JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

Received 14 November 2019
Accepted 27 February 2020

Edited by M. Yabashi, RIKEN SPring-8 Center, Japan

‡ These authors contributed equally to this work.

Keywords: time-resolved X-ray solution scattering; time-resolved X-ray liquidography; simulation; signal-to-noise ratio; solvent cage.

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

Estimating signal and noise of time-resolved X-ray solution scattering data at synchrotrons and XFELs

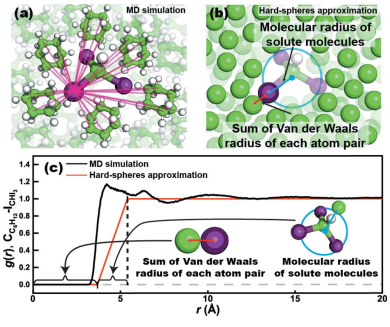
Jungmin Kim,[a,b]‡ Jong Goo Kim,[a,b]‡ Hosung Ki,[a,b]‡ Chi Woo Ahn[a,b] and Hyotcherl Ihee[a,b]*

[a]Department of Chemistry and KI for the BioCentury, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea, and [b]Center for Nanomaterials and Chemical Reactions, Institute for Basic Science (IBS), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea.
*Correspondence e-mail: hyotcherl.ihee@kaist.ac.kr

Elucidating the structural dynamics of small molecules and proteins in the liquid solution phase is essential to ensure a fundamental understanding of their reaction mechanisms. In this regard, time-resolved X-ray solution scattering (TRXSS), also known as time-resolved X-ray liquidography (TRXL), has been established as a powerful technique for obtaining the structural information of reaction intermediates and products in the liquid solution phase and is expected to be applied to a wider range of molecules in the future. A TRXL experiment is generally performed at the beamline of a synchrotron or an X-ray free-electron laser (XFEL) to provide intense and short X-ray pulses. Considering the limited opportunities to use these facilities, it is necessary to verify the plausibility of a target experiment prior to the actual experiment. For this purpose, a program has been developed, referred to as *S-cube*, which is short for a *Solution Scattering Simulator*. This code allows the routine estimation of the shape and signal-to-noise ratio (SNR) of TRXL data from known experimental parameters. Specifically, *S-cube* calculates the difference scattering curve and the associated quantum noise on the basis of the molecular structure of the target reactant and product, the target solvent, the energy of the pump laser pulse and the specifications of the beamline to be used. Employing a simplified form for the pair-distribution function required to calculate the solute–solvent cross term greatly increases the calculation speed as compared with a typical TRXL data analysis. Demonstrative applications of *S-cube* are presented, including the estimation of the expected TRXL data and SNR level for the future LCLS-II HE beamlines.

1. Introduction

Understanding the solution-phase reaction mechanism is central to the field of chemistry. Fast processes in solutions involving short-lived intermediates are often investigated by time-resolved spectroscopy, but the associated molecular structural changes in general are not direct observables of time-resolved spectroscopy, which relies on the transitions between energy states. In this regard, time-resolved X-ray solution scattering (TRXSS), also known as time-resolved X-ray liquidography (TRXL), has been established as a powerful tool for studying molecular structural dynamics in the liquid solution phase (Kjaer *et al.*, 2019; Haldrup *et al.*, 2012, 2019; Salassa *et al.*, 2010; Kim *et al.*, 2012; Ahn *et al.*, 2018; Biasin *et al.*, 2018; Canton *et al.*, 2015; Kim, Kim, *et al.*, 2016; Kong *et al.*, 2019; Leshchev *et al.*, 2018; Berntsson *et al.*, 2017; Josts *et al.*, 2018; Kim, Yang, *et al.*, 2016; Kim, Ganesan, *et al.*, 2018; Rimmerman *et al.*, 2017; Arnlund *et al.*, 2014; Kim, Muniyappan, *et al.*, 2016; Malmerberg *et al.*, 2015; Haldrup *et*



al., 2009; Christensen *et al.*, 2009; Cammarata *et al.*, 2006, 2008; Kong *et al.*, 2007; Ihee *et al.*, 2005; Plech *et al.*, 2004). In a typical experiment, an optical laser pulse is used to initiate a reaction and after a time delay an X-ray pulse is sent to probe the reaction progress via X-ray scattering. TRXL has been applied to a wide range of molecules ranging from small molecules such as organometallic compounds (Biasin *et al.*, 2018; Kim, Kim, *et al.*, 2016; Kong *et al.*, 2019; Leshchev *et al.*, 2018; Canton *et al.*, 2015; Haldrup *et al.*, 2012; Salassa *et al.*, 2010; Haldrup *et al.*, 2019; Kjaer *et al.*, 2019; Kong *et al.*, 2007) and hydrocarbons (Ahn *et al.*, 2018; Kim *et al.*, 2012; Ihee *et al.*, 2005; Davidsson *et al.*, 2005) to macromolecules such as proteins (Berntsson *et al.*, 2017; Josts *et al.*, 2018; Kim, Yang, *et al.*, 2016; Kim, Ganesan, *et al.*, 2018; Rimmerman *et al.*, 2017; Arnlund *et al.*, 2014; Kim, Muniyappan, *et al.*, 2016; Malmerberg *et al.*, 2011, 2015; Konuma *et al.*, 2011; Westenhoff *et al.*, 2010; Cho *et al.*, 2010; Andersson *et al.*, 2009; Cammarata *et al.*, 2008). Due to the successful applications of TRXL to a wide range of molecules and reactions, applications of TRXL are expected to increase in the future.

Nevertheless, TRXL is not an omnipotent technique, and it has two major limitations. The first is the weak sensitivity to the solute compared with time-resolved spectroscopy and the second is the limited number of beamlines for TRXL. Generally, the scattering signal from the solute is weaker than the spectroscopic signal in time-resolved spectroscopy because the total scattering signal is dominated by the solvent, which generally exists in greater amounts than the solute and usually does not participate in the net reaction (Neutze *et al.*, 2001; Plech *et al.*, 2004; Kim *et al.*, 2009; Haldrup *et al.*, 2010). Due to the relatively weak signal, it is necessary to use extremely intense X-ray pulses to collect data successfully. Accordingly, TRXL experiments are conducted at large-scale facilities that can produce such intense X-ray pulses, such as third-generation synchrotrons, which offer approximately 100 ps-long X-ray pulses, or at X-ray free-electron lasers (XFELs), which provide sub-100 fs-long X-ray pulses. Access to these facilities is generally competitive.

Due to these limitations, it is crucial to estimate the plausibility of a target TRXL experiment theoretically prior to performing the actual experiment. The X-ray scattering signal from a liquid solution sample is composed of four signals; the solute-only signal, the solute–solvent cross signal (cage signal), the solvent-only signal and noise, as illustrated in Fig. 1. Because the structural dynamics information of a reaction is mostly contained in the solute-only signal, the relative magnitude of the solute-only signal with respect to the total signal is the determining factor for the successful application of TRXL. In addition, if the expected signal level relative to the experimental noise level, that is, the signal-to-noise ratio (SNR), can be predicted prior to the actual beam time, it will allow an estimation of the accumulation time required to obtain a sufficient SNR suitable for the purpose of the experiment. Accordingly, it would greatly facilitate the planning of experiments and increase the success rate.

In this work, we present a program which can aid in the forecasting of the quality of experimental TRXL data. This program, named *S-cube* (S^3), which is short for a *Solution Scattering Simulator*, is a MATLAB-based graphical user interface (GUI) application. On the basis of the given experimental conditions, *e.g.* the intensity of the X-ray sources, the target solute and solvent, and the data accumulation time, *S-cube* calculates the solute-only signal, the solute–solvent cross signal, the solvent-only signal and the expected noise level for a desired target reaction for investigation. We show how *S-cube* can be used to estimate the plausibility of a TRXL experiment or to design a successful experiment. In particular, we demonstrate how *S-cube* can be used to simulate the signal level for a chemical reaction with a variety of experimental conditions. In addition, we use *S-cube* to show the expected signal level of a TRXL experiment at the upcoming LCLS-II HE beamlines (LCLS, 2018; LCLS-II, 2018).

