JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

Real-space analysis of radiation-induced specific changes with independent component analysis

Dominika Borek,[a,b] Raquel Bromberg,[a] Johan Hattne[a,c] and Zbyszek Otwinowski[a]*

[a]Department of Biophysics, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA, [b]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA, and [c]Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA 20147, USA. *Correspondence e-mail: zbyszek@work.swmed.edu

Received 27 July 2016
Accepted 19 December 2017

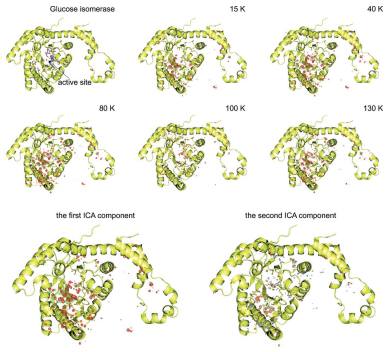
Edited by E. F. Garman, University of Oxford, England

Keywords: independent component analysis (ICA); singular value decomposition (SVD); radiation damage; tunnelling.

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

A method of analysis is presented that allows for the separation of specific radiation-induced changes into distinct components in real space. The method relies on independent component analysis (ICA) and can be effectively applied to electron density maps and other types of maps, provided that they can be represented as sets of numbers on a grid. Here, for glucose isomerase crystals, ICA was used in a proof-of-concept analysis to separate temperature-dependent and temperature-independent components of specific radiation-induced changes for data sets acquired from multiple crystals across multiple temperatures. ICA identified two components, with the temperature-independent component being responsible for the majority of specific radiation-induced changes at temperatures below 130 K. The patterns of specific temperature-independent radiation-induced changes suggest a contribution from the tunnelling of electron holes as a possible explanation. In the second case, where a group of 22 data sets was collected on a single thaumatin crystal, ICA was used in another type of analysis to separate specific radiation-induced effects happening on different exposure-level scales. Here, ICA identified two components of specific radiation-induced changes that likely result from radiation-induced chemical reactions progressing with different rates at different locations in the structure. In addition, ICA unexpectedly identified the radiation-damage state corresponding to reduced disulfide bridges rather than the zero-dose extrapolated state as the highest contrast structure. The application of ICA to the analysis of specific radiation-induced changes in real space and the data pre-processing for ICA that relies on singular value decomposition, which was used previously in data space to validate a two-component physical model of X-ray radiation-induced changes, are discussed in detail. This work lays a foundation for a better understanding of protein-specific radiation chemistries and provides a framework for analysing effects of specific radiation damage in crystallographic and cryo-EM experiments.

1. Introduction



X-ray exposure in diffraction experiments not only generates diffraction intensities which are essential for structure solution and are acquired in reciprocal space but also drives hundreds of chemical reactions in real space, which progress simultaneously across the whole crystal, with variable rates and through variable chemistries (Patten & Gordy, 1960; Rao & Hayon, 1974; Hawkins & Davies, 2001; Xu & Chance, 2007; Farver & Pecht, 1997; Rao *et al.*, 1983; Jones *et al.*, 1987; Garman & Weik, 2013; Weik, Kryger *et al.*, 2001; Weik *et al.*, 2000; Pimblott & LaVerne, 2007; Garrison, 1987). X-ray photon-driven reactions start with the production of a primary high-energy photoelectron. Each primary photoelectron generates hundreds of secondary low-energy electrons (LEEs)

(Ziaja *et al.*, 2005; Pimblott & LaVerne, 2007; Nave, 1995; Emfietzoglou *et al.*, 2003) which can migrate across multiple unit cells in a crystal, causing positive ionizations at the start and negative at the end (Alizadeh & Sanche, 2012*a*,*b*; Nave, 1995; Gonzalez & Nave, 1994). The paths of the secondary electrons will be affected by electric fields in the crystal, so we expect a uniform starting distribution for holes and a potentially non-uniform distribution for negatively charged states. Later these ionized states can migrate further *via* tunnelling and diffusion, with their migration influenced by electric fields in the crystal. The final consequence can be a recombination of electrons and holes, or the creation of covalent alterations of the protein and ligand structures. Even if recombination neutralizes charges, rearrangements of atomic structures during the process may in the end create minor structural alterations, particularly in water networks (Ball, 2008). The sum of a large number of such effects will result in an accumulating Gaussian displacement of atoms in an initial crystal structure. After application of the Fourier transform, these accumulated displacements will contribute to diffraction intensity decay, with the scaling *B*-factor representing the variance of these displacements and linearly increasing with absorbed radiation dose. The historical convention in crystallography is to multiply the three-dimensional variance of displacement by $8\pi^2/3$ when expressing it as the *B*-factor.

Covalent changes will create a signal of specific radiation-induced damage in the vicinity of the covalent structure alterations. While the locations of the origins of secondary electrons are randomly and uniformly distributed in the crystal lattice, their consequences are modulated by the migration of charged states. Both electrons and holes can migrate by tunnelling, which is temperature-independent, and at around 100 K they may start to have an additional component of motion driven by diffusion. The precise patterns of specific radiation damage will extend beyond covalent rearrangements, as they can create shifts in parts of protein structures and create complex alterations in the network of hydrogen-bonded water molecules (Weik, Kryger *et al.*, 2001; Florusse *et al.*, 2004; Pizzitutti *et al.*, 2007; Sterpone *et al.*, 2010; Bellissent-Funel *et al.*, 2016).

The decay of diffraction intensities for protein crystals was noticed early on in crystallography (Blake & Phillips, 1962). Corrections for diffraction decay have been used for several decades during data scaling, sometimes parameterized directly in reciprocal space as scaling *B*-factors (Otwinowski & Minor, 1997; Evans, 2011; Otwinowski *et al.*, 2003; Wilson & Yeates, 1979) and sometimes parameterized in detector space (Kabsch, 2010). The decay correction compensates for the majority of radiation-induced changes in diffraction intensities, for which reason it is a standard procedure when solving a crystal structure. Over time, more attention has been paid to the non-decay component of changes in diffraction intensities whose features and consequences need to be analysed and modelled in real space. In particular, multiple experiments identified disulfide bridges and Glu and Asp side-chains (Weik *et al.*, 2000; Petrova *et al.*, 2010; Liebschner *et al.*, 2013; Burmeister, 2000), halogenated nucleic acid bases (Ennifar *et al.*, 2002), Hg^{2+} (Ramagopal *et al.*, 2005) and many others as structural components that are on average more sensitive than the rest of the structure to radiation-induced chemistries.

The existence of these specific radiation-induced changes has multiple consequences for crystallographic procedures. The structural non-isomorphism produced by radiation may alter diffraction so that the desired phasing signal is obstructed. Specific radiation-induced changes may also be important for the structural interpretation of a studied molecule. Significant effort has already been made to correct specific radiation-induced changes in reciprocal space (Borek *et al.*, 2007, 2010, 2013; Diederichs *et al.*, 2003). In particular, the authors of this publication developed an approach using dimensionality reduction to correct for radiation-induced non-isomorphism in reciprocal space so that structure solution can be achieved even in the presence of a significant level of specific radiation-induced changes (Borek *et al.*, 2013). To achieve this task, we considered specific radiation-induced changes to be a departure from the average displacement described by the scaling *B*-factor, and we modelled these changes in reciprocal space in a dose-dependent manner (Borek *et al.*, 2007). This two-component model was validated by singular value decomposition (SVD) analysis (Borek *et al.*, 2013) based on the acquisition of multiple full data sets from the same crystal. The analysis indicated that a single SVD component dominates the data variance, and only occasionally can other weaker non-noise components be identified. This approach could be used for zero-dose extrapolation that effectively corrects for radiation-induced changes in reciprocal space. SVD analysis also allows for the interpolation or extrapolation of diffraction intensities to any dose of choice by the effective noise-filtering feature of SVD dimensionality reduction. Our work is consistent with the zero-dose extrapolation procedures introduced by others (Diederichs *et al.*, 2003), but it introduces more effective noise suppression.

In modelling radiation damage in real space, significant effort has also been made to analyse the chemical reactions that contribute to specific radiation-induced changes (Weik, Ravelli *et al.*, 2001, 2003; Ramagopal *et al.*, 2005; Leiros *et al.*, 2006; Fioravanti *et al.*, 2007; Fütterer *et al.*, 2008; Macedo *et al.*, 2009), with several approaches proposed to model them in real space that generally rely on some version of calculating the difference maps between the heavily damaged state and the state that was as close as possible to the original structure (Weik *et al.*, 2000; Ravelli & Garman, 2006; Ramagopal *et al.*, 2005; Gerstel *et al.*, 2015; Bury *et al.*, 2016). These difference maps are equivalent to the maps calculated with the SVD-derived components (Borek *et al.*, 2013), except that they have higher noise levels.

Here we present an exploratory data analysis that allows for the use of complex multidimensional crystallographic data sets to answer questions about specific radiation-induced changes. There are many methods for factorizing large multi-dimensional data sets. The choice of the method is driven by the properties of the data and the scientific questions. Specific radiation-induced changes are difference signals which are

distributed in real space and which are frequently weak. Therefore, their analysis can be very sensitive to the noise level and so noise-filtering by appropriate procedures is essential. Dimensionality reduction can be used not only to separate signal from noise but also as a tool in interpreting structural signals. SVD or principal component analysis applied in reciprocal space to Gaussian mixtures of signals (Borek *et al.*, 2013) are examples of methods of dimensionality reduction which result in noise-filtering and can provide a full answer if they identify a single component of signal variation. When analysing multidimensional signals such as specific radiation-induced changes for which we expect a non-Gaussian distribution in real space, other methods such as independent component analysis (ICA) might be better suited. Here we present and discuss two examples of using ICA to analyse patterns of X-ray-induced specific radiation damage.

2. Methods

2.1. Crystallization, data collection and data processing

In this analysis, with the exception of the 100 K data set for glucose isomerase (GI), previously collected data sets were used (Borek *et al.*, 2007, 2013; Banumathi *et al.*, 2004). However, a description of the crystallization and data collection conditions is provided here, as this information is important for understanding the results presented in this work.

All proteins were crystallized by the vapour diffusion method, as described before (Borek *et al.*, 2007, 2013). Briefly, 2 µl of thaumatin (Sigma) solution of 36 mg ml^{-1} in 29 m*M* HEPES, pH 7.0, and 10 m*M* $CaCl_2$ were mixed 1:1 with 2 µl of well solution (v = 0.5 ml) containing 0.75 *M* KNa tartrate, 0.1 *M* citrate buffer, pH 6.5, and 10% (*v*/*v*) glycerol to form 4 µl drops. A crystal of dimensions 0.15 mm × 0.15 mm × 0.25 mm was cryo-protected by dipping it in well solution supplemented to a final concentration of 27% (*v*/*v*) of glycerol.

Glucose isomerase from *Streptomyces rubiginosus* (Hampton Research) was dialyzed several times against dH_2O and concentrated to 25 mg ml^{-1}. GI crystals of about 0.25 mm × 0.25 mm × 0.2 mm dimensions grew from drops (2 µl) consisting of 1 µl of protein solution diluted at a 1:1 ratio (*v*/*v*), with 1 µl of reservoir solution (1 ml) consisting of 0.1 *M* $CaCl_2$, 16–22% MPD (*v*/*v*) and 0.1 *M* Tris-HCl, pH 7.0. Crystals for data collected at 15 K, 40 K, 80 K and 130 K were lifted from crystallization drops and cryo-cooled in liquid propane without adding additional cryo-protectant other than the MPD already present in the crystallization solution. Then, for collecting the data sets at 15 K, 40 K and 80 K, the crystals in the solidified propane were placed in a 100 K N_2 cryo-stream to allow for thawing of the propane, which has a melting temperature of 85 K. This was followed by placing the crystals in liquid nitrogen and remounting them in a setup with a helium cryostat (Hanson *et al.*, 1999). The crystal of GI that was measured at 100 K was initially cryo-cooled by plunging it in liquid nitrogen. Although we tried to remove manganese ions by dialysis in all cases, both difference $2mF_o - DF_c$ and anomalous difference maps unambiguously identified incompletely occupied manganese ions bound to the protein, either carried over from the protein purification or contributed by impurities of the salts used in the crystallization.

