron Rad.* **14**, 24–33.
- Berglund, G. I., Carlsson, G. H., Smith, A. T., Szöke, H., Henriksen, A. & Hajdu, J. (2002). *Nature*, **417**, 463–468.
- Burmeister, W. P. (2000). *Acta Cryst.* **D56**, 328–341.
- Bury, C. S., Carmichael, I. & Garman, E. F. (2017). *J. Synchrotron Rad.* **24**, 7–18.
- Bury, C., Garman, E. F., Ginn, H. M., Ravelli, R. B. G., Carmichael, I., Kneale, G. & McGeehan, J. E. (2015). *J. Synchrotron Rad.* **22**, 213–224.
- Bury, C. S., McGeehan, J. E., Antson, A. A., Carmichael, I., Gerstel, M., Shevtsov, M. B. & Garman, E. F. (2016). *Acta Cryst.* **D72**, 648–657.
- Clavel, D., Gotthard, G., von Stetten, D., De Sanctis, D., Pasquier, H., Lambert, G. G., Shaner, N. C. & Royant, A. (2016). *Acta Cryst.* **D72**, 1298–1307.
- De la Mora, E., Carmichael, I. & Garman, E. F. (2011). *J. Synchrotron Rad.* **18**, 346–357.
- Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Acta Cryst.* **D59**, 903–909.
- Dubnovitsky, A. P., Ravelli, R. B. G., Popov, A. N. & Papageorgiou, A. C. (2005). *Protein Sci.* **14**, 1498–1507.
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* **D66**, 486–501.
- Fioravanti, E., Vellieux, F. M. D., Amara, P., Madern, D. & Weik, M. (2007). *J. Synchrotron Rad.* **14**, 84–91.
- Fox, G. C. & Holmes, K. C. (1966). *Acta Cryst.* **20**, 886–891.
- Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N. & Alber, T. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 16247–16252.
- Helliwell, J. R. (1988). *J. Cryst. Growth*, **90**, 259–272.
- Juergens, D. H. & Weik, M. (2011). *J. Synchrotron Rad.* **18**, 329–337.
- Keedy, D. A. *et al.* (2015). *eLife*, **4**, e07574.
- Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* **D52**, 826–828.
- Kmetko, J., Warkentin, M., Englich, U. & Thorne, R. E. (2011). *Acta Cryst.* **D67**, 881–893.
- Leiros, H.-K. S., Timmins, J., Ravelli, R. B. G. & McSweeney, S. M. (2006). *Acta Cryst.* **D62**, 125–132.
- Matsui, Y., Sakai, K., Murakami, M., Shiro, Y., Adachi, S., Okumura, H. & Kouyama, T. (2002). *J. Mol. Biol.* **324**, 469–481.
- Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. (2005). *Acta Cryst.* **D61**, 1227–1237.
- Owen, R. L., Axford, D., Nettleship, J. E., Owens, R. J., Robinson, J. I., Morgan, A. W., Doré, A. S., Lebon, G., Tate, C. G., Fry, E. E., Ren, J., Stuart, D. I. & Evans, G. (2012). *Acta Cryst.* **D68**, 810–818.
- Owen, R. L., Paterson, N., Axford, D., Aishima, J., Schulze-Bries, C., Ren, J., Fry, E. E., Stuart, D. I. & Evans, G. (2014). *Acta Cryst.* **D70**, 1248–1256.
- Ravelli, R. B. & McSweeney, S. M. (2000). *Structure*, **8**, 315–328.
- Southworth-Davies, R. J., Medina, M. A., Carmichael, I. & Garman, E. F. (2007). *Structure*, **15**, 1531–1541.
- Ten Eyck, L. F. (1973). *Acta Cryst.* **A29**, 183–191.
- Tickle, I. J. (2012). *Acta Cryst.* **D68**, 454–467.
- Ursby, T. & Bourgeois, D. (1997). *Acta Cryst.* **A53**, 564–575.
- Urzhumtsev, A., Afonine, P. V., Lunin, V. Y., Terwilliger, T. C. & Adams, P. D. (2014). *Acta Cryst.* **D70**, 2593–2606.
- Weik, M., Ravelli, R. B., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 623–628.
- Winn, M. D. *et al.* (2011). *Acta Cryst.* **D67**, 235–242.
- Zeldin, O. B., Brockhauser, S., Bremridge, J., Holton, J. M. & Garman, E. F. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 20551–20556.
- Zeldin, O. B., Gerstel, M. & Garman, E. F. (2013). *J. Appl. Cryst.* **46**, 1225–1230.