2. Results and discussion

2.1. Simulation of the TRXL signal using *S-cube*

The TRXL signal consists of the solute-only term, solute–solvent cross (cage) term, solvent-only term and noise as shown in Figs. 1(*a*) and 1(*b*). Therefore, these four terms should be considered in the data simulation. Fig. 1(*c*) schematically illustrates how each of the four signals can be calculated. In the *S-cube* simulation, the simulated signal $[\Delta S_{\rm sim}(q)]$ is obtained by summing the solute-only signal $[\Delta S_{\rm solute}(q)]$, the solvent-only signal $[\Delta S_{\rm solvent}(q)]$, the solute–solvent cross signal $[\Delta S_{\rm cage}(q)]$ and the noise $[\Delta S_{\rm noise}(q)]$,

$$\Delta S_{\rm sim} = \Delta S_{\rm solute} + \Delta S_{\rm cage} + \Delta S_{\rm solvent} + \Delta S_{\rm noise}. \quad (1)$$

The key aspects of TRXL signal simulation using *S-cube* are as follows: (i) $\Delta S_{\rm cage}(q)$ is calculated much more quickly than in a typical program by treating molecules as hard spheres instead of relying on molecular dynamics (MD) simulations, which is the most time-consuming step, and (ii) the expected experimental noise level can be estimated in addition to theoretical signal components of the solute-only, solvent-only and solute–solvent cross signals. Thus, $\Delta S_{\rm cage}(q)$ and $\Delta S_{\rm noise}(q)$ are discussed in this section, and the details are described in the *Methods* section and elsewhere (Neutze *et al.*, 2001; Plech *et al.*, 2004; Kim *et al.*, 2009; Haldrup *et al.*, 2010).

$\Delta S_{\rm cage}(q)$ is calculated from the sine Fourier transform of pair-distribution functions (PDFs), $g(r)$, between atoms in solute and solvent in the following equation (Dohn *et al.*, 2015; Kim *et al.*, 2009),

$$\Delta S_{\rm cage}(q) = \sum_{i,j} f_i(q) f_j(q) \frac{N_i N_j}{V} 4\pi \times \left[\int_0^{R_{\rm box}} \Delta g_{i,j}(r) \frac{\sin(qr)}{qr} r^2 \, {\rm d}r \right], \quad (2)$$

where i and j are indices of atom types in solute and solvent molecules, respectively, f_i and f_j are atomic form factors for atom-types i and j, N_i and N_j are the numbers of atoms for atom-type i and j, and V is the box volume for MD simulations. In typical TRXL data analysis, PDFs are obtained from MD

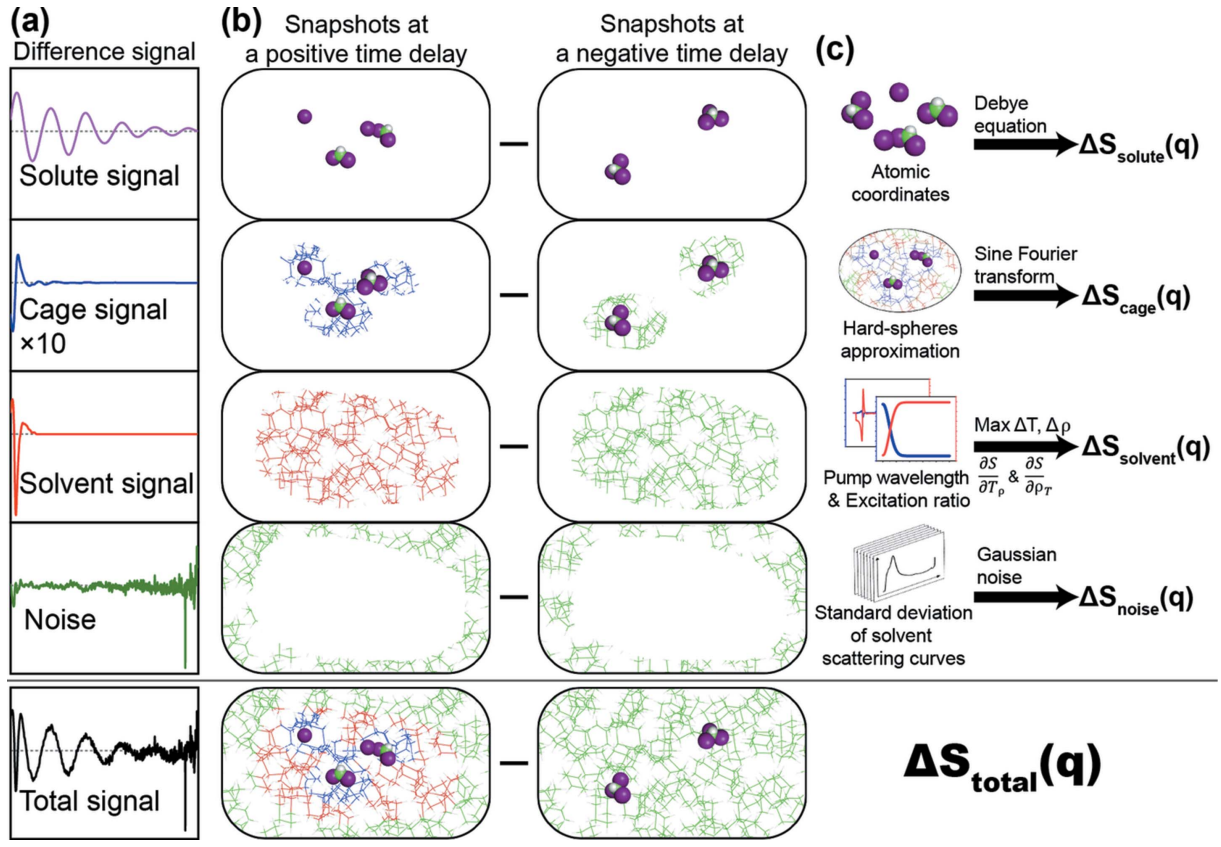


Figure 1
The four signals [solute (solute-only), cage (solute–solvent cross), solvent (solvent-only) and noise] that comprise the difference signal are schematically illustrated. (*a*) The difference curves corresponding to the four contributing signals and the total signal. (*b*) Snapshots of solute and solvent molecules for each signal. The snapshot for the solute signal represents the structure change of the solute molecules induced by a pump pulse. The snapshot for the cage signal represents the structure change of the solvent molecules surrounding the solute caused by the structural change of the solute molecules. The snapshot for the solvent signal represents the structural changes due to temperature and density changes of the solvent from heating caused by excited solute molecules. (*c*) Schematic diagram of the simulation of each signal in *S-cube*. The solute signal is calculated from the concentration and structure of the reactants and products using the Debye equation. For calculation of the solvent signal, maximum changes of temperature (ΔT) and density ($\Delta\rho$) are calculated from the energy of the pump laser and the number of light-absorbing solute molecules. Subsequently, the solvent signal is obtained from the sum of each product that are maximum $\Delta T \times (\delta S/\delta \mathrm{T})_\rho$ and $\Delta\rho \times (\delta S/\delta\rho)_T$. The noise is acquired by considering the simulation environments and $\sigma_{\mathrm{solvent}}$. The cage signal can be calculated using the sine Fourier transform from radial distribution functions of MD simulations or the hard-spheres approximation.

simulations [see Fig. 2(*a*)], which demand a large amount of computation. $g_{ij}(r)$ is the PDF which is calculated for every pair between types of atom *i* in solute and types of atom *j* in solvent. To substantially reduce the computation time, *S-cube* does not use MD simulations and instead adopts a simplified $g_{ij}(r)$ for atom pair *i* in solute and *j* in solvent using a hard-spheres approximation method [see Fig. 2(*b*)]. As shown in Fig. 2(*c*), the hard-spheres approximation method simplified the $g_{ij}(r)$ a trapezoidal function using the following equation (see Fig. 2),