The GI data sets were collected at the sector 19ID beamline at the Structural Biology Center in the Advanced Photon Source, Argonne National Laboratory, Argonne, IL, USA. A separate single crystal was used for each temperature point. Data sets for crystals cooled to 15 K, 40 K and 80 K were acquired with an open helium cryostat (Hanson *et al.*, 2001), while data sets for crystals cooled to 100 K and 130 K used a nitrogen cryostat. All data sets besides the 100 K data set, which was collected with the Pilatus 3X 6M detector, used the SBC detector (Naday *et al.*, 1998; Westbrook & Naday, 1997). All data were indexed and integrated using the *HKL2000* suite of programs (Otwinowski & Minor, 1997; Otwinowski *et al.*, 2012). The mosaicity of crystals varied between 0.2° and 0.4°, indicating a good microscopic order of the crystal lattice. To include recent advances in data scaling (Alkire *et al.*, 2016; Otwinowski *et al.*, 2012) and to assure the consistency of the reciprocal lattice definition, the previous GI data sets (Borek *et al.*, 2007) were rescaled using *HKL2000* (Otwinowski & Minor, 1997) with the 15 K data set used as a reference point. The aim of this operation was to assure that the grids used to calculate the electron density maps in real space, which are the input to the ICA procedure described in §2.3, have the same dimensions. For all temperatures other than 100 K, the same number of data sets as in the earlier analysis was used (Borek *et al.*, 2007), *i.e.* 6 × 240° for 15 K, 8 × 240° for 40 K, 4 × 240° for 80 K and 4 × 240° for 130 K, while for the 100 K data set we acquired one data set consisting of 500 × 0.36° data. The data collection and re-scaling statistics are summarized in Table 1.

The 22 consecutive diffraction data sets acquired from the single crystal of thaumatin were measured at the National Synchrotron Light Source, Brookhaven National Laboratory, beamline X9B, using the ADSC Quantum 4 CCD detector and the crystal was cooled at 100 K with an Oxford Cryosystems device. All data sets were indexed, integrated and scaled with the *HKL2000* suite, as described earlier (Borek *et al.*, 2013; Banumathi *et al.*, 2004) and kindly provided by Dr Zbyszek Dauter as a set of 22 .sca files representing merged data, each file corresponding to a single data set. We scaled these 22 .sca files together to assess the signal-to-noise ratio prior to the ICA and to estimate the scaling *B*-factor change during the experiment. The resulting scaling statistics are included in Table 1 so that the text can be followed without referring to the previous work.

2.2. Dose considerations

The dose estimations were derived from the scaling *B*-factor (Otwinowski & Minor, 1997), also called B_{rel} (Kmetko *et al.*, 2006; Borek *et al.*, 2007) (Table 1). A change in B_{rel} of 1 Å^2 corresponds to a ∼1 MGy dose at 100 K, as determined by Kmetko *et al.* (2006). The scaling *B*-factor, particularly when

Table 1
Data collection and processing statistics for GI and thaumatin.

	Multi-temperature glucose isomerase					Dose-fractionated thaumatin
Temperature (K)	15	40	80	100	130	100
Wavelength (Å)	0.9792	0.9792	0.9792	1.5406	0.9792	0.9792
Space group	*I*222					$P4_12_12$
Unit cell (Å)	a = 92.85	a = 92.85	a = 92.86	a = 93.03	a = 93.05	$a = b$ = 57.75
	b = 98.06	b = 98.05	b = 97.78	b = 97.78	b = 98.30	c = 150.11
	c = 102.28	c = 102.32	c = 102.41	c = 102.50	c = 102.50	
Scaling B-factor change ΔB (Å^2)	0.52	0.91	0.88	0.4	1.22	3.0
Dose estimation based on scaling B-factor corrected for temperature dependence (MGy)	0.74	0.99	0.50	0.4	0.50	3.0
Resolution of data used in the ICA analysis (Å)	50.0–1.70 (1.73–1.70)					50.00–1.45 (1.46–1.45)
Number of data sets and data ranges used in data analysis	6 × 240°	8 × 240°	4 × 240°	1 × 180° (500 × 0.36°)	4 × 240°	22 × 90°
No. of unique reflections	49609/2567	51581/2683	51081/2676	49983/2500	51588/2667	45611 (1063)
Completeness (%)	96.2/100	99.9/100	99.1/100	99.9/99.6	99.4/98.8	98.6 (95.6)
$\langle I\rangle/\langle\sigma_I\rangle$/last shell	125/58	184/68	81/42	61/34	54/29	179/18
Multiplicity of observations/last shell	43.1/37.0	70.8/61.2	35.3/28.4	6.5/6.5	38.6/39.8	41.1/42.0
$R_{\rm pim}$ (%)	0.6/1.3	0.3/1.0	0.8/1.3	1.1/2.4	1.1/1.8	0.4/5.1
CC1/2	1.000/0.999	1.000/0.999	1.000/0.999	0.999/0.997	0.999/0.998	1.000/0.993

applied to complete data sets, remains an excellent proxy of dose, not only because it avoids complications related to uncertainties of the crystal size and composition, beam size and beam profile, all of which can introduce significant uncertainty into the absolute dose estimations, but also because it is calculated together with the other scaling corrections, so that the data used to estimated $B_{\rm rel}$ values are already corrected for systematic effects other than decay.

As shown in Table 1, the estimated doses for the GI data acquired at 15 K, 40 K, 80 K and 130 K varied between 0.5 and 1 MGy. Those data sets were acquired under exactly the same conditions, on the same day and with a very stable beam of a size larger than the size of the crystals. The differences in the scaling B-factor between the data sets result from the dependence of overall decay on temperature, details of which have been determined in our previous work (Borek *et al.*, 2007). The 100 K data set was collected many years later with a different experimental setup. However, the scaling B-factor has a very low uncertainty and the estimate of 0.4 MGy is consistent with the changes observed in the data and in real space. The signal-to-noise level in the 100 K data set is lower, presumably due to the presence of systematic errors that could not be filtered out because of the modest multiplicity of observations (Table 1). We included this data set to show how robust ICA is to the presence of noise in data.

For thaumatin, the scaling B-factor increase for all 22 data sets acquired at 100 K is about 3.0 Å^2, which corresponds to 3 MGy. The much higher dose of 15 MGy, provided in the original work (Banumathi *et al.*, 2004), was based on calculations using nominal beamline flux. The discrepancies between nominal and actual beam flux are a common problem.

As we discuss in §2.3, ICA is scale-independent and so uncertainties in dose determination have no bearing on the results of the analysis presented here. All inputs to ICA and the resulting ICA maps can be found in the supporting information (Table S1).

2.3. Independent component analysis

All data scaling programs correct for resolution-dependent decay of intensity, which represents a component of radiation-induced changes that is global in reciprocal space and uniform in real space (Borek *et al.*, 2007, 2010, 2013). Therefore, the intensity decay component is effectively excluded from the ICA and its interpretation in real space which we performed here.

Regarding the specific radiation damage, more than one component of specific radiation-induced changes is needed for ICA or else the result is equivalent to the results of SVD. We performed two variants of ICA. The first analysis was performed with a group of data sets collected at different temperatures for GI (Borek *et al.*, 2007), where a single (linear) component of radiation damage was identified at each temperature separately. However, we hypothesized that these single components, analysed across multiple temperatures, could be expressed as the sum of temperature-dependent and temperature-independent contributors. The second type of analysis was carried out for the case of radiation damage in thaumatin, in which, in an earlier analysis in reciprocal space, we found by means of SVD two components, both representing specific radiation-induced changes occurring at different rates with respect to dose (Borek *et al.*, 2013). Our goal here was to interpret these components in terms of their radiation chemistry in real space.

The exact value of dose in our experiments was not estimated directly from the experimental parameters, as described in §2.2. However, this does not affect our analysis, because our SVD and ICA are not dependent on the scale of the data provided as input, although each method is independent in a somewhat different manner. The inverse of the dose value contributes to elements of the matrices in both decompositions as a scaling factor of the input data. ICA is completely independent of the relative scales of inputs, *i.e.* the same

decomposition to components will be obtained regardless of relative errors in scale (dose) estimation. SVD is somewhat sensitive to the relative scale factors in the input data. However, SVD was not used as a separate data mining technique for the GI data sets, while, for the thaumatin data sets, uncertainty about the relative scale factor was the same for all data sets, as they were scaled together before SVD was applied. That type of overall scale factor uncertainty does not affect SVD results. In contrast to ICA, the dose estimation may affect SVD when each data set has its own uncertain scale factor derived from dose estimation. For such cases, there will be a change in the relative weights associated with the contribution from each data set used in SVD, but even that type of uncertainty would not generate additional components; it would only change the relative contribution of inputs to the components.

To facilitate our analysis, we calculated consensus difference maps representing specific changes induced by radiation damage. These maps were calculated by analysing, for each reflection separately, the dose-dependent change of intensity already corrected for overall decay (Borek *et al.*, 2007, 2010, 2013; Otwinowski *et al.*, 2003). The dose-dependent change of intensity, or equivalently the change in structure factor amplitudes ($I\tilde{F}^2$), can be expressed either as the first derivative ($\mathrm{d}F/\mathrm{d}D$) or higher-order derivatives ($\mathrm{d}^2F/\mathrm{d}D^2$ or $\partial^2F/\partial D\partial T$), where F is the structure factor amplitude, D is dose and T is temperature. The fast Fourier transform (FFT) of such derivatives, calculated separately for each index and multiplied by $\exp(i\varphi_c)$, will result in electron density maps that represent rates of change of electron density with respect to dose or to dose and temperature ($\mathrm{d}\rho/\mathrm{d}D$, $\mathrm{d}^2\rho/\mathrm{d}D^2$, $\partial^2\rho/\partial D\partial T$). *Scalepack* can produce such derivatives for each index based on merging multiple observations. Additionally, higher-order derivatives or some combinations of them may appear in SVD and ICA, in which *Scalepack* outputs are used as an input. The interpretation of results might become challenging for more complex combinations of such derivatives, but the Taylor series property allows us to simplify the interpretation. We can express electron density corresponding to such a combination of derivatives as a Taylor expansion at the dose at which the derivative terms cancel and call such a component an extrapolation to that dose. Therefore, here we refer to maps as native-like if they represent the electron density at some dose value, while maps derived directly from derivatives are called difference maps because, despite their being in principle differential maps, their physical interpretation is the same as the change in electron density. If the specific radiation-induced changes are linear with dose (Borek *et al.*, 2013), a $\mathrm{d}\rho/\mathrm{d}D$ map describes simply localized differences between the most and the least damaged state of the crystal lattice. The positions of the peaks in this map describe which atoms undergo the specific radiation-induced reactions, while the heights of the peaks correspond to the relative rates of electron density changes with dose and they are proportional to the changes in atom occupancies. Non-linear components of the $\mathrm{d}^2\rho/\mathrm{d}D^2$ type should be interpreted according to the definition of the second derivative. The positions of peaks in this map describe which atoms undergo specific radiation-induced reactions, while the heights of the peaks correspond to acceleration and/or deceleration of the rates of the specific radiation-induced electron density changes with dose.

For the multi-temperature ICA analysis for GI, five difference maps ($\mathrm{d}\rho/\mathrm{d}D$) representing specific radiation-induced changes at 15 K, 40 K, 80 K, 100 K and 130 K were generated. We did not include $2mF_o - DF_c$ coefficients in this analysis because in this case the data were collected from different crystals and the primary interest was in the changes of radiation-induced patterns rather than in minor structural effects which may have resulted from variability in cryo-cooling and other variance in the experimental conditions. Initially, the 1XIB Protein Database (PDB) model was refined against 1.1 Å data merged from three different GI crystals not included in the analysis presented here. These were collected at 15 K and 100 K and have data of very good quality, with minimal radiation damage. That re-refined 1.1 Å model was used to calculate phases and these phases were used to calculate separate maps with amplitudes from 15 K, 40 K, 80 K, 100 K and 130 K. The *R*-factor values without additional temperature-specific refinement were in the range 11.9% to 13.3%, while the *R*-free factors were ∼0.3% higher for every data set.

To confirm the lack of non-isomorphism between crystals measured at different temperatures, we performed the full refinement to convergence using *REFMAC* (Murshudov *et al.*, 2011) with the reference model. The reference model was obtained by refining the initial 1XIB-based 1.1 Å model, from which the phases for the ICA were derived, against a 1.7 Å set of structure factor amplitudes, obtained by averaging intensities from data sets acquired at all temperatures. The model was manually rebuilt and inspected with *COOT* (Emsley & Cowtan, 2004) and validated with *MolProbity* (Davis *et al.*, 2004, 2007). The refinement and validation statistics are provided in Table 2. This refinement was followed by five separate refinements to convergence, each with the same set of *R*-free reflections as the one used for the refinement of the reference model, against the structure factor amplitudes corresponding to the specific temperatures. The *R* and *R*-free values for each model are tabulated in Table 2. *LSQKAB* (Kabsch, 1976) was then used to assess differences between temperature-specific models and the reference model. The RMSD values (Table 2) vary from 0.029 to 0.061 Å and correspond to 1.7% to 3.6% of the diffraction resolution (1.7 Å). These values of RMSD together with the estimated overall coordinate error (ESU, Table 2) of 0.039 to 0.077 Å indicate that the models and the data sets corresponding to these models are highly isomorphous. For that reason, we have deposited a single model, the reference model, together with six sets of structure factor amplitudes: five corresponding to data sets acquired at separate temperatures and one corresponding to the averaged data set (PDB code 5VR0).

For the thaumatin case, in which 22 data sets were collected on a single crystal, four maps were included in the ICA: two of the $\mathrm{d}\rho/\mathrm{d}D$ / $\mathrm{d}^2\rho/\mathrm{d}D^2$ type and two of the $2mF_o - DF_c$ (ρ) type. The first two maps were calculated using amplitudes defined

Table 2
Refinement and validation statistics for the glucose isomerase models refined against structure factor amplitudes merged across all temperatures and at separate temperatures.

The RMSD values between models refined against data sets acquired at different temperatures and the deposited model refined against the temperature-averaged data set serve as confirmation of the isomorphism between different crystals. FOM values provide confirmation that the phase error is minimal.

	Multi-temperature glucose isomerase models					Averaged model
Temperature (K)	15	40	80	100	130	All merged together
Refinement						
Resolution range (Å)	50.01–1.70	50.01–1.70	50.01–1.70	50.01–1.70	50.05–1.70	50.01–1.70
No. of unique reflections (test set)	46918 (2433)	48550 (2530)	48458 (2530)	47448 (2475)	48591 (2531)	48877 (2547)
Mean *B*-factor (Å^2)	10.52	10.59	10.43	13.69	10.49	11.04
R_{work}/R_{free} (%)	12.79/15.39	12.77/15.10	12.83/15.30	12.64/15.20	13.02/15.71	12.60/15.07
FOM	0.9283	0.9319	0.9264	0.9333	0.9177	0.9315
R.m.s deviations						
Bond lengths (Å)	0.009	0.009	0.009	0.009	0.009	0.009
Bond angle (°)	1.403	1.402	1.407	1.397	1.416	1.401
Estimated overall coordinate error based on *R*-free value (Å)	0.077	0.073	0.074	0.075	0.076	0.073
Estimated overall coordinate error based on maximum likelihood (Å)	0.040	0.040	0.040	0.039	0.041	0.039
Validation						
Molprobity score	1.48	1.51	1.49	1.50	1.50	1.51
Molprobity clashscore	5.16	5.31	5.00	4.85	5.47	5.31
Ramachandran outliers (%)	0.3	0.3	0.3	0.3	0.3	0.3
Ramachandran favoured (%)	96.9	96.6	96.6	96.4	96.9	96.6
Non-isomorphism assessment						
RMSD against the averaged model (Å)						
All atoms	0.036	0.035	0.040	0.061	0.038	N/A
Protein atoms	0.029	0.029	0.036	0.043	0.033	N/A

by the first and the second eigenvectors, obtained with SVD (Borek *et al.*, 2013). Each eigenvector had contributions from linear and second-order non-linear specific radiation-induced changes with dose. Additionally, we included two sets of $2mF_o - DF_c$ coefficients, calculated by *REFMAC* (Murshudov *et al.*, 1997, 1999, 2011) using one of the best models representing the cryo-cooled structure of thaumatin (PDB 1RQW) and structure factor amplitudes extrapolated to two different doses: one close to zero-dose and the second close to mid-dose (∼2 MGy). There was a noticeable decrease of *R*-free values for both sets, from an *R*-free of 17.8% to 14.0% for the zero-dose extrapolated amplitudes, and from an *R*-free of 19.5% to 14.5% for amplitudes corresponding to mid-dose. Thus, the input to ICA consisted of four electron density maps. We expected to obtain three significant components from the ICA analysis corresponding to the zero-dose extrapolated map and two components of specific radiation-induced damage. The inclusion of two states corresponding to different doses allowed for testing how noise affects ICA.

Files containing scaled intensity values corresponding to all maps were transformed to structure factor amplitudes using *CTRUNCATE* (French & Wilson, 1978), and binary map files covering the entire asymmetric unit were calculated using the same grid for each data file: 120 × 120 × 312 for the thaumatin data sets and 168 × 176 × 180 for the GI data sets, in both cases using phases derived as described above. To assure data were uniformly measured with sufficient signal-to-noise, the resolution limit for the ICA analysis was set to 1.7 Å for GI and 1.45 Å for thaumatin. The map files, upon removing the header, were combined into a matrix, such that each map file was a row in the new matrix, and this matrix was then used as an input into the *FastICA* package, version 2.5 [http://research.ics.aalto.fi/ica/fastica/ (Hyvärinen, 1999)] within Matlab R2016a (*MATLAB and Statistics Toolbox Release 2016a*, The MathWorks, Inc.) with the default settings for eigenvalue filtering. The separating matrix (**W**) and the estimated independent components resulting from it were calculated by minimizing the mutual information so that the components were as independent as possible (Hyvärinen, 1999). The number of independent components to be calculated was set to the number of rows in the matrix using the optional 'numOfIc' parameter in the routine, but the output was limited to the number of components that could be identified. The *FastICA* package was applied to data from GI (five maps) and thaumatin (four maps). For thaumatin, three components were obtained, as expected. For GI, ICA identified only two components, even when asked for more. Each component, represented by the set of values on the same grid that was used in the original map calculation, was transformed back to a format that could be displayed in *COOT* (Emsley & Cowtan, 2004) by converting the data to a binary format with a CCP4 header added by a short custom-written procedure. In addition, pairs of components were displayed in scatter plots, one component against another for all points in the asymmetric unit. We inspected the results for two target functions in *FastICA*: kurtosis and skewness. The choice of the target function in *FastICA* is driven by noise considerations and the nature of the components for which a search is being made.

The more nonlinear target function (kurtosis) makes ICA less noise sensitive, but may overweight common large deviations in what are sought as independent components. For example, this could result from heavy atoms contributing large signal both to native structures and to the specific radiation change, which was definitely true for the thaumatin case. In such a case, skewness, being a less nonlinear target function, provided a more cleanly interpretable result. In the field of ICA, even less nonlinear target functions than skewness are used, but when they were applied we found them to be too sensitive to the amount of noise present in the data. We also analysed the two-dimensional scatter plots to assess whether ICA produced statistically independent components. The essence of using ICA is to adjust the target functions to make the results most interpretable. The amount of bias this can introduce is limited by the very small number of choices for target functions.

2.4. Other components of data analysis and interpretation

Structural figures were made with *PYMOL* (Schrodinger, 2015), with labels added in Adobe Illustrator (Adobe Systems Inc.). Figs. 2 and 4 were made with MATLAB and EXCEL (Microsoft Corp.). All figures were labelled and resized in Adobe Illustrator.

3. Results and discussion

3.1. ICA as a method of real-space feature extraction

ICA transforms a mixture of signals by decomposing it into components that are as independent as possible. This is done by optimizing selected higher-order moments of the signals' distributions, *e.g.* skewness (third moment of a distribution) and kurtosis (fourth moment of a distribution). In principle, ICA produces results that are an approximation of results obtained with maximum entropy (Cichocki *et al.*, 1998; Learned-Miller & Fisher, 2004; Hyvärinen, 1999). Optimization of statistical independence means that ICA can only be used for linear or approximately linear mixtures of non-Gaussian signal distributions (Hyvärinen, 1999). In contrast to ICA, SVD analysis targets data variance (second moment of a distribution), and so it decorrelates mixtures of Gaussian distributions well; however, it is less appropriate for mixtures of non-Gaussian signals. Nevertheless, even with such mixtures, SVD can still be used to assess a number of measurable components present in the data. For this reason, SVD is frequently a prerequisite for the ICA procedure as the data-whitening step (Vicente *et al.*, 2007), and it is employed in the *FastICA* procedure that was used here. Feature extraction in ICA is achieved by optimizing contrast in the component while minimizing mutual information between components, which is different from the condition of orthogonality for components generated by SVD. Unless skewness is used as a target, the sign of an ICA-identified component is arbitrary as it is in SVD analysis, and, similarly to SVD and other non-parametric methods, ICA-derived components require physical interpretation.

In crystallography, the choice between SVD and ICA is dictated by properties of the expected signal in either real or reciprocal space. By Parseval's theorem, the second moments of distribution are preserved by the Fourier transform (FT), and therefore SVD performed in real and reciprocal space produces equivalent results represented by the same number of components. Thus, changes to amplitudes in reciprocal space resulting from experimental errors, due to this FT property, cause changes in real space that are scrambled over the whole volume of the unit cell. We expect that, due to the central limit theorem, these experimental errors will have a Gaussian distribution in real space. However, the higher moments of distribution used in ICA are not preserved by the FT, and therefore the space appropriate for data analysis needs to be chosen by taking into account the nature of the signals' mixture.

Signals arising from chemical events such as specific radiation-induced changes are expected to be non-Gaussian in real space. Firstly, non-Gaussianity arises from the presence in the crystal lattice of ordered atoms in combination with areas of bulk solvent. Secondly, most chemical signals will be localized to specific volumes of the structure, which also results in a non-Gaussian distribution of changes in electron density. Those real-space non-Gaussian signals in SVD analysis will be reduced to their second moments, which will be combined with second moments arising from contributions from experimental errors. Therefore, SVD analysis in such a case may result in components that mix contributions from various sources. This is also because, by definition, SVD does not impose statistical independence of components, but only a lack of correlation between components, *i.e.* only second cross-moments of decomposition are zero. In contrast, in ICA, one of the higher-order cross-moments is required to be zero in the decomposition for all components, because if the analysed signals are statistically independent, a single zero-valued higher cross-moment implies that all others must be zero as well. For Gaussian errors, the expected values of all higher moments of the distributions are zero, and therefore ICA is also less sensitive to contributions from errors. Due to these two properties, *i.e.* resistance to errors and statistical independence, ICA is more suitable for the analysis of the mixtures of non-Gaussian signals that was selected here.

In reciprocal space, such mixtures appear rarely. For instance, if multiple crystals with variable twinning fractions are measured, one can expect that the measured diffraction intensities will be suitable for ICA, which in principle in such a case could use all those data sets together and de-twin them with a single-step procedure. However, the more typical crystallographic case suitable for ICA is represented by some localized changes in real space, for which one expects variable patterns of behaviour in different data sets. For instance, studies of how the distribution of rotamers in a particular protein depends on temperature in multi-temperature data sets can create a case suitable for such analysis. A related method is restricting SVD analysis in real space to a region specified by a mask (Rajagopal *et al.*, 2004; Schmidt *et al.*, 2003). This mask could be provided by human input, but this

would introduce subjectivity into the analysis. Alternatively, the mask could be defined automatically from the data variance, and such a procedure would have interesting analogies to ICA.

In principle, very large non-isomorphism, *e.g.* due to a large change in the unit-cell parameters, may become either an ICA component or a contributor to an ICA component. This is not the case in our analysis, because such non-isomorphism would manifest in a very specific manner in electron density maps of ICA components; the difference electron density maps for the majority of the atoms in the structure would show the shift, involving translation and/or rotation, correlated with the unit-cell volume change. In GI, which is crystallized with a tetramer on intersection of two-fold axes in *I*222, the rotational shift is not possible, and we do not observe translational shifts consistent with non-isomorphism. In the case of thaumatin, when multiple data sets have been acquired on the same crystals, significant non-isomorphism was not expected unless different volumes of the crystal were exposed. In any case, we do not observe in any of the maps corresponding to the ICA components features that would indicate the presence of structural non-isomorphism. A method of correcting for structural non-isomorphism has been presented by Ren *et al.* (2013), where such a correction was performed in the context of the SVD-based analysis of native electron density maps.

In our case, ICA identified components that we could interpret as corresponding to specific radiation-induced changes.