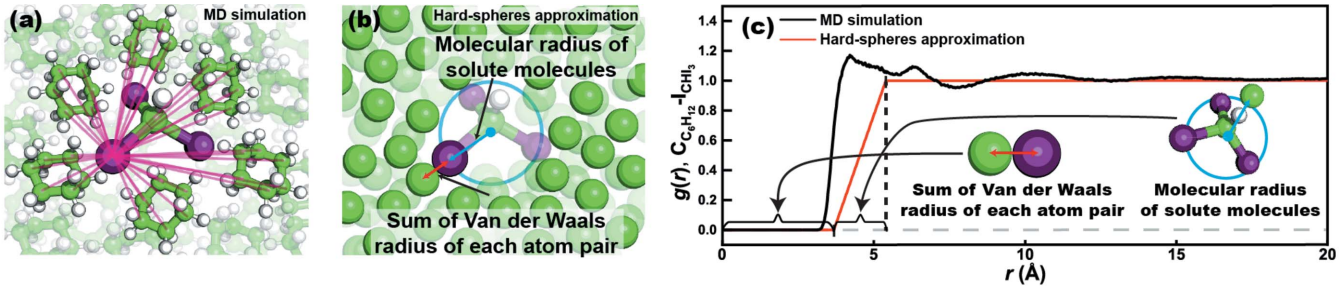


Figure 2
Comparison between the MD simulation and the hard-spheres approximation method for CHI_3 in cyclohexane. (*a*) Schematic diagram for the MD simulation. The transparent magenta lines represent interatomic distances between the iodine atom in CHI_3 and carbon atoms in cyclohexane. (*b*) Schematic diagram for the hard-spheres approximation. (*c*) Comparison of the pair distribution functions, $g(r)$, between iodine in CHI_3 and carbon in cyclohexane from an MD simulation and the hard-spheres approximation.

$$C_s = \sum_{l=1}^{k} X_l/k, \tag{3.1}$$

$$R_s = \sum_{l=1}^{k} ||X_l - C_s||^2/k, \tag{3.2}$$

$$g_{ij}(r) = \begin{cases} 0 & (r < V_{ij}), \\ (1/R_s)(r - V_{ij}) & (V_{ij} < r < V_{ij} + R_s), \\ 1 & (V_{ij} + R_s < r), \end{cases} \tag{3.3}$$

where V_{ij} is the sum of the van der Waals radii of the ith atom and jth atoms, and R_s is the molecular radius of the solute molecule, C_s is the centroid of the solute molecule, X_l is the position of the lth atom in the solute molecule, and k is the total number of atoms in the solute molecule. In equation (3.3), the simplified $g_{ij}(r)$ is calculated in the following three regimes: (1) when $r < V_{ij}$, $g_{ij}(r)$ is set to 0 due to the impenetrability of the hard sphere; (2) when $V_{ij} + R_s < r$, the interaction between the solute and solvent is approximated to be zero and thus $g_{ij}(r)$ is set to 1; (3) when $V_{ij} < r < V_{ij} + R_s$, $g_{ij}(r)$ increases linearly from 0 to 1 as a function of r. Figs. 3 and S5 compare $\Delta S_{cage}(q)$ calculated from MD simulations and the hard-spheres approximation method for various reactions. Although fine shapes cannot be reproduced by the hard-spheres approximation, the agreement with regard to the q range above 1 Å^{-1} in terms of the overall trend and amplitude appears sufficient for the purpose of estimating the contribution of the solute–solvent cross term relative to the other terms.

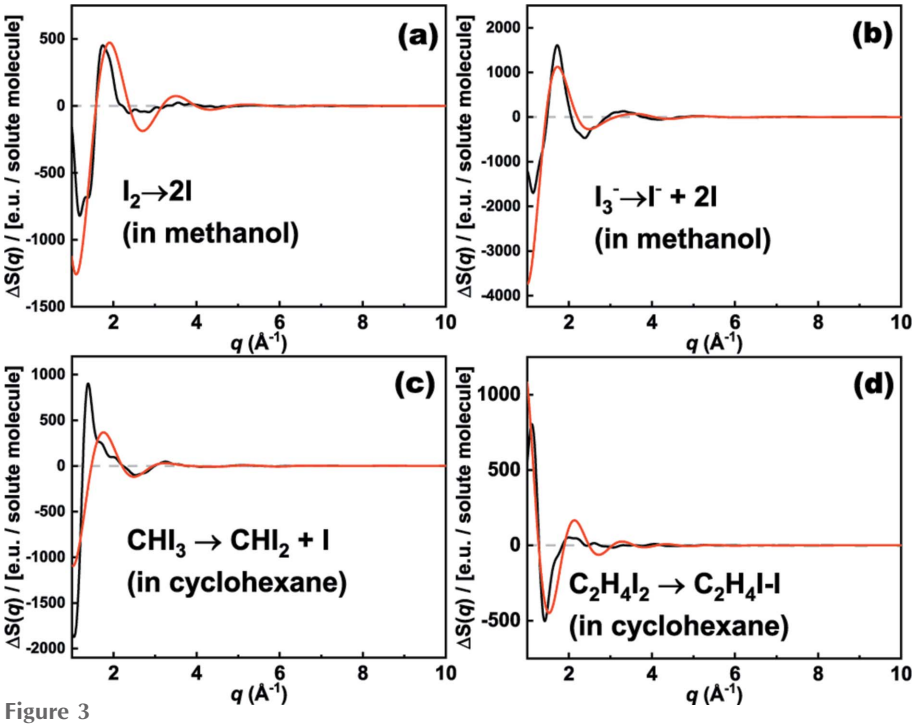


Figure 3
Comparison of solute–solvent cross terms obtained from MD simulations (black curves) and the hard-spheres approximation method (red curves) for various reactions. The difference scattering intensity, $\Delta S(q)$, is for one solute molecule and divided by the scattering intensity of a single electron. As a result, $\Delta S(q)$ is in electron units (e.u.) per solute molecule.

In a typical data analysis, $\Delta S_{noise}(q)$ is not calculated and only the other three signals are calculated to generate the theoretical curve to be compared with the experimental data. Because the purpose of *S-cube* is to predict the plausibility of a target experiment, the estimation of $\Delta S_{noise}(q)$ becomes important.

To simulate ΔS_{noise} of an experiment, as it is well known that the noise consists of several components having different physical origins, it is indeed ideal to take into account all the components of the noise. The noise components can be classified into two major categories: random noise and systematic noise. Random noise gives stochastic fluctuation of measured intensities around their true values. One of the representative components that consist of random noise is quantum noise. Quantum noise, which is also known as 'Poisson noise' or 'shot noise', originates from the quantum nature of scattered X-ray photons. Due to the Poisson nature of the noise, the amplitude of quantum noise is determined to be proportional to the square root of the scattering intensity, regardless of the experimental condition. In most cases, the quantum noise dominates the entire experimental noise, except for the two representative cases:

(1) When the scattering signal is too weak so that each detector pixel cannot receive a sufficient number of photons. In this case, readout noise, which emerges from the conversion process of the intensity of incident light to the electric signal, can be the most dominant source of the experimental noise as the amplitude of the quantum noise decreases due to the decrease of the intensity of scattering intensities.

(2) When the readout rate of the detector is too fast. It is known that readout noise of the detector abruptly increases with the increase of the readout rate of the signal when normalized to the data collection time. Accordingly, the readout noise can be the governing component of the entire noise with an experiment with a high readout rate.

Nevertheless, consideration of all of these noise components complicates the simulation algorithm unnecessarily, and thus can potentially confuse the users of the simulation code. Also, in a typical TRXL experiment, the incident photon flux is high and as a result the Poisson noise due to the scattered photons impinging on the detector is larger than those from dark noise and readout noise from the detector by orders of magnitude. For this reason, the *S-cube* simulation only considers the contribution of the Poisson noise when estimating the amplitude of the noise in the experimental data. Therefore, for the simulations in *S-cube*, we allow the user to choose between two different methods of noise estimation. As the first method,

which is for a more accurate estimation of the noise, one can collect solvent scattering data under similar experimental conditions to the experiment to be simulated in advance so that *S-cube* can use the obtained solvent data as the reference. Use of the experimental data measured under similar experimental conditions as a reference for the simulation indeed guarantees a precise estimation of the amplitude of the experimental noise as the contributions of all the possible components of the noise are reflected in the real data. Nevertheless, considering the limited accessibility to the X-ray sources for a TRXL experiment, this method is impractical for general users preparing an experiment. Accordingly, we also provide a second, the simpler, but more general method as an option for *S-cube* simulation. In this alternative method, the noise level is estimated based on reference solvent scattering data which are provided by default in *S-cube*. For this estimation, it is assumed that quantum noise, which dominates $\Delta S_{\rm noise}$ for a typical experimental condition, is considered as the only source of noise and other sources of noise are ignored for the purpose of simplicity, but rough estimation of the quality of the experimental data. This consideration allows to quickly estimate the SNR of an experiment by only considering the intensity of incident X-rays, as proven in the supporting information. Nevertheless, at the current stage, the number of solvent scattering data provided by *S-cube* is not yet enough to cover all the general experimental conditions, and therefore there is room for update. We will continue to add the solvent data for as many different conditions as possible in order to further improve the reliability of the *S-cube* simulation. All the simulations demonstrated in this work were performed using this second method. A more detailed theoretical background for the second method is as follows.

As discussed in the *Methods* section, the total standard deviation of the scattering intensity, $\sigma_{\rm solution}(q)$, can be approximated to be equal to the standard deviation of the solvent, $\sigma_{\rm solvent}(q)$, because the number of solvent molecules is high enough to neglect the other two contributions from the solute molecules and cages. In other words, regardless of the type of solute, $\sigma_{\rm solution}(q)$ can be approximated by $\sigma_{\rm solvent}(q)$ which can be measured from a separate experiment on a neat solvent.

Assuming that each scattering photon is independent and that the variance of the scattering intensity is proportional to the scattering intensity, the variance of the scattering intensity is proportional to the number of incident scattering photons to the sample per scattering curve (N). Thus, $\sigma_{\rm solvent}(q)$ can be used to estimate the standard deviation of scattering intensity for a simulated condition, $\sigma_{\rm solution}(q)$, using $N_{\rm solvent}$ and $N_{\rm simulation}$ calculated from the experimental parameters of the X-ray pulse repetition rate (f), detector exposure time (D) and the number of photons per pulse (n) using the following equation,

$$\begin{aligned} \sigma_{\rm simulation}(q) &= \sigma_{\rm solvent}(q)\left(N_{\rm simulation}/N_{\rm target}\right)^{1/2}, \\ N_{\rm simulation} &= f_{\rm simulation}\, D_{\rm simulation}\, n_{\rm simulation}, \\ N_{\rm target} &= f_{\rm target}\, D_{\rm target}\, n_{\rm target}. \end{aligned} \tag{4}$$

S-cube receives the required experimental parameters including $\sigma_{\rm solvent}$ (including $f_{\rm solvent}$, $D_{\rm solvent}$ and $n_{\rm solvent}$), $f_{\rm target}$, $D_{\rm target}$ and $n_{\rm target}$ as input through the graphical user interface depicted in Fig. S1. Some examples of $\sigma_{\rm solvent}(q)$ for experiments conducted at various X-ray facilities are shown in Fig. S2. Meanwhile, scattering photon counting statistics can be assumed to be a Poisson process (Kirian *et al.*, 2011; Schindler *et al.*, 2016; Sedlak *et al.*, 2017), which is popularly chosen to describe the noise of the scattering intensity with a specific standard deviation using Gaussian (Bernadó *et al.*, 2007; Förster *et al.*, 2008; Pinfield & Scott, 2014; Schindler *et al.*, 2016; Stovgaard *et al.*, 2010) or Poisson noise (Schneidman-Duhovny *et al.*, 2010; Sedlak *et al.*, 2017).

$\Delta S_{\rm noise}$ is calculated by considering a normal probability distribution which has $\sigma_{\rm target}$ as its standard deviation as follows,

$$\begin{aligned} I &= td/2D, \\ \Delta S_{\rm noise} &= (1/I)\textstyle\sum_{k=1}^{I} r_k[0, \sigma_{\rm target}(q)]. \end{aligned} \tag{5}$$

Here, I is the number of difference curves for the simulation, t is the nominal accumulation time (in seconds), d is the duty cycle which is the fraction of the actual accumulation time to the nominal accumulating time (t), and $r_k[0, \sigma_{\rm target}]$ is the kth randomly generated Gaussian noise. The Gaussian noise was generated by sampling random numbers with a normal probability distribution which has zero as the mean value and has $\sigma_{\rm target}$ as its standard deviation.

2.2. Comparison with experimental data

First, we checked the validity of the *S-cube* simulation by comparing simulated difference scattering curves from *S-cube* for a model reaction and actual experimental data from synchrotrons and XFELs. For the model reaction, the photolysis of iodoform (CHI_3) in cyclohexane, where the experimental data corresponding to the reaction were previously analyzed and reported by our group (Ahn *et al.*, 2018), was selected. According to the study, there are two parallel reaction channels that contribute to the reaction intermediates at a time delay of 100 ps, which are the isomer channel (from the excited CHI_3 to the CHI_2I isomer) and the radical channel (from the excited CHI_3 to the CHI_2 and I radicals) with the molar fraction of the two intermediates being 40:60 at the time delay [as shown in the inset of Fig. 4(*e*)].

For comparison, we simulated the averaged difference scattering curves corresponding to the model reaction while varying the number of difference scattering curves for averaging, using the same experimental parameters used in the experiment and with $\sigma_{\rm solvent}$ as obtained from an earlier study (Ahn *et al.*, 2018). The difference curves contain all relevant contributions including the heating signal (the solvent-only term) although the heating signal for this case with cyclohexane as solvent is negligible in the displayed q range. Fig. 4 compares the simulated and experimental difference scattering curves. The experimental data, obtained at a synchrotron, and the simulated data are generally consistent regardless of the number of difference curves. For a proper

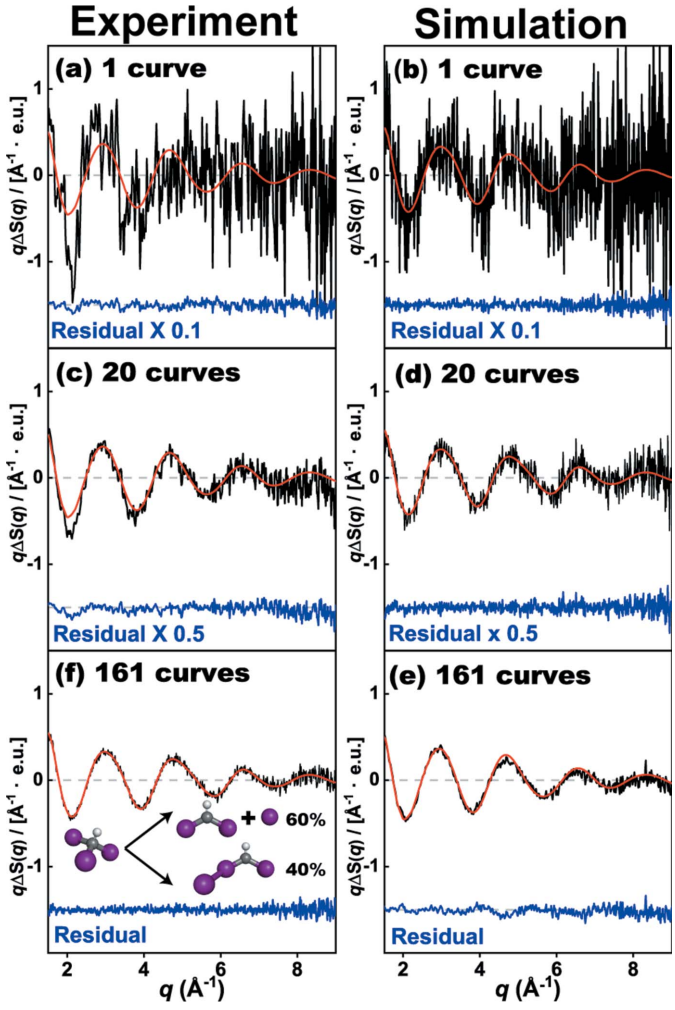


Figure 4
Comparison of experimental (*a*, *c*, *e*) and simulated data (*b*, *d*, *f*) for the difference curve at the time delay of 100 ps for iodoform dissolved in cyclohexane collected at the ID09 beamline of ESRF. The experimental data and simulation results are averaged from one curve (*a*, *b*), 20 curves (*c*, *d*) and 161 curves (*e*, *f*), respectively. In experimental data (*a*, *c*, *e*), the black lines are experimental curves. The red lines are theoretical curves, which are calculated $q\Delta S(q)$ from the results reported by Ahn *et al.* (2018). The blue lines are residual between experimental and theoretical curves. In simulation results (*b*, *d*, *f*), the black, red and blue curves are simulation results with noise, simulation results without noise, and the residuals between simulation results with and without noise, respectively. Note that the residuals in (*a*) and (*b*) are multiplied by 0.1 and those in (*c*) and (*d*) by 0.5. The residuals in experiment and simulation represent noise. The *y*-axis indicates $q\Delta S(q)$, the difference scattering intensity multiplied by *q*. The difference scattering intensity, $\Delta S(q)$, is scaled by the number of solvent molecules and has the unit of electron units (e.u.).

comparison, the experimental data are scaled to the simulated data using a scaling factor which is determined by following the scheme depicted in Figs. S3 and S4. The level of the residual (blue line), which is the deviation of the measured (or simulated) data from the corresponding expected values and which thereby represents the experimental noise of the experimental data, is quite well reproduced by the simulated data. The overall consistency of the residuals from the experiment and the simulation confirms that *S-cube* can satisfactorily simulate the degree of experimental noise.