3.2. ICA identifies temperature-dependent and temperature-independent components of radiation-induced changes in real space

In the case of GI for data sets measured at 15 K, 40 K, 80 K, 100 K and 130 K, SVD identified only one component at each temperature for specific radiation damage. We attributed the lack of second-order effects to the dose in every data set being rather modest, *i.e.* the scaling *B*-factor indicated a dose below 2 MGy (Table 1), which effectively prevents the appearance of any second-order effects, particularly for a protein such as GI which does not have disulfide bridges or other features indicating sensitivity to specific radiation chemistry. ICA was used only to assess the temperature dependence of the specific radiation damage rates with respect to dose and revealed two components. To perform ICA, for each temperature the maps of specific radiation damage ($d\rho/dD$) (Fig. 1*a*) were calculated separately and introduced into the ICA procedure described in §2.3, which produced two ICA components. Presented here is the interpretation of the results obtained with kurtosis as the target function. Firstly, we visually compared ($d\rho/dD$) maps representing specific radiation-induced changes of electron density at each temperature (Fig. 1*a*) with the map representing the first ICA component (Fig. 1*b*) and the map representing the second ICA component (Fig. 1*c*). This was done as a quick validation that the ICA maps have interpretable features.

Then we analysed the magnitudes of contributions from each data set and their sign [Figs. 2(*a*) and 2(*b*)]. For the first high-contrast ICA component, the contributions from each data set decreased with increasing temperature of data

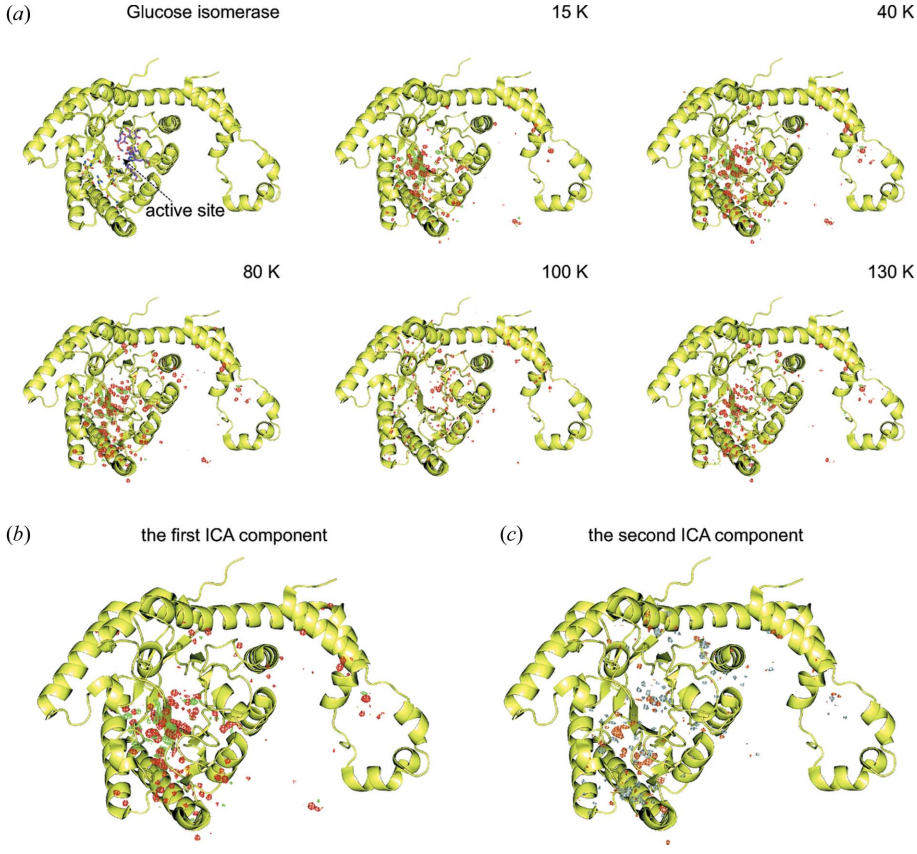


Figure 1
The results of the ICA for glucose isomerase (GI) mapped onto 1XIB. The metal-binding residues in the active site are labelled with magenta, while the residues involved in substrate binding are labelled yellow. (*a*) $d\rho/dD$ maps of specific radiation-induced changes at different temperatures. All maps are contoured at $\pm 5\sigma$, with the exception of the 100 K map, which is contoured at $\pm 4.5\sigma$. Red colour represents loss of electron density and green colour represents gain of electron density. (*b*) Map of the first component of the ICA contoured at $\pm 5\sigma$ with red colour representing loss of electron density and green colour representing gain of electron density. The map represents the common part of the specific radiation-induced changes between all data analysed. As this is the common part between the data collected at different temperatures, it is equivalent to the changes that are independent of temperature, so it is interpreted as a map of ($d\rho/dD$) type. (*c*) $\partial^2\rho/\partial DT$ map of the second component of the ICA, contoured at $\pm 4\sigma$ and representing those specific radiation-induced changes that vary with temperature. Grey colour denotes the decrease in the rate of specific radiation damage with change of temperature, while orange denotes the increase in the rate of specific radiation damage with change of temperature.

collection and all contributions have the same sign. Based on this analysis, we concluded that the first higher-contrast component represents the temperature-independent signals. Firstly, the decreasing contribution of data sets with increasing temperature of data collection is consistent with the expectation that diffusion, which is temperature-dependent and is severely restricted at temperatures below 100 K, will start contributing more with the increase in temperature at the cost of the temperature-independent component. Secondly, the same sign of contributions from data sets acquired at different temperatures indicates that the first ICA component corresponds to the common part between data sets, which is again expected for the temperature-independent component. The changing signs of the contributions from the specific radiation-induced changes corresponding to different temperatures for the second component indicate that the component represents the differences between the data acquired at different temperatures [Figs. 1(*a*), 1(*c*) and 2(*b*)]. This is consistent with the expectation that the temperature-dependent component will be significantly different between the 15 K temperature and higher temperatures, and therefore we interpreted the second ICA component as temperature-dependent.

We included the 100 K data set, which has a lower signal-to-noise ratio (Table 1), because it provides an excellent test of the method's robustness and its resistance to data contaminated by noise. As Figs. 2(*a*) and 2(*b*) show, the 100 K data set has no contribution to either ICA components. The data set collected at 40 K also contributes little, but for a different reason. It has an excellent signal-to-noise ratio (Table 1) and it is very consistent with both the 15 K and 80 K data sets in terms of the specific radiation damage patterns [Figs. 1(*a*) and 1(*b*)]. This data set does not contribute significantly to the ICA results because it does not carry independent information that would allow ICA to increase the contrast in the two determined components. In other words, the information in this data set is a linear combination of information in the 15 K and 80 K data sets. This property of ICA is very different from SVD, which in such a situation, *i.e.* three highly correlated data sets, would produce a component with approximately equal contributions from all three data sets.

To test our interpretation of the reasons for which the 40 K data set did not contribute significantly, contributions from each data set to two identified ICA components for a less non-linear ICA target, *i.e.* skewness [Figs. 2(*a*) and 2(*b*)], were calculated. The expectation was that, for skewness, the contribution from the 40 K data set to the first component would increase at the cost of contributions from other temperatures. Fig. 2(*a*) shows that such a modest increase indeed occurred.

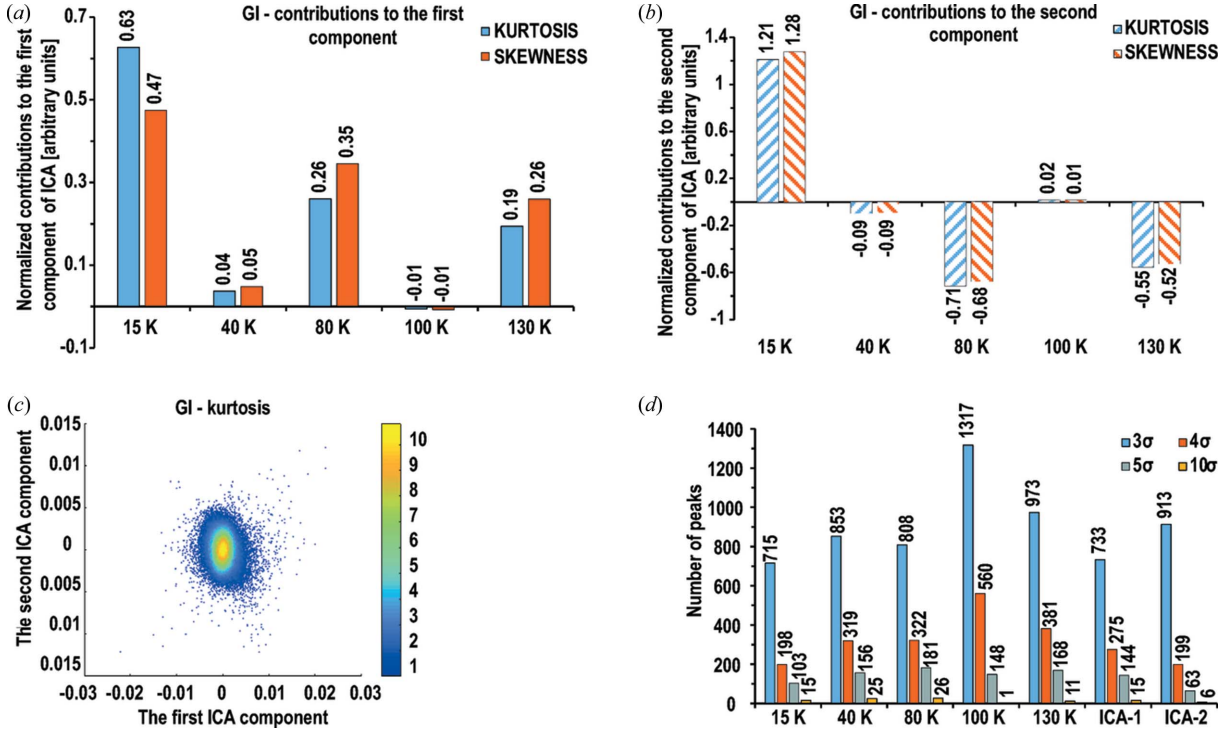


Figure 2
Properties of the ICA analysis for GI. (*a*, *b*) Two plots of normalized contributions from GI data sets to the first (*a*) and second (*b*) ICA components, each for two target functions. The components' maps are normalized to a RMSD of 1 so the *y*-axis of the right-hand plot has a higher value due to the opposite signs of correlated contributors. (*c*) Scatter plot of ICA results for GI with kurtosis as the ICA-optimized target, projected onto the plane defined by the first and the second ICA components. The units are arbitrary because the method is scale-independent. However, their ratio carries information about the relative magnitude of the components. The colour gradient represents the density of the points. The plot shows symmetrical distribution along both axes, which indicates statistical independence between the first ICA component, representing temperature-independent radiation-induced structural changes in the GI structure, and the second ICA component, representing temperature-dependent radiation-induced structural rearrangements. (*d*) Number of peaks at different thresholds present in dρ/dD maps for each temperature separately and for the maps of two ICA components.

Table 3
List of the 15 highest peaks in the maps representing components of the ICA.

The deposits 5VR0 for glucose isomerase and 1RSW for thaumatin were used for numbering the atoms. The numbering in parentheses for GI indicates the number of the corresponding water molecule in 1XIB. The number of peaks exceeding the $\pm 5\sigma$ threshold is listed as Total. NL means not listed. The sign of the effect, which is assigned arbitrarily by ICA, was selected so that '−' represents loss of electron density in the position of an atom and '+' represents gain of electron density in the new position of an atom. The disulfides in thaumatin are numbered as follows: A, C9–C204; B, C56–C66; C, C71–C77; D, C121–C193; E, C126–C177; F, C134–C145; G, C149–C158; H, C159–C164.

Peak	Glucose isomerase		Thaumatin	
	ICA component I	ICA component II	ICA component I	ICA component III
1	H_2O 815 (481, 1XIB) (-22.18σ)	Mn^{2+} 391 (-12.15σ)	C126 Sγ (-45.37σ) (E)	C149 Sγ (-20.68σ) (G)
2	H_2O 515 (585, 1XIB) (-19.93σ)	H_2O 815 ($+11.76\sigma$)	C158 Sγ (-32.84σ) (G)	C159 Sγ (-18.10σ) (H)
3	H_2O 514 (584, 1XIB) (-17.55σ)	H_2O 815 (-11.44σ)	C149 Sγ (-32.40σ) (G)	C177 Sγ ($+15.44\sigma$) (E)
4	T90 Oγ1 (-16.82σ)	Cl^- 15 (H_2O 504, 1XIB) (-10.82σ)	C145 Sγ (-29.61σ) (F)	C177 Sγ (-13.66σ) (E)
5	E141 Oε2 (-16.66σ)	D101 Oδ1 (-10.07σ)	C204 Sγ (-26.79σ) (A)	C66 Sγ (-12.93σ) (B)
6	H_2O 443 (467, 1XIB) (-13.63σ)	S302 Oγ1 (-10.03σ)	C164 Sγ (-24.43σ) (H)	C164 Sγ (-12.37σ) (H)
7	W137 C_{mc} ($+13.25\sigma$)	Mn^{2+} 390 (-9.62σ)	C66 Sγ (-21.42σ) (B)	C164 Sγ ($+11.67\sigma$) (H)
8	W137 N_{mc} ($+12.79\sigma$)	T119 Oγ1 (-9.42σ)	C121 Sγ (-20.90σ) (D)	C149 Sγ (-9.97σ) (G)
9	H_2O 763 (631, 1XIB) (-12.48σ)	Y225 OH (-8.43σ)	C71 Sγ (-17.59σ) (C)	C204 Sγ (-8.35σ) (A)
10	Mn^{2+} 390 ($+12.35\sigma$)	Y225 OH ($+8.06\sigma$)	C159 Sγ ($+15.81\sigma$) (H)	C56 Sγ (-8.29σ) (B)
11	H_2O 399 (406, 1XIB) (-12.31σ)	K149 Nζ (-7.28σ)	C193 Sγ ($+14.06\sigma$) (D)	C193 Sγ (-8.12σ) (D)
12	M370 Sδ (-12.17σ)	M370 Sδ (-7.17σ)	M112 Sδ (-12.87σ)	C158 Sγ (-8.00σ) (G)
13	D101 Oδ2 (-11.47σ)	H_2O 893 (611, 1XIB) ($+6.83\sigma$)	H_2O 1004 ($+12.52\sigma$)	C134 Sγ (-7.49σ) (F)
14	T119 Oγ1 ($+11.32\sigma$)	Y221 Cε1 ($+6.82\sigma$)	D101 Oδ1 (-11.44σ)	M112 Sδ (-6.86σ)
15	T91 Oγ1 (-11.10σ)	T105 Oγ1 (-6.61σ)	C193 Sγ ($+10.12\sigma$) (D)	C177 Sγ (-6.35σ) (C)
⋮				
29	Mn^{2+} 391 (-9.32σ)	NL	NL	NL
Total	144	63	140	40

To assess the statistical independence of the components, two-dimensional scatter plots of the ICA results projected on the space defined by the first and second ICA components (Fig. 2*c*) were calculated. The plot shows a symmetrical distribution along both axes, which confirms statistical independence between the first ICA component and the second ICA component. These analyses were followed by the interpretation of patterns of specific radiation-induced changes on maps for both ICA components.

The 144 peaks exceeding the threshold of $\pm 5\sigma$ in the temperature-independent ICA component clearly identified the area in which radiation-induced changes are clustered [Table 3, Fig. 1(*b*)]. The region of the most significant changes is near the substrate binding site and most of the changes affect ordered solvent molecules [Table 3, Figs. 3(*a*) and 3(*b*)]. The map of the second component that represents specific radiation-induced changes that depend on temperature shows 63 peaks [Fig. 1(*c*)] exceeding the threshold of $\pm 5\sigma$ [Table 3, Figs. 3(*a*) and 3(*b*)]. The second component has lower contrast due to a lower level of signal [Fig. 2(*b*)]. Therefore, the analysis was focused only on the highest peaks for the second component. These largest changes in the second ICA component are clustered in the same area as those for the first component [Figs. 1(*b*) and 1(*c*)]. The temperature-dependent changes sometimes have an opposite sign for their peaks when compared with the temperature-independent changes in the same area [Figs. 3(*b*) and 3(*c*)]. This indicates that the reactions that are driving the temperature-independent effects have different rates from the migration of excited states by diffusion. The diffusion may increase specific radiation-induced damage in a particular site, but at the same time may take away migrating charged states from other specific locations. Therefore, the temperature modulation effect can be both positive and negative with respect to how the specific radiation-induced changes proceed with dose.

As noted previously (Borek *et al.*, 2007), Mn^{2+} positions are not the highest peaks on the maps of specific radiation damage; although the peaks localized at their positions are significant in the map of the first ICA component ($+12.35\sigma$ for Mn^{2+} 390 and -9.32σ for Mn^{2+} 391), they are more pronounced in the map of the second ICA component (-9.62σ for Mn^{2+} 390 and -12.15σ for Mn^{2+} 391). The map of the first ICA component also contains a negative peak at -8.1σ for Mn^{2+} 390 near the positive $+12.35\sigma$ peak, with the position of Mn^{2+} 390 refined in between two peaks.

The observed peaks are consistent with Mn^{2+} 390 undergoing a small shift in position due to local reorganization of the hydrogen bond network in a temperature-independent manner. The data are not conclusive regarding the change of redox state of this manganese cation, but it is likely that the shift is accompanied by reduction because the positive peak is higher than the negative one. With increased temperature, the shift of Mn^{2+} 390 is less pronounced and the reduction seems to be stronger, as the peak associated with this cation is one of the strongest ones on the map of the second component. The behaviour of Mn^{2+} 391 is more difficult to interpret because this peak is not very significant in the map of the first temperature-independent component and it has a strong temperature dependence.

The surface electrostatic potential in the area affected by specific radiation damage in the structure of GI indicates a strong negative potential arising from six carboxylic acids that cover the binding site and coordinate two Mn^{2+} cations (Nicoll *et al.*, 2001). Although the two Mn^{2+} cations provide a +4

charge together, the carboxylic acids tilt the electrostatic balance towards the negative potential. The negative electrostatic potential of the active site results in the expectation that the relative contribution to specific radiation-induced changes in this area from electrons and other negatively charged species will be lower than the contribution from positively charged species.

Additionally, any analysis of the potential mechanisms behind the observed patterns of specific radiation-induced changes needs to be consistent with the relative changes in occupancies of the affected species. The peaks on the ICA maps are proportional to the number of electrons, so for the analysis of relative changes in occupancies one needs to normalize their height by taking into account their atomic form factors. We considered the height of the peaks for O, N and C (Table 3) as directly proportional to the change in occupancy, because these atoms have very similar atomic number. However, when analysing the relative change of occupancy for Mn and S atoms with respect to the occupancies of O, N and C atoms, we considered the $\sim 3\times$ and $\sim 2\times$ higher atomic numbers for Mn and S, respectively. Therefore, although the Mn^{2+} peaks are statistically significant on the ICA maps, the changes in their occupancies are only $\sim 15\%$ of the change in the occupancies of the most affected solvent molecules listed in Table 3. In addition, the solvent molecules directly coordinated by Mn^{2+} have not been affected by specific radiation changes significantly, and therefore we concluded that reorganization at Mn^{2+} is not a dominant driver of the changes that we observed.

The area neighbouring the substrate binding site, with Thr90, Thr91, Thr119 and Trp137 (Fig. 3*b*), holds the highest peaks on the maps of the first and second ICA components. All the highest peaks are clustered at and near several water molecules, which in turn interact through WAT457 with Glu141. One possible interpretation of the patterns observed in this area described by the first ICA component is radiolysis of the neighbouring side-chains of Thr90, Thr91, Thr119 and decarboxylation of Glu141. These processes might be driven by radicals generated from solvent molecules as discussed later in this section. Such radiolytic reactions would destabilize the network of hydrogen bonds in the water clathrate (H_2O: 399, 443, 514, 515, 815), which would lead to a correlated dispersal of the molecules involved in the clathrate [Figs. 3(*a*) and 3(*b*)]. Another possibility would be to consider these patterns as arising directly from the action of secondary electrons (LEEs, in some fields called ballistic electrons). However, this hypothesis is inconsistent with the observed patterns of specific radiation-induced changes in our ICA

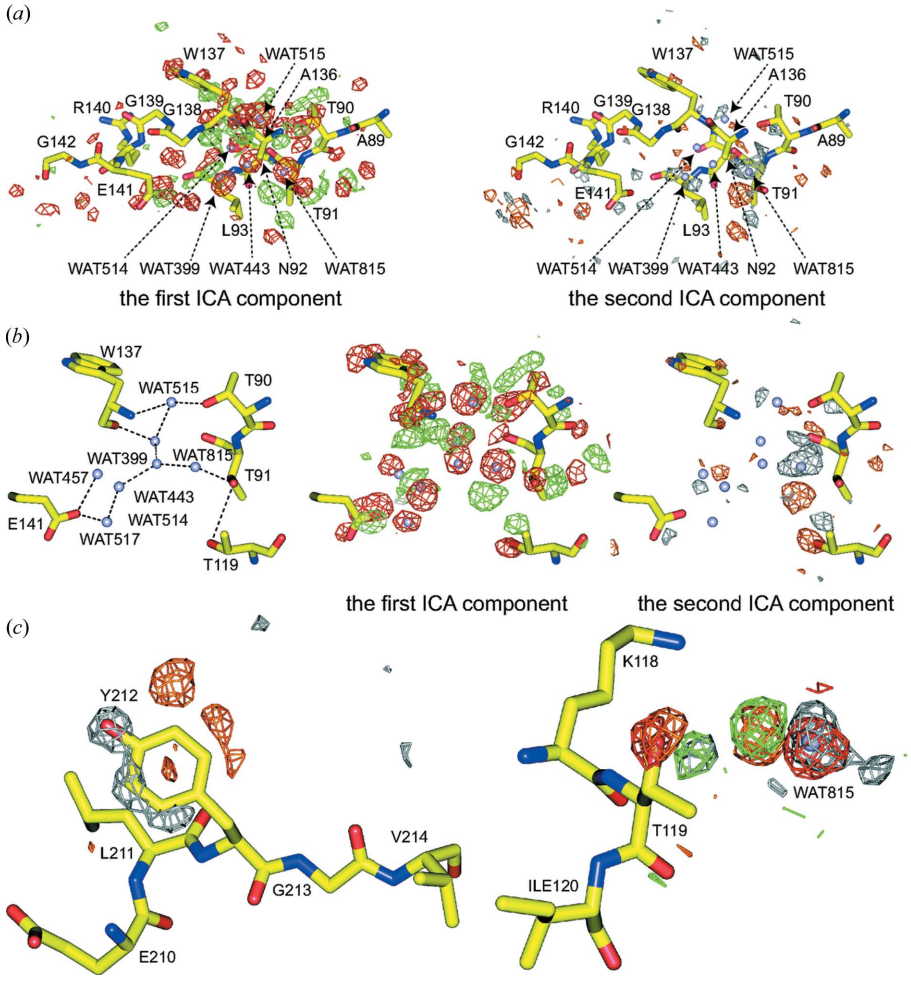


Figure 3
(*a*) The radiation-induced temperature-independent reorganization of the main chain near the substrate-binding cavity for GI. The colour scheme is the same as in Fig. 1. All changes represented by the first ICA component are contoured at $\pm 5\sigma$, and all changes represented by the second ICA component are contoured at $\pm 4\sigma$. The panel on the left shows the first component of the ICA decomposition that corresponds to the specific temperature-independent radiation-induced changes. (*b*) Close-up of the changes in the water clathrate near the most affected area. Damage to Thr90, Thr91 and Thr119 and decarboxylation of Glu141 may induce rearrangements of water molecules that result in the movement of the main chain and side-chains near Trp137. The panel on the left shows this volume without electron density maps, while the panels in the middle and on the right show the maps corresponding to the first and second components of the ICA mapped on the same volume. (*c*) Two examples of the redistribution of specific radiation-induced effects with temperature that may have consequences for the map interpretation. The panel on the left shows the Tyr212 side-chain which undergoes transition to a new position in a temperature-dependent manner. The panel on the right shows the vicinity of WAT815. When both the first and second components of ICA show the same sign, *i.e.* the water molecule transitions towards Thr119 in this case, it means that the movement of electron density is attenuated with temperature. The changes around the OH of Thr119 have opposite signs. The temperature-independent changes represented by the first ICA component show the transfer of electron density represented by the OH position of Thr119 towards WAT815. The local anti-correlation with the second component indicates that this process is enhanced by temperature.

maps. In particular, we do not observe the expected correlations between specific radiation-induced changes and absorption cross sections of damaged atoms and between patterns of damage and local electrostatics (Alizadeh & Sanche, 2012*a*). This indicates that mechanisms of reactions leading to observed patterns have to involve some form of transport for electron-gain and electron-loss centres. The major electron loss centres created by ionization of proteins are trapped at low temperature (Jones *et al.*, 1987). Therefore, we attribute the observed reactions to hopping of electron holes generated either from solvent (Box, 1972; Box *et al.*, 1969, 1972; Sevilla *et al.*, 1979; Burmeister, 2000; Hidaka *et al.*, 2004) or from a protein chain. Temperature independence narrows the possibilities to H^+ for direct tunnelling and/or hopping for longer distances (Aubert *et al.*, 2000; Winkler & Gray, 2015). Whether the temperature-independent component of the GI ICA analysis represents proton hopping will have to be established in further experiments accompanied by appropriate computational simulations.

3.3. ICA identifies a set of components in multi-data-set experiments for thaumatin

For the case of thaumatin, our goal was to interpret the complexity of the radiation-induced structural changes that may happen with variable rates. Therefore, as described in §2.3, we included as inputs to ICA SVD-identified linear and non-linear components of specific radiation-induced structural changes, and also $2mF_o - DF_c$ maps corresponding to zero-dose and to an intermediate dose of ∼2 MGy, to see how the complexity in terms of signal contrast of the native structure is affected by radiation damage. The significance of this investigation is derived from the signal contrast being highly important for phasing and structure interpretation in crystallography. In particular, it was expected that the zero-dose extrapolated map would be clearly extracted as a result of the procedure, because it was explicitly introduced into the decomposition.

The highest contrast component resulting from ICA corresponds to the initial rate of specific radiation damage, which presents as the fastest changing electron density ($d\rho/dD$) (Fig. 4*a*). In the case of thaumatin, this represents breakage of disulfide bridges (Fig. 5*a*). The variation in the heights of the negative peaks for sulfur positions indicates differences in the rates of disulfide bridge reduction for each of the eight disulfides in the structure (Sutton *et al.*, 2013). Most of these bridges are damaged faster than other parts of the structure, as the list of the peaks indicates (Table 3). In the map of the highest-contrast component, 140 peaks exceeding the threshold of $\pm 5\sigma$ were identified, with the most significant peak reaching -45σ (Table 3).

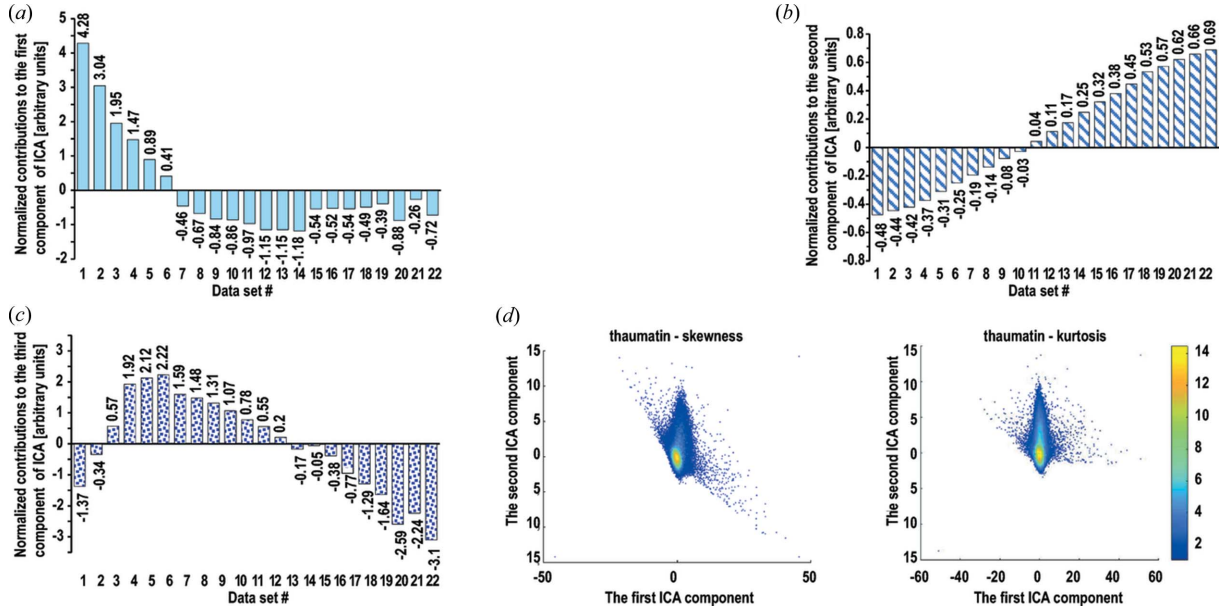


Figure 4
Properties of the ICA analysis for thaumatin. (*a*, *b*, *c*) Three plots of contributions from 22 thaumatin data sets for three ICA components. Dose increases monotonically with the data set number. (*a*) Contributions to the first component indicate that the first component resembles well the ($d\rho/dD$) estimator for initial changes. (*b*) Contributions from each data set to the second component of ICA. The monotonically rising pattern of contributions in combination with significantly higher absolute values for data sets acquired at higher doses indicates that the $2mF_o - DF_c$-type map generated by the second component corresponds to the dose being ∼1.5–2 times higher than the one achieved in the experiment. (*c*) Contributions from thaumatin data sets to the third component of ICA. The parabolic shape of the dependence indicates a large contribution from the ($d^2\rho/dD^2$) (non-linear) estimator. According to that shape, the rate of specific radiation changes with dose imaged by the third component corresponds to the combination of rate changes ($d^2\rho/dD^2$) at higher doses together with initial deceleration ($d\rho/dD$) at higher doses. (*d*) Scatter plots of ICA results for thaumatin with kurtosis (right) and skewness (left) as the ICA-optimized targets. The ICA results are projected onto the plane defined by the first and the second ICA components. The units are arbitrary because the method is scale-independent. However, their ratio carries information about the relative magnitude of components. The colour gradient represents the density of the points. Both targets show a slightly asymmetrical distribution along both axes, which indicates that full statistical independence between the ICA components was not achieved.

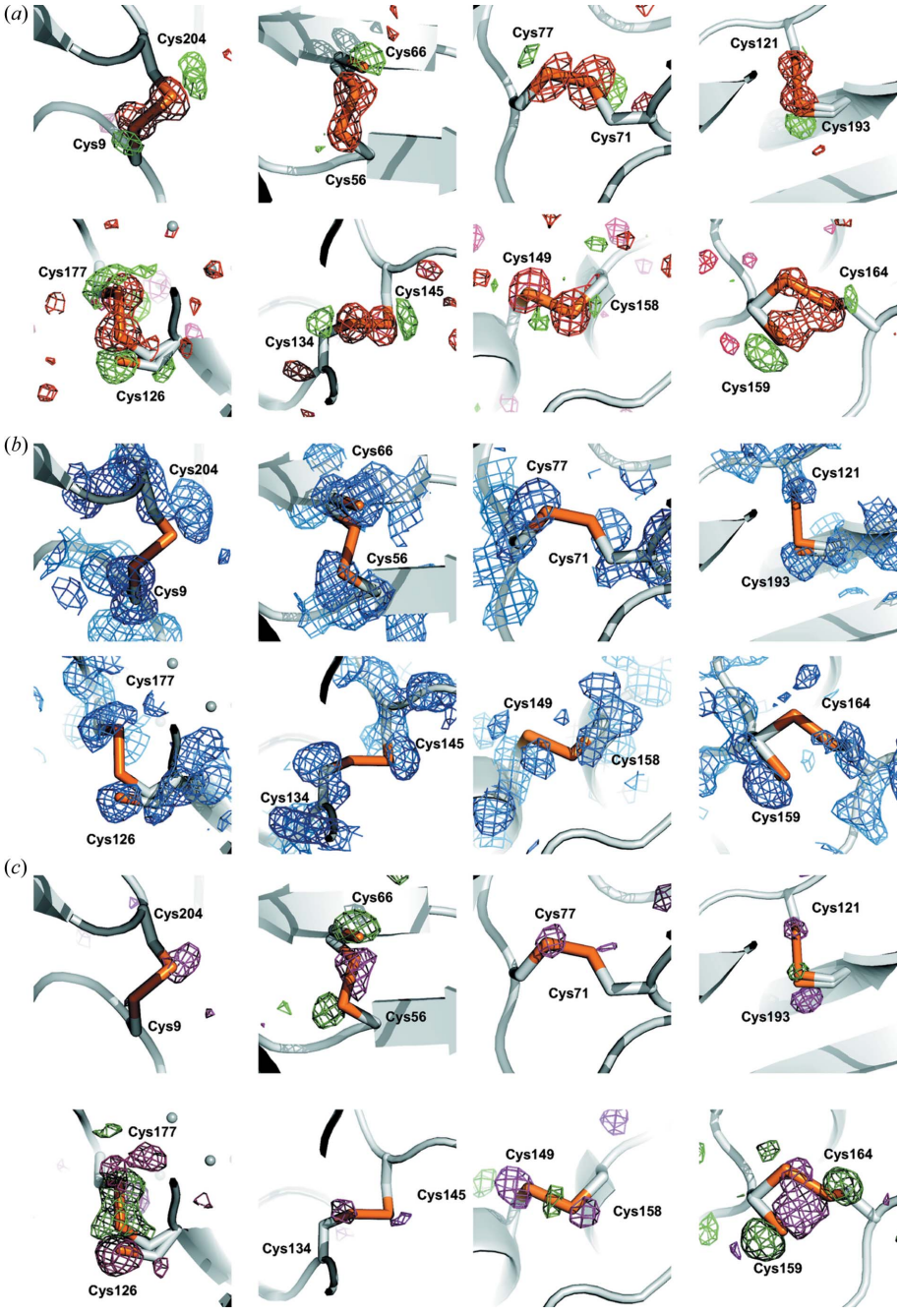


Figure 5
Density maps calculated from the ICA components for thaumatin. (*a*) The first component of ICA for all eight disulfide bridges of thaumatin. The map was contoured at $\pm 5\sigma$ and corresponds to the (dρ/dD) component of the specific radiation-induced changes. The red colour represents loss of electron density, while the green colour represents gain of electron density. (*b*) The second component of the ICA for all disulfide bridges. The map was contoured at $+1.5\sigma$ and coloured blue because it represents a $2mF_o - DF_c$-type map corresponding to the thaumatin structure with reduced disulfide bridges. Therefore, this component corresponds to large dose, as indicated also by Fig. 4(*b*). (*c*) The third component of the ICA. The map was contoured at $\pm 4\sigma$ and corresponds to the non-linear with dose (d$^2\rho$/dD^2) component of the specific radiation-induced changes. The highest non-linearity was expected for the disulfide bridges undergoing the fastest initial damage, described in the map of the first component and in Table 3. Therefore, the colouring scheme for the map of the third component was selected so that magenta shows higher deceleration in the changes of electron density with dose, as described by the first component (*a*), while dark green shows slower deceleration in the changes of electron density with dose, as described by the first component (*a*). At larger doses, it is expected that the map of the third component would correspond to the purer (d$^2\rho$/dD^2) map; however, because ICA did not provide a fully independent component, this map represents a mixture of a small contribution from (dρ/dD) and a large contribution from (d$^2\rho$/dD^2).

An unexpected and interesting result was that ICA identified a $2mF_o - DF_c$-like electron density map that corresponded to a state at a higher absorbed dose than experimentally recorded as the second-highest contrast component [Figs. 4(*b*) and 5(*b*)]. Such a type of map had to appear because of the linear input-output relationships in ICA; however, what was interesting was the stage of the experiment to which it corresponded. This ICA component map results from a linear combination of input electron densities, selected so that the contrast between the protein and the solvent regions is highest. The map shows the structure with all disulfides either fully or partially reduced [Fig. 5(*b*)]. A zero-dose extrapolated map of this type was expected, because in the first approximation the radiation-induced changes should disorganize the structure and we naively anticipated that this would decrease the contrast. The distribution of contributions from each data set to the second component of ICA in Fig. 4(*b*) has two features: (i) the gradient from the first to the last data set and (ii) an excess of positive contributions from more damaged data sets over the negative contributions from initial data sets. These two features together indicate that the map of the second ICA component is extrapolated to a dose higher than the maximum dose in the experiment, *i.e.* 3 MGy, and thus it does not represent any of the $2mF_o - DF_c$ maps which we introduced to ICA.

Upon consideration, it was realized that, first, by applying a resolution-dependent correction for decay, a large contribution to structural disorganization had already been removed from the analysis. The presence of specific radiation-induced changes at the surface of the protein can increase the overall map contrast if such features are being attenuated within the protein core. Cryo-cooling suspends large-scale structural rearrangements within the crystal lattice, but this does not eliminate localized rearrangements. Specific damage that starts with a covalent change at one point in the structure can create a chain reaction of local rearrangements. Our conclusion from this

analysis is that such chain reactions are relatively limited within the structurally constrained protein core, but there is enough energy to rearrange less constrained structures on the protein surface, which can subsequently become trapped in some alternative meta-stable conformation. In cases where there are multiple alternative meta-stable conformations, the statistical average will result in ordered parts becoming specifically disordered upon X-ray exposure. The tendency of these happening on the surface creates an overall increase in contrast between the protein core and the increasingly disordered solvent and solvent-exposed side-chains. Based on the ICA results [Figs. 5(*b*) and 5(*c*)], it was estimated that the increase of the contrast is ∼1% of the native signal at 3 MGy. Such a level of changes is usually too low to be observed as a significant change in $2mF_o - DF_c$ electron density for solvent-exposed side-chains, but the sum of these effects has a high impact on phase improvement procedures that involve solvent flattening. This may be the reason why, in our experience, structure solution methods work very well for radiation-damaged crystals, as such phase improvement can be an important part of the process and partially damaged crystal structures with more flattened solvent may produce more accurate phase estimates, compensating for errors associated with the need for higher exposure.

Because this radiation-induced solvent flattening clearly does not change the hydrophobic core structure, even in the case of a disulfide-containing protein such as thaumatin, the improvement in phasing may appear operationally more important than the increased structural disorder in the solvent-exposed area of the protein. Simplistically, one may think that this represents a choice of improving the flatness of the solvent with a high dose and so improving the intermediate steps of structure solution at the cost of potential loss of information in the result. However, one can always change the so-called native experimental structure factors during the structure determination process. There is no absolute requirement that the same target be used across the whole process of solving a structure, and there may be good reasons why one set of native structure intensities would be used at the beginning and another set at a later stage where the solvent-flattened phases contribute much less to the quality of the maps, and in a typical process these phases may even have no contribution at all.

In addition, in the current experimental reality in which dose-slicing-based data collection can be easily achieved, the zero-dose extrapolation is more stable and accurate than it was in the past. This will facilitate approaches in which, after completing phasing with data corresponding to more exposed structures, the structure can be refined and reinterpreted against zero-dose extrapolated intensities for detailed analysis of biological questions.

The third highest contrast component corresponds to the non-linearity of the specific radiation damage process with dose $d^2\rho/dD^2$ [Fig. 4(*c*)]. For the third component, the most pronounced changes were again observed for disulfide bridges [Fig. 5(*c*), Table 3]. The third component should be interpreted according to the definition of the second derivative, as the rate of change of the rate of change of the electron density with dose. This means that the acceleration and/or deceleration of the rates of the specific radiation-induced electron density changes with dose defined by the first ICA component can be considered. We initially hoped for full separation of disulfide bridge breakage from other slower reactions. The two-dimensional scatter plots were inspected, where changes in component 1 were plotted against changes in component 2, to assess the statistical independence of the identified components for two possible ICA target functions, *i.e.* skewness and kurtosis [Fig. 4(*d*)]. Kurtosis as a target function emphasizes the biggest deviations from a distribution, but, if there are not enough of these deviations, statistical independence may not be achieved. Skewness as a target function emphasizes analysis of deviations for intermediate values, but for that reason analysis based upon it becomes more sensitive to noise in data.

For thaumatin, these plots are not fully symmetrical for the highest peaks, which indicates the partial statistical dependence between the two ICA components, as both are associated with the same sub-structure, *i.e.* disulfide bridges. This is consistent with the disulfide bridges being only partially reduced at the end of the experiments, and with data noise preventing the ICA from extrapolating too far into the larger dose range where they would presumably become fully reduced and the signals would become independent. Therefore, it is expected that extending data collection to larger doses would have resulted in a better separation of the ICA components.

3.4. ICA as a method for visualizing pathways of tunnelling, hopping and diffusion

The ICA analysis of multi-temperature data sets for GI identified two components: one temperature-independent and one temperature-dependent. The independence of the rate of the reaction on temperature is the hallmark of tunnelling reactions and, at zero-temperature, reactions can only occur by direct tunnelling or multi-step hopping.

Although the contribution of quantum mechanical tunnelling and hopping to kinetic or chemical and biochemical reactions is well established (Beratan *et al.*, 1992; Tezcan *et al.*, 2001; Gray & Winkler, 2005; Warren *et al.*, 2012; Winkler & Gray, 2014, 2015), visualizing tunnelling pathways remains a challenge. Our method has the potential to address this problem.

Tunnelling is expected to contribute significantly to mechanisms of reactions that occur in response to X-ray exposure in cryo-cooled crystals (O'Neill *et al.*, 2002). This expectation is based on radiation-induced reactions in crystals proceeding through the redox chemistry that involves radicals, with significant participation of radicals from water radiolysis (Bednarek *et al.*, 1998*a*,*b*; Gupta *et al.*, 2016). Most experiments in crystallography happen nowadays at cryo-temperatures in which diffusion is limited, and that may increase the contribution of tunnelling even further.

In an X-ray experiment, every primary event generates hundreds of secondary ionizations in which electrons and electron holes are formed. These initial events happen in the first approximation uniformly across the crystal space, as non-hydrogen atoms present in the crystal have very similar cross sections for generation of secondary electrons. Secondary electrons do travel distances larger than the unit cell and their trajectories are affected by the charge of the environment, *i.e.* they can come to rest preferentially around positive charges (Alizadeh & Sanche, 2012*a*). In contrast, electron holes, *i.e.* positive charges formed after an electron is ejected, are part of the whole atom, so the momentum that this atom acquires corresponds to much lower energy. Therefore, the migration of holes can only occur by diffusion, tunnelling and hopping. The rates of tunnelling are very strongly tied to the chemistry of interactions in the sample, because these interactions form energetic barriers that need to be overcome for chemical reactions to proceed. In protein crystals, the energetics of localized charges are very strongly tied to the local hydrogen-bonded networks, particularly in solvent areas where rearrangements require only individual sub-angstrom proton motions that may also occur by tunnelling.

Because ICA provides an unbiased method of separating the common part of multiple-temperature data sets (temperature-independent component) from differences in response to temperature (temperature-dependent component), it provides a unique opportunity to generate maps that show what happens in a temperature-independent manner. This landscape can be probed under different chemical conditions, including isotopically labelled crystals, at different temperatures and at different X-ray doses. There is a high probability that such maps will be too rich in detail to be interpretable without additional data. However, when correlated with other information such as computer simulations and mutational studies, they should provide access to currently unavailable information.

Differentiating diffusion from tunnelling contributions requires careful and precise studies of the temperature dependence of the radiation damage process. Multi-temperature data collection extended to very low temperatures combined with appropriate feature selection methods such as the ICA presented here is one possible method that can be used for this purpose. We still need to improve our understanding of the specifics of the involved chemistry, in particular water radiolysis, to better identify potential artefacts in the data and to mitigate their consequences. When the process of water radiolysis is well understood, it will even be possible to take advantage of radiation-induced changes as a source of chemical events (Xu & Chance, 2004) at temperatures below 100 K where tunnelling dominates.

4. Conclusions

The improved ability to collect multiple data sets from macromolecular crystals allows for more elaborate and hierarchical data analysis, which can provide better control over the structure determination of complex molecules. The benefits may include low noise, zero-dose extrapolation of data, analysis of data variance to improve phasing, and better understanding of protein surface chemistry. Having multiple observations of intensity of the unique *hkl* provides new opportunities in data analysis, and specifically permits more complex approaches to decomposition of merged intensities into factors that describe different sources of systematic effects, including different sources of non-isomorphism.

Here we presented an approach that permits systematic inspection of the sources of these effects in real space. In contrast to the SVD method which we described before (Borek *et al.*, 2013; Otwinowski *et al.*, 2003) and which separates data into uncorrelated components, ICA defines the statistically independent components. This is a much more stringent condition than the lack of correlation and allowed new types of questions to be approached. In the case of analysing specific radiation-induced changes in real space, the condition of statistical independence allows for the selection of the features that correspond to the highest contrast in the possible linear data combinations, defined by the distinct components determined by ICA.

Although this work is only an initial exploration into the space of possible questions, we determined that for multi-temperature data sets, ICA clearly reveals two components, with the temperature-independent component being a major one [Figs. 2(*a*)–2(*c*)]. This work introduces the possibility of using multi-temperature crystallography for studying tunnelling and hopping processes. The analysis of dose series that was made for a single crystal of thaumatin revealed differential rates of reactions driven by specific radiation-induced changes in solvent and solvent-exposed areas *versus* the rates of reactions in the core of the structure upon excluding from this core radiation-sensitive elements described by faster progressing processes [Fig. 4(*b*)].

These results also have bearing on the ongoing and unresolved discussion of whether one should model side-chains that do not have clean electron densities due to being highly solvent-exposed or rather trim such side-chains or else model them in multiple conformations. Our results may indicate that the difference in opinions about the visibility of side-chains on the surface may depend on the approach to data collection, because, depending on the level of X-ray exposure in a particular data set, solvent-exposed side-chains, even those that cannot undergo decarboxylation, may be more or less visible. For example, the electron density of the side-chain of Tyr212 of GI (Fig. 3*c*) undergoes a correlated shift in a temperature-dependent manner into a new position. On $2mF_o - DF_c$ maps, that process would manifest itself as the appearance of two overlapping conformations with occupancies depending on the temperature of data collection, on dose and possibly on time as well. An instance of such behaviour for Tyr has already been documented at 100 K (Bury *et al.*, 2017). One approach to resolving these problems could be to use zero-dose extrapolation more broadly to determine whether the state of side-chains, including their occupancies, is affected by radiation-induced changes.

The ICA variant presented here will be useful for studies that rely on differences of electron density maps in real space. We expect that studies of dose-rate effects in the radiation damage field will be particularly suitable for such analyses. Ligand-identification experiments, where one attempts to identify ligand-binding pockets in fragment soaks or separate the densities associated with mixtures of ligands, is another example. Multi-temperature data studies, in which the change of temperature provides a probe to study conformational rearrangements or distributions of conformers, are yet another type of experiment where ICA will be useful.

Our results may also have a high impact on emerging methods of single-particle reconstruction in cryo-EM, in which dose-fractionation data collection is becoming standard (Campbell *et al.*, 2012; Cheng *et al.*, 2015). It is expected that radiation-induced changes in cryo-EM proceed through the same mechanism as in X-ray crystallography and in both types of experiments are driven by high-energy electrons that generate Auger electrons (Henderson, 1990, 1995), with their associated ionization causing chemical damage. In cryo-EM, high-energy electrons are delivered directly, while for X-ray experiments they are generated by the photoelectric effect, with the primary absorption events producing insignificant radiation damage. In cryo-EM, using high dose is one of the few paths to overcoming the basic limitations of particle alignment. We observed here the preservation of the core structure for thaumatin under radiation dose, which may indicate that using full dose from properly resolution-weighted averaging of the image stack should be done for initial structure solution and determination of the relative alignment of particles. Only then should one attempt to reweight the data more toward the initial exposures, to see if more structural details can be identified in a less damaged but noisier map. One can also create averaged maps associated with each time-point in the image stack, and apply ICA to such multiple reconstructions. These maps should be calculated with the same alignment as before, as the core should not be much affected, so systematic errors from using higher dose images should be low, while the increasing noise from using only initial images is highly detrimental to the accuracy of the alignment if such images will be used to re-determine the particles' orientation (Cheng *et al.*, 2015).

Acknowledgements

The authors are indebted to the 19ID and 19BM beamline staff: Randy Alkire, Marianne Cuff, Norma Duke, Steve Ginell, Youngchang Kim, Jerzy Osipiuk and Frank Rotella for their help in the data collection and analysis. We thank Zbigniew Dauter from the Center for Cancer Research, National Cancer Institute, NIH, for providing the data sets for thaumatin used in this analysis. Use of the Argonne National Laboratory Structural Biology Center beamlines at the Advanced Photon Source was supported by the US Department of Energy, Office of Biological and Environmental Research, under Contract No. DE-AC02-11357. We would also like to thank our anonymous referees.

Funding information

Funding for this research was provided by: the National Institutes of Health (grant Nos. R01GM053163, R01GM117080 and R01GM118619).

References

Alizadeh, E. & Sanche, L. (2012*a*). *Chem. Rev.* **112**, 5578–5602.
Alizadeh, E. & Sanche, L. (2012*b*). *Radiat. Prot. Dosimetry*, **151**, 591–599.
Alkire, R. W., Rotella, F. J., Duke, N. E. C., Otwinowski, Z. & Borek, D. (2016). *J. Appl. Cryst.* **49**, 415–425.
Asunción Vicente, M., Hoyer, P. O. & Hyvärinen, A. (2007). *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 896–900.
Aubert, C., Vos, M. H., Mathis, P., Eker, A. P. M. & Brettel, K. (2000). *Nature (London)*, **405**, 586–590.
Ball, P. (2008). *Chem. Rev.* **108**, 74–108.
Banumathi, S., Zwart, P. H., Ramagopal, U. A., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1085–1093.
Bednarek, J., Plonka, A., Hallbrucker, A. & Mayer, E. (1998*a*). *J. Phys. Chem. A*, **102**, 9091–9094.
Bednarek, J., Plonka, A., Hallbrucker, A. & Mayer, E. (1998*b*). *Radiat. Phys. Chem.* **53**, 635–638.
Bellissent-Funel, M. C., Hassanali, A., Havenith, M., Henchman, R., Pohl, P., Sterpone, F., van der Spoel, D., Xu, Y. & Garcia, A. E. (2016). *Chem. Rev.* **116**, 7673–7697.
Beratan, D. N., Onuchic, J. N., Winkler, J. R. & Gray, H. B. (1992). *Science*, **258**, 1740–1741.
Blake, C. C. F. & Phillips, D. C. (1962). *Biological Effects of Ionizing Radiation At the Molecular Level*, IAEA, Vienna, Austria, STI/PUB/60, pp. 183–191. International Atomic Energy Agency.
Borek, D., Cymborowski, M., Machius, M., Minor, W. & Otwinowski, Z. (2010). *Acta Cryst.* D**66**, 426–436.
Borek, D., Dauter, Z. & Otwinowski, Z. (2013). *J. Synchrotron Rad.* **20**, 37–48.
Borek, D., Ginell, S. L., Cymborowski, M., Minor, W. & Otwinowski, Z. (2007). *J. Synchrotron Rad.* **14**, 24–33.
Box, H. C. (1972). *Annu. Rev. Nucl. Sci.* **22**, 355–382.
Box, H. C., Budzinski, E. E. & Lilga, K. T. (1972). *J. Chem. Phys.* **57**, 4295–4298.
Box, H. C., Lilga, K. T., Budzinski, E. E. & Derr, R. (1969). *J. Chem. Phys.* **50**, 5422–5423.
Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.
Bury, C. S., Carmichael, I. & Garman, E. F. (2017). *J. Synchrotron Rad.* **24**, 7–18.
Bury, C. S., McGeehan, J. E., Antson, A. A., Carmichael, I., Gerstel, M., Shevtsov, M. B. & Garman, E. F. (2016). *Acta Cryst.* D**72**, 648–657.
Campbell, M. G., Cheng, A., Brilot, A. F., Moeller, A., Lyumkis, D., Veesler, D., Pan, J., Harrison, S. C., Potter, C. S., Carragher, B. & Grigorieff, N. (2012). *Structure*, **20**, 1823–1828.
Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. (2015). *Cell*, **161**, 438–449.
Cichocki, A., Douglas, S. C. & Amari, S. (1998). *Neurocomputing*, **22**, 113–129.
Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, J. S. (2007). *Nucleic Acids Res.* **35**, W375–W383.
Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2004). *Nucleic Acids Res.* **32**, W615–W619.
Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Acta Cryst.* D**59**, 903–909.
Emfietzoglou, D., Karava, K., Papamichael, G. & Moscovitch, M. (2003). *Phys. Med. Biol.* **48**, 2355–2371.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Ennifar, E., Carpentier, P., Ferrer, J.-L., Walter, P. & Dumas, P. (2002). *Acta Cryst.* D**58**, 1262–1268.
Evans, P. R. (2011). *Acta Cryst.* D**67**, 282–292.
Farver, O. & Pecht, I. (1997). *J. Biol. Inorg. Chem.* **2**, 387–392.
Fioravanti, E., Vellieux, F. M. D., Amara, P., Madern, D. & Weik, M. (2007). *J. Synchrotron Rad.* **14**, 84–91.
Florusse, L. J., Peters, C. J., Schoonman, J., Hester, K. C., Koh, C. A., Dec, S. F., Marsh, K. N. & Sloan, E. D. (2004). *Science*, **306**, 469–471.
French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.
Fütterer, K., Ravelli, R. B. G., White, S. A., Nicoll, A. J. & Allemann, R. K. (2008). *Acta Cryst.* D**64**, 264–272.
Garman, E. F. & Weik, M. (2013). *J. Synchrotron Rad.* **20**, 1–6.
Garrison, W. M. (1987). *Chem. Rev.* **87**, 381–398.
Gerstel, M., Deane, C. M. & Garman, E. F. (2015). *J. Synchrotron Rad.* **22**, 201–212.
Gonzalez, A. & Nave, C. (1994). *Acta Cryst.* D**50**, 874–877.
Gray, H. B. & Winkler, J. R. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 3534–3539.
Gupta, S., Feng, J., Chan, L. J. G., Petzold, C. J. & Ralston, C. Y. (2016). *J. Synchrotron Rad.* **23**, 1056–1069.
Hanson, B. L., Martin, A., Harp, J. M., Parrish, D. A., Bunick, C. G., Kirschbaum, K., Pinkerton, A. A. & Bunick, G. J. (1999). *J. Appl. Cryst.* **32**, 814–820.
Hawkins, C. L. & Davies, M. J. (2001). *BBA Bioenergetics*, **1504**, 196–219.
Henderson, R. (1990). *Proc. R. Soc. B*, **241**, 6–8.
Henderson, R. (1995). *Q. Rev. Biophys.* **28**, 171–193.
Hidaka, H., Koike, T., Kurihara, T. & Serpone, N. (2004). *New J. Chem.* **28**, 1100–1106.
Hyvärinen, A. (1999). *IEEE Trans. Neural Netw.* **10**, 626–634.
Jones, G. D. D., Lea, J. S., Symons, M. C. R. & Taiwo, F. A. (1987). *Nature (London)*, **330**, 772–773.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kmetko, J., Husseini, N. S., Naides, M., Kalinin, Y. & Thorne, R. E. (2006). *Acta Cryst.* D**62**, 1030–1038.
Learned-Miller, E. G. & Fisher, J. W. (2004). *J. Mach. Learn. Res.* **4**, 1271–1295.
Leif Hanson, B., Harp, J. M., Kirschbaum, K., Parrish, D. A., Timm, D. E., Howard, A., Alan Pinkerton, A. & Bunick, G. J. (2001). *J. Cryst. Growth*, **232**, 536–544.
Leiros, H.-K. S., Timmins, J., Ravelli, R. B. G. & McSweeney, S. M. (2006). *Acta Cryst.* D**62**, 125–132.
Liebschner, D., Dauter, M., Brzuszkiewicz, A. & Dauter, Z. (2013). *Acta Cryst.* D**69**, 1447–1462.
Macedo, S., Pechlaner, M., Schmid, W., Weik, M., Sato, K., Dennison, C. & Djinović-Carugo, K. (2009). *J. Synchrotron Rad.* **16**, 191–204.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–255.
Naday, I., Ross, S., Westbrook, E. M. & Zentai, G. (1998). *Opt. Eng.* **37**, 1235–1244.
Nave, C. (1995). *Radiat. Phys. Chem.* **45**, 483–490.
Nicoll, R. M., Hindle, S. A., MacKenzie, G., Hillier, I. H. & Burton, N. A. (2001). *Theor. Chim. Acta*, **106**, 105–112.
O'Neill, P., Stevens, D. L. & Garman, E. (2002). *J. Synchrotron Rad.* **9**, 329–332.
Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Acta Cryst.* A**59**, 228–234.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Otwinowski, Z., Minor, W., Borek, D. & Cymborowski, M. (2012). *International Tables for Crystallography Vol. F*, 2nd (revised) ed., pp. 282–295. John Wiley and Sons, Ltd.
Patten, F. & Gordy, W. (1960). *Proc. Natl Acad. Sci. USA*, **46**, 1137–1144.
Petrova, T., Ginell, S., Mitschler, A., Kim, Y., Lunin, V. Y., Joachimiak, G., Cousido-Siah, A., Hazemann, I., Podjarny, A., Lazarski, K. & Joachimiak, A. (2010). *Acta Cryst.* D**66**, 1075–1091.
Pimblott, S. M. & LaVerne, J. A. (2007). *Radiat. Phys. Chem.* **76**, 1244–1247.
Pizzitutti, F., Marchi, M., Sterpone, F. & Rossky, P. J. (2007). *J. Phys. Chem. B*, **111**, 7584–7590.
Rajagopal, S., Schmidt, M., Anderson, S., Ihee, H. & Moffat, K. (2004). *Acta Cryst.* D**60**, 860–871.
Ramagopal, U. A., Dauter, Z., Thirumuruhan, R., Fedorov, E. & Almo, S. C. (2005). *Acta Cryst.* D**61**, 1289–1298.
Rao, D. N. R., Symons, M. C. R. & Stephenson, J. M. (1983). *J. Chem. Soc. Perkin Trans. 2*, p. 727.
Rao, P. S. & Hayon, E. (1974). *J. Phys. Chem.* **78**, 1193–1196.
Ravelli, R. B. & Garman, E. F. (2006). *Curr. Opin. Struct. Biol.* **16**, 624–629.
Ravelli, R. B., Leiros, H. K., Pan, B., Caffrey, M. & McSweeney, S. (2003). *Structure*, **11**, 217–224.
Ren, Z., Chan, P. W. Y., Moffat, K., Pai, E. F., Royer, W. E., Šrajer, V. & Yang, X. (2013). *Acta Cryst.* D**69**, 946–959.
Schmidt, M., Rajagopal, S., Ren, Z. & Moffat, K. (2003). *Biophys. J.* **84**, 2112–2129.
Schrodinger, LLC (2015). *The pyMOL Molecular Graphics System*, Version 1.8.
Sevilla, M. D., D'Arcy, J. B. & Morehouse, K. M. (1979). *J. Phys. Chem.* **83**, 2893–2897.
Sterpone, F., Stirnemann, G., Hynes, J. T. & Laage, D. (2010). *J. Phys. Chem. B*, **114**, 2083–2089.
Sutton, K. A., Black, P. J., Mercer, K. R., Garman, E. F., Owen, R. L., Snell, E. H. & Bernhard, W. A. (2013). *Acta Cryst.* D**69**, 2381–2394.
Tezcan, F. A., Crane, B. R., Winkler, J. R. & Gray, H. B. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 5002–5006.
Warren, J. J., Ener, M. E., Vlček, A. Jr, Winkler, J. R. & Gray, H. B. (2012). *Coord. Chem. Rev.* **256**, 2478–2487.
Weik, M., Kryger, G., Schreurs, A. M. M., Bouma, B., Silman, I., Sussman, J. L., Gros, P. & Kroon, J. (2001). *Acta Cryst.* D**57**, 566–573.
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.
Weik, M., Ravelli, R. B., Silman, I., Sussman, J. L., Gros, P. & Kroon, J. (2001). *Protein Sci.* **10**, 1953–1961.
Westbrook, E. M. & Naday, I. (1997). *Methods Enzymol.* **276**, 244–268.
Wilson, K. & Yeates, D. (1979). *Acta Cryst.* A**35**, 146–157.
Winkler, J. R. & Gray, H. B. (2014). *J. Am. Chem. Soc.* **136**, 2930–2939.
Winkler, J. R. & Gray, H. B. (2015). *Q. Rev. Biophys.* **48**, 411–420.
Xu, G. Z. & Chance, M. R. (2004). *Anal. Chem.* **76**, 1213–1221.
Xu, G. Z. & Chance, M. R. (2007). *Chem. Rev.* **107**, 3514–3543.
Ziaja, B., London, R. A. & Hajdu, J. (2005). *J. Appl. Phys.* **97**, 064905.