To verify the applicability of *S-cube* to experiments at XFELs, we measured and used the experimental data corresponding to the same model reaction at PAL-XFEL (Kang *et al.*, 2017) for comparison with the *S-cube* data. $\sigma_{\rm solvent}(q)$ was also measured in the same experiment, and it was used for the *S-cube* simulation in the experiment. The resulting experimental and simulated data show excellent agreement, as shown in Fig. 5. In Fig. 5(*a*), the average of 460 experimental difference scattering curves is shown together with the corresponding residual representing the noise level of the averaged experimental curve. More detailed experimental and simulated parameters are as follows. *f*: 1 kHz; *D*: 1.5 s; *n*: 5×10^8 for Figs. 5(*a*)–5(*f*). Measurement times: 0.0008 h for Fig. 4(*b*), 0.1288 h for Fig. 5(*d*). The same *q* bins are used for both experimental and simulated data. For comparison, the simulated difference scattering curve and related noise level are depicted in Fig. 5(*b*). The amplitude of the simulated noise

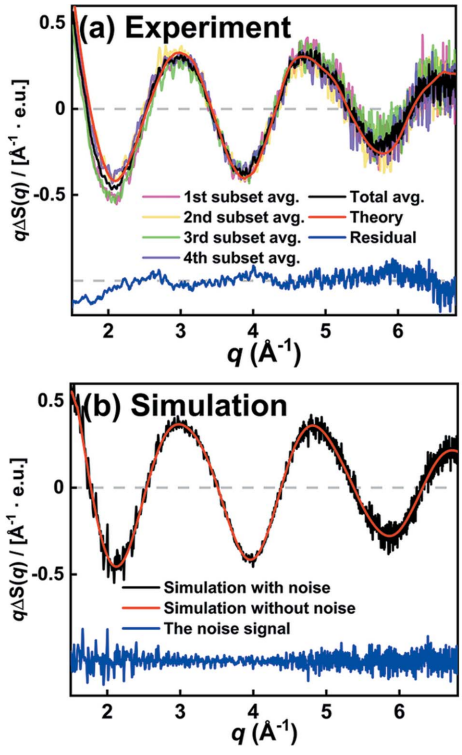


Figure 5
Comparison of (*a*) experimental TRXL data and (*b*) *S-cube* simulation for iodoform in cyclohexane at a time delay of 100 ps after excitation at 267 nm, which was measured at the XSS beamline of PAL-XFEL. In (*a*), the black curve is the averaged difference scattering curve from 460 difference scattering curves, the red curve is the smoothed difference scattering curve from the averaged difference scattering curve, the blue curve is the residual between the two curves, and the transparent pink, yellow, green and purple curves are the averaged curves from subsets of 460 difference scattering curves, each of which consists of 115 difference scattering curves. In (*b*), the black, red and blue curves are simulation results with noise, simulation results without noise, and the residuals between simulation results with and without noise, respectively. The residuals in (*a*) and (*b*) represent noise. The *y*-axis indicates $q\Delta S(q)$, the difference scattering intensity multiplied by *q*. The difference scattering intensity, $\Delta S(q)$, is scaled by the number of solvent molecules and has the unit of electron units (e.u.).

deviates from the experimental values at around $q = 2.0$ Å^{-1}. To verify the origin of the discrepancy, we calculated the average curves of four different subsets of the experimental data, each of which consists of 115 experimental difference scattering curves [partial averages, shown in Fig. 5(*a*)]. Although the partial averages show reasonable agreement overall, a noticeable inconsistency between the partial averages can be observed at around $q = 2.0$ Å^{-1}. The deviation of each partial average from the overall average shows a tendency such that the deviation is either all positive or all negative at $q = 2.0$ Å^{-1} and adjacent q-points. This tendency of the fluctuation of difference scattering intensities in q-space is in clear contrast to the expected behavior of the quantum noise, which should randomly fluctuate between positive and negative around zero. Such correlations in the fluctuation of the signal in the adjacent q-points indicate that there is another noise source which contributes to the regular fluctuation of the signal, other than the quantum noise from the detector. We attribute the other source of the signal fluctuation to systematic noise arising from fluctuations in the experimental conditions, such as the thickness of the liquid jet or the intensity of the X-ray pulse (Haldrup, 2014; Ki *et al.*, 2019; van Driel *et al.*, 2015).

As manifested by the fluctuations of the partial averages, the difference scattering curves measured from the experiment inevitably are contaminated by systematic noise unless the experiment is performed in an ideally constant condition.

Accordingly, $\sigma_{\rm solvent}(q)$ contains quantum noise and systematic noise from the fluctuations of the experimental conditions. As a result, if $\sigma_{\rm solvent}(q)$ is used to represent the quantum noise alone, it would overestimate the quantum noise. We do not consider this overestimation to be a major problem because it would set the upper limit for the worst case, reflecting the potential systematic noise. Moreover, Fig. 5 shows that the noise in the simulation and experiment converges as the number of averaged curves increases.

When conducting a TRXL experiment, the diffraction signal from the sample can be either separately collected for each X-ray pulse in a shot-by-shot manner or accumulated for multiple X-ray pulses in an integration mode. The difference of the noise level depending on the two experimental modes can be inferred from the comparison between the standard deviation of experimental data measured at ESRF and PAL-XFEL (blue and magenta curves in Fig. S2, respectively), obtained by the integration and shot-by-shot schemes, respectively. Compared with the images measured in the integration mode, the images measured in the shot-by-shot manner in general display higher fluctuation of scattering intensities. There are two main reasons for this. One is that when the signal from multiple X-ray pulses are accumulated in an image, the effect of the fluctuation of the experimental conditions such as the X-ray pulse intensities and the thickness of the liquid jet are averaged over the period of accumulation. The other reason is that the overall electronic readout noise is higher for the shot-by-shot mode when normalized by the data collection time because the electronic readout noise per image is about the same.

2.3. Simulation of time-resolved curves

A typical TRXL experiment yields time-resolved data measured at a number of time delays. The concentrations of reaction intermediates change over time delays, and a kinetic analysis of time-resolved data can extract such information and species-associated difference curves, which are often subject to further structural analyses. In this regard, it is desirable to simulate time-resolved difference curves as a function of the time delays. One of the most important parameters to be determined for a TRXL experiment is the number of time delays to cover the desired time range for a given measurement time, as the number of time delays during a limited measurement time exists in a trade-off relationship with the quality of the data, that is, the SNR of the data at each time delay in the q-space. As both a sufficient number of time delays and the quality of the data are prerequisites for the retrieval of reliable kinetics and structural changes, determining the optimal number of time delay points for a given measurement time plays an important role in a successful experiment.

For such a purpose, *S-cube* allows the user to estimate the quality of the experimental data from a given set of experimental parameters such as the duration of the beam time, the duty cycle and the number of time delays, and thereby facilitates the determination of the optimal number of time delays to be measured. As a demonstrative example, we simulated the TRXL data using *S-cube* for the photodissociation reaction of CHI_3 for different numbers of time delays. The reaction scheme and the time-dependent concentrations of intermediates, which were reported from the previous TRXL study on the reaction, are shown in Fig. 6. Based on the time-dependent concentration profile, we simulated the experimental curves for two different number of time delays. For the simulation parameters, 24 h of beam time and 70% of duty cycle were assumed. In addition, it was assumed that the measurement time is the same for each time delay. The other parameters such as $\sigma_{\rm solvent}(q)$ and the intensity of the X-ray pulses were set to be the same as in the simulations shown in Fig. 4. Fig. 6(*c*) shows the quality of the experimental data for four selected time delays, −3 ns, 100 ps, 30 ns and 1 µs, when the data are measured for a hundred time delays. It can be seen that the quality of the experimental data is sufficient enough to resolve the small changes in the shape of the signal. By contrast, in the case when the data are measured for 4000 time delays, the SNR of the data becomes much worse so that it is difficult to discern the differences between the experimental data for different time delays [see Fig. 6(*d*)]. The noise level of the experimental data for two different numbers of time delays and the difference between the data for different time delays are shown in Fig. 6(*e*) for comparison. For 100 time delays, as the noise level of the data is much smaller than the difference between the experimental data, the change of the shape of the signal can be clearly resolved which eventually allows the change of the molecular structures during the reaction to be retrieved. However, when the data are measured at 4000 time delays for the same given measurement

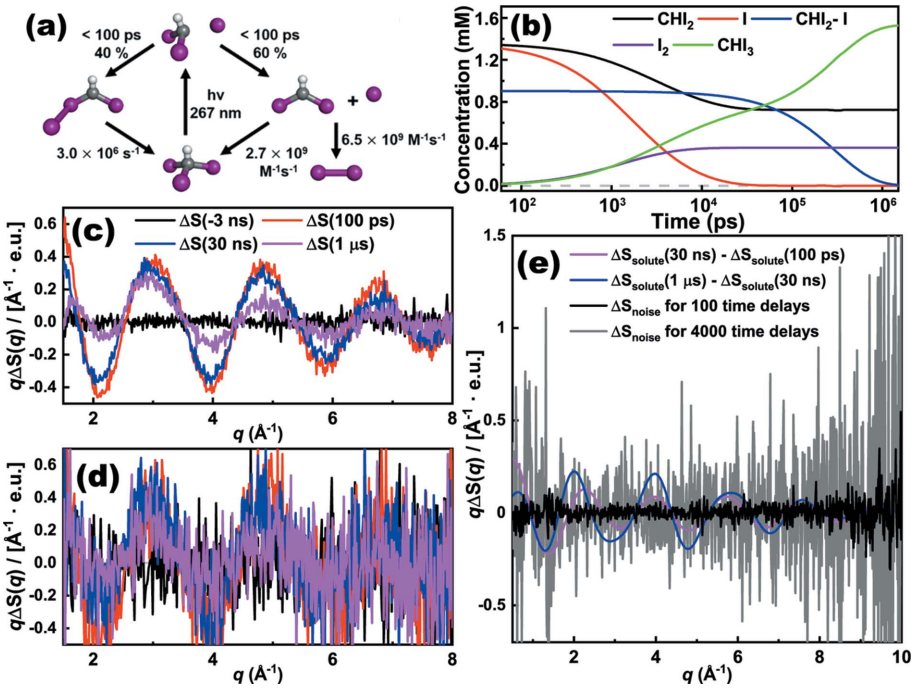


Figure 6
TRXL data simulation by *S-cube* for the photolysis reaction of CHI_3. (*a*) Kinetics scheme used for the simulation. (*b*) Time-dependent concentrations (solid lines) according to the kinetics. (*c*, *d*) Time-resolved difference curves simulated assuming that the data are collected for 24 h of beam time at a synchrotron (ESRF) with 70% of duty cycle. The curves at four selected time delays, −3 ns, 100 ps, 30 ns and 1 µs, are shown among the entire series consisting of (*c*) 100 or (*d*) 4000 time delays. (*e*) Comparison of the noise level of the experimental data for (*c*) and (*d*) and the difference between the data at different time delays without any consideration of experimental noise. The *y*-axis for (*c*, *d*, *e*) indicates $q\Delta S(q)$, the difference scattering intensity multiplied by q. The difference scattering intensity, $\Delta S(q)$, is scaled by the number of solvent molecules and has the unit of electron units (e.u.).

time, it is expected that such a change of the molecular structures would not be retrieved due to the low SNR of the experimental data. As shown through this demonstrative example, *S-cube* can be used to simulate TRXL data for a series of time delays with varying the number of time delays. Apparently, by inspecting the quality of the simulated TRXL data, one can determine the optimal number of time delays to achieve the goal of the experiment. Thus *S-cube* can be a useful tool for conducting a successful experiment within a given beam time.

2.4. Effect of heavy-atom labeling

Because *S-cube* is a program that simulates TRXL data, we demonstrate the application of *S-cube* by simulating TRXL data for a series of target molecules that have photoinduced structural changes. Here, we selected the *cis-trans* photoisomerization of azobenzene (AB) as a model of photoinduced structural change and simulated TRXL data corresponding to the structural change using *S-cube*. This reaction is one of the most representative photoisomerization reactions, and it has been studied for decades with various spectroscopic techniques (Bortolus & Monti, 1979; Schultz *et al.*, 2003; Stuart *et al.*, 2007). Despite the fact that *cis-trans* isomerization of AB has received much attention (Bandara & Burdette, 2012; Schultz *et al.*, 2003), it has been a challenge to examine the structural change during the reaction by using TRXL. The major obstacle during an investigation using TRXL is that AB consists of only light atoms, *i.e.* without any heavy atoms. As the scattering intensity from a molecule is proportional to the square of the number of electrons in the molecule, the weak scattering signal from AB makes it difficult to obtain the signal directly associated with the structural change of the molecule with a sufficient SNR. For the same reason, only a highly limited number of TRXL studies have been reported on molecules composed only of light atoms (Kim *et al.*, 2009; Leshchev *et al.*, 2018). Nevertheless, as an archetypal means of overcoming the low scattering cross section of such molecules, heavy-atom labeling using heavily scattering labels such as bromine with the molecules to enhance the scattering cross section has been proposed (Ihee, 2009; Ihee *et al.*, 2010; Ki *et al.*, 2017; Mathew-Fenn *et al.*, 2008). Using *S-cube*, we quantitatively evaluated how the heavy-atom labeling enhances the TRXL signal for the photoisomerization of AB. For this purpose, we also examined the TRXL signal corresponding to the photoisomerization of a di-bromo derivative of AB, 4,4′-dibromo azobenzene (Br_2AB).

As shown in Fig. 7(*a*), using *S-cube*, we simulated the TRXL signal arising from the photoisomerization reaction of 50 m*M* AB in cyclohexane with a 10% excitation ratio. In Fig. 7(*b*), we also simulated the TRXL signal for the photoisomerization reaction of Br_2AB under the same experimental conditions. To examine the effect of the heavy-atom labeling on the TRXL signal in terms of the noise level, we ignored structural differences other than the existence of heavy atoms between AB and Br_2AB by simply replacing two H atoms with two Br atoms in the structure of AB [see Fig. 7(*b*)]. For the simulations shown in Figs. 7(*a*) and 7(*b*), $\sigma_{solvent}(q)$ at the XSS beamline of PAL-XFEL was used.

Fig. 7(*a*) shows the TRXL signal simulation results of the photoisomerization of AB for PAL-XFEL. The solute signal is much lower than the noise estimated from $\sigma_{solvent}(q)$ in PAL-XFEL, indicating that the experiment would yield data sufficient for elucidating the molecular structural change occurring during the reaction. On the other hand, Fig. 7(*b*), showing the simulation result of the photoisomerization of the heavy atom substituent Br_2AB, indicates that the solute signal is dominant compared with the noise. Therefore, it can be seen from the comparison of Figs. 7(*a*) and 7(*b*) that the heavy-atom labeling

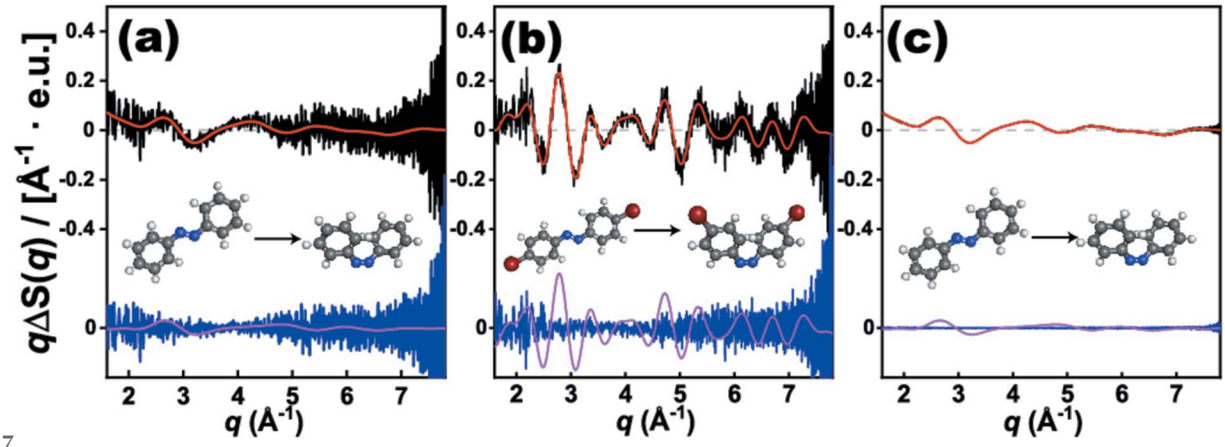


Figure 7
Comparison of *S-cube* simulation of photoisomerization of azobenzene (AB) at PAL-XFEL and LCLS-II HE and 4,4′-dibromo azobenzene (Br_2AB) at PAL-XFEL. The black, red and blue curves are simulation results with noise, simulation results without noise, and the residuals between simulation results with and without noise, respectively, and the magenta lines are solute-only signals. (*a*) *S-cube* simulation of photoisomerization of AB in cyclohexane at PAL-XFEL with accumulation time of 60 s and repetition rate (f) of 30 Hz, which corresponds to 540 difference scattering curves. (*b*) *S-cube* simulation of photoisomerization of Br_2AB in cyclohexane at PAL-XFEL with accumulation time of 60 s and f of 30 Hz, which corresponds to 540 difference scattering curves. (*c*) *S-cube* simulation of photoisomerization of AB in cyclohexane at LCLS-II HE with accumulation time of 60 s, f of 1 MHz, which corresponds to 18 000 000 difference scattering curves. $\sigma_{solvent}(q)$ at LCLS-II HE is scaled from $\sigma_{solvent}(q)$ at PAL-XFEL with each value of photons per curve which are 3×10^{10} for LCLS-II HE and 1×10^{12} for PAL-XFEL. The y-axis indicates $q\Delta S(q)$, the difference scattering intensity multiplied by q. The difference scattering intensity, $\Delta S(q)$, is scaled by the number of solvent molecules and has the unit of electron units (e.u.).

clearly offers great potential to observe structural changes at PAL-XFEL.

2.5. The prospect of LCLS-II HE

Although heavy-atom labeling of molecules can be an effective means by which to overcome the low scattering signals arising from molecules consisting of only light atoms (Andersson *et al.*, 2008; Malmerberg *et al.*, 2011), one cannot completely rule out the possibility of unexpected structural distortion due to the attached heavy atoms. Therefore, it remains desirable to obtain the signal from an intact molecule without heavy-atom labeling. In this regard, we considered the possibility of obtaining TRXL data without heavy-atom labeling in the future XFELs as the performance of state-of-the-art XFEL is evolving remarkably. One typical example showing the evolution of the performance of XFEL is LCLS-II HE (Schoenlein *et al.*, 2017), which is expected to have an f value of 1 MHz, representing a substantial improvement by a factor of tens of thousands compared with PAL-XFEL with a minute loss of n by a factor of dozens (Kim, Kim *et al.*, 2018). In terms of N, there would be an improvement by a factor of several thousand, which would eventually lead to a significant improvement of the SNR of the experimental signal. Accordingly, it is expected that the evolution of the performance of the XFELs would open up new experimental possibilities, such as the ability to capture the small signals that could not be resolved at currently available XFELs. Hence, we suggest that *S-cube* can be utilized to quantitatively predict the quality of future experimental data which can be obtained using these XFELs at higher performance levels. To demonstrate this aspect, we estimated the experimental signal corresponding to the photoisomerization of AB using *S-cube* for experiments at one of the representative future XFELs, LCLS-II HE. The following two assumptions applied when estimating $\Delta S_{noise}(q)$ from the experiment at LCLS-II HE. The first is that $\sigma_{solvent}(q)$ is proportional to the square root of N. By using this assumption, we estimated $\sigma_{target}(q)$ at LCLS-II HE from $\sigma_{solvent}(q)$ at PAL-XFEL using the following formula, which is closely related to equation (4),

$$\sigma_{\mathrm{LCLS\,II\,HE}}(q) = \sigma_{\mathrm{PAL\,XFEL}}(q)\left(\frac{N_{\mathrm{LCLS\,II\,HE}}}{N_{\mathrm{PAL\,XFEL}}}\right)^{1/2} = \sigma_{\mathrm{PAL\,XFEL}}(q)\left(\frac{f_{\mathrm{LCLS\,II\,HE}}\, D_{\mathrm{LCLS\,II\,HE}}\, n_{\mathrm{LCLS\,II\,HE}}}{f_{\mathrm{PAL\,XFEL}}\, D_{\mathrm{PAL\,XFEL}}\, n_{\mathrm{PAL\,XFEL}}}\right)^{1/2}. \quad (6)$$

The second assumption is that $\sigma_{target}(q)$ stems solely from the quantum noise. However, it should be noted that there are other sources of noise as well, *e.g.* systematic noise due to fluctuations of the experimental conditions, that contribute to $\sigma_{target}(q)$. As demonstrated in Fig. 5, the presence of such systematic noise makes it difficult for the second assumption to be strictly valid when estimating actual experimental data. Nevertheless, as discussed in Fig. 5, even an estimation under a flawed assumption still allows a reasonable prediction of the feasibility of certain ground-breaking experiments at new X-ray source facilities, as the noise level in *S-cube* is purposely overestimated under this assumption and thus can be regarded as able to provide an upper limit in terms of the noise level.

For the simulation shown in Fig. 7(*c*), the $\sigma_{solvent}$ used for the simulations, of which the results are shown in Fig. 7, indeed cannot be directly measured from LCLS-II HE, but estimated from experimental data measured at PAL-XFEL. For the estimation, $\sigma_{solvent}$ from PAL-XFEL was simply scaled on the basis of the different number of incident photons. The n values at PAL-XFEL and LCLS-II HE were set to 1×10^{12} and 3×10^{10}, respectively, and the corresponding f values were considered to be 30 Hz and 1 MHz (Kim, Kim *et al.*, 2018). An

accumulation time of 36 s was used in both simulations for the PAL-XFEL and LCLS-II HE cases, after which the overall 540 and 1.8×10^7 difference scattering curves were averaged to yield the resulting simulated difference scattering curves. Fig. 7(*c*) shows the result of the simulation of the experiment at LCLS-II HE. When the result is compared with that from PAL-XFEL, despite the fact that the measurement times for both simulated data are identical, there is a considerable difference in the SNR of the data. The curve obtained from LCLS-II HE shows a much clearer signal than that from PAL-XFEL due to the large value of *f* of LCLS-II HE (1 MHz). The quantitative SNR estimation by *S-cube* indicates that the improved performance of XFEL will allow the resolution of structural changes of molecules with light atoms, such as the photoisomerization of AB.

3. Conclusions

In this work, we introduce *S-cube*, which can simulate the solute-only signal, the solute–solvent cross signal, the solvent-only signal and the noise for a target reaction, a desired solvent, a target beamline and a given data collection time. The purpose of *S-cube* is to allow the user to calculate the expected noise level routinely compared with the scattering signal, the relative magnitude of the solute-only signal against the solvent-only signal and the simplified cage signal using the hard-spheres approximation. We expect that *S-cube* well help in the design of successful TRXL experiments at both synchrotrons and XFEL beamlines.

4. Distribution

S-cube is an app that can be used in MATLAB for free. The *S-cube* program (doi:10.5281/zenodo.3637919) is distributed in https://zenodo.org/badge/latestdoi/207274974 as well as the GitHub repository (https://github.com/Jkim9486/Scube).

5. Methods

5.1. Detailed procedure of the TRXL data simulation using *S-cube* and its theoretical background

For randomly oriented molecules, the X-ray scattering intensity from chemical species (reactants, intermediates and products) can be calculated by the Debye equation using the molecular structures of chemical species,

$$S_k(q) = \sum_n f_n^2(q) + \sum_n \sum_{m \neq n} f_n(q) f_m(q) \frac{\sin(q r_{nm})}{q r_{nm}}, \quad (7)$$

where q is the momentum transfer vector between the incident and elastically scattered X-ray waves, $S_k(q)$ is the X-ray scattering intensity for the chemical species k, the indices m and n include all atoms in the chemical species, r_{nm} is the distance between the nth and mth atoms, and $f_n(q)$ and $f_m(q)$ are correspondingly the atomic form factors of the n-type atoms and m-type atoms.

The solute-only signal is one of the components of the TRXL signal stemming from the structural changes of the solute molecules. Some of the reactant molecules excited by the pump pulse transform into intermediates or products. The associated changes in the intramolecular atomic coordination affect the solute-only signal, which can be calculated according to the following equation,

$$\Delta S_{\text{solute}} = \frac{r_{\text{str}}\, c_{\text{solu}}}{c_{\text{solv}}} \left(S_{\text{r}} - S_{\text{p}}\right). \quad (8)$$

Here c_{solu} is the concentration of the solute; c_{solv} is the concentration of the solvent, which is used to scale the amplitude of the solute-only signal to one mole of the solvent molecules; $S_{\text{r}}(q)$ and $S_{\text{p}}(q)$ are scattering intensities of the reactant and product solute molecules, respectively, which are calculated from the atomic coordinates of the reactants and products using the Debye equation; and r_{str} is the ratio of the solute molecules that are converted to the product through the reaction to the total number of solute molecules.

The solute–solvent cross signal, which is also called the cage signal, results from the interference between the atoms in the solute molecules and those in the solvent molecules. Therefore, the solute–solvent cross signal is sensitive to structural changes of the solvation cage surrounding the solute molecules [see Fig. 1(*b*)]. Typically, the solute–solvent cross signal is obtained from the sine-Fourier transform of pair distribution functions, which can be calculated from MD simulations (see Fig. S3). For the TRXL simulation, it is necessary to obtain the solute–solvent cross signal from the MD simulations for every solute structure. On the other hand, because the scattering intensity is proportional to the square of the electron of the atom, the solute–solvent cross signal becomes negligible if the solute molecule is heavy enough. Because MD simulations for obtaining the solute–solvent cross signal require a large amount of calculation power compared with the other three signals and given that the purpose of *S-cube* is not to provide an accurate theoretical curve but a rapid estimation of the expected signal level relative to the noise level, we decided to use an approximation method rather than relying on the time-consuming MD simulations. Specifically, *S-cube* uses simplified pair distribution functions, which are approximated by trapezoidal functions. More specifically, the pair distribution function between the pair of two different elements A and B was approximated as a trapezoidal function of which the value is zero for distances shorter than the sum of the van der Waals radius of A and B and is equal to unity for distances longer than the sum of the van der Waals radius of A and B and the size of the solute molecule, r_{s}, increasing linearly for distances between the two. The size of the solute molecule, r_{s}, was approximated using the following formulae,

$$c_{\text{s}} = \sum_{i=1}^{N_{\text{A}}} \frac{r_i}{N_{\text{A}}}, \qquad r_{\text{s}} = \sum_{i=1}^{N_{\text{A}}} \frac{\left|r_{\text{i}} - c_{\text{s}}\right|}{N_{\text{A}}}, \quad (9)$$

where c_{s} is the center of the solute molecule, r_i is the atomic coordinates of the ith atom in the solute molecule, N_{A} is the number of atoms in the solute molecule and $|r_i - c_{\text{s}}|$ is the

distance of the ith atom in the solute molecule from the center of the molecule. Equation (9) is similar to the formula for obtaining the radius of gyration of a molecule but is different in that the contribution of each atom by its atomic mass is ignored in order to focus solely on the distribution of the positions of the atoms.

The solvent-only signal arises from the changes in the temperature and density of the solvent which originate from the heat released from the light-absorbing solute molecule. The solvent-only signal can be expressed by the following equation,

$$\Delta S_{\rm solvent} = \Delta T\left(\frac{\partial S}{\partial T}\right)_{\rho} + \Delta\rho\left(\frac{\partial S}{\partial \rho}\right)_{T}. \tag{10}$$

In equation (10), the two differentials $(\delta S/\delta T)_{\rho}$ and $(\delta S/\delta\rho)_{T}$ or commonly employed solvents such as acetonitrile, cyclohexane, methanol, ethanol, dichloromethane chloroform and carbon tetrachloride are well documented in the literature (Kim *et al.*, 2009; Cammarata *et al.*, 2006). For calculation of the solvent signal, maximum changes of temperature (ΔT) and density ($\Delta\rho$) are calculated from the energy of the pump laser and the number of light-absorbing solute molecules by assuming that all energy absorbed by the excited molecules ($Q_{\rm max}$) is transferred as heat to the solvent without energy loss. This assumption is not strictly accurate, but provides the maximum amount of Q to avoid underestimating the solvent signal,

$$Q_{\rm max} = \frac{h\nu N_{\rm A}(r_{\rm str} + r_{\rm heat})\, c_{\rm solu}}{c_{\rm solv}}. \tag{11}$$

Equation (11) is the maximum heat transformed from the excited solute to the solvent where, h (J s^{-1}) is Planck's constant, ν (s^{-1}) is the frequency of the pump laser, $N_{\rm A}$ is Avogadro's number, $r_{\rm heat}$ is the ratio of the excited solute molecules that do not undergo a subsequent structural transition and only release the absorbed energy as heat relative to the total number of excited solute molecules. The maximum temperature and density changes, $\Delta T_{\rm max}$ and $\Delta\rho_{\rm max}$, are obtained from $Q_{\rm max}$ assuming isochoric and isobaric processes, respectively, via the following equations,

$$\Delta T_{\rm max} = \frac{Q_{\rm max}}{C_{\rm v}}, \tag{12}$$

$$\Delta\rho_{\rm max} = \frac{\rho_0(-\alpha_{\rm p})\, Q_{\rm max}}{C_{\rm p}}. \tag{13}$$

In equations (12) and (13), $C_{\rm v}$ (J mol^{-1} K^{-1}) and $C_{\rm p}$ (J mol^{-1} K^{-1}) are the heat capacities at a constant volume and pressure, respectively; $\alpha_{\rm p}$ (K^{-1}) is the volumetric thermal expansion coefficient, also known as the isobaric dilation constant; and ρ_0 is the density of the solvent. The time dependence of the solvent-only signal in equation (10) can be precisely calculated if the energy levels of all intermediates and products are provided. For a fast calculation, as an approximation, ΔT is set to $\Delta T_{\rm max}$ up to 10 ns and to decrease linearly to reach zero by 3 µs, while $\Delta\rho$ is set to zero up to 10 ns and is set to increase linearly to $\Delta\rho_{\rm max}$ by 3 µs.

The scattering intensity of the solution is represented by the sum of the solute-only, solute–solvent cross and solvent-only signals in the following equation,

$$\begin{aligned} S_{\rm solution}(q) &= S_{\rm solute}(q) + S_{\rm cage}(q) + S_{\rm solvent}(q) \\ &\simeq S_{\rm solvent}(q). \end{aligned} \tag{14}$$

In this equation, $S_{\rm solution}(q)$ denotes the total scattering intensity from the solution and $S_{\rm solute}(q)$, $S_{\rm cage}(q)$ and $S_{\rm solvent}(q)$ denote the components of the total scattering intensities which originate from the contributions of the solute, the solvent molecules (cage) surrounding the solute molecule and the bulk solvent, respectively. In general, $S_{\rm solvent}(q)$ dominates the signal. Because the variance of the scattering intensity is proportional to the scattering intensity TRXL simulation, it can be expressed as follows (Kim *et al.*, 2009),

$$\begin{aligned} \sigma^2_{\rm solution}(q) &= \sigma^2_{\rm solute}(q) + \sigma^2_{\rm cage}(q) + \sigma^2_{\rm solvent}(q) \\ &\simeq \sigma^2_{\rm solvent}(q). \end{aligned} \tag{15}$$

In equation (15), σ_k is the standard deviation of $S_k(q)$. To simplify the estimation, the total standard deviation, $\sigma_{\rm solution}(q)$, can be approximated to be equal to $\sigma_{\rm solvent}(q)$ because the number of solvent molecules is high enough to neglect the other two contributions from the solute molecules and cages, $\sigma_{\rm solute}(q)$ and $\sigma_{\rm cage}(q)$, respectively. In other words, regardless of the type of solute, $\sigma_{\rm sol}(q)$ can be approximated by $\sigma_{\rm solvent}(q)$ which can be measured in a separate experiment on a neat solvent.

6. Related literature

The following references, not cited in the main body of the paper, have been cited in the supporting information: Wulff *et al.* (2003, 2007).

Acknowledgements

This work was supported by the Institute for Basic Science (IBS-R004). The experiment for comparison of simulation and experiment in synchrotrons was performed in the ID09 beamline of the European Synchrotron Radiation Facility (ESRF). We acknowledge the ESRF for provision of synchrotron radiation facilities, and we would like to thank Sungjun Park, Yunbeom Lee, Matteo Levantino and Michael Wulff for their assistance in using the ID09B beamline and their support for the experiment. The experiment for comparison of simulation and experiment in XFELs was performed at XSS-FXL end-station of PAL-XFEL (proposal No. 2017-2nd-XSS-001) funded by the Ministry of Science and ICT of Korea. We appreciate Yunbeom Lee and Eun Hyuk Choi for insightful discussions. We thank Hanui Kim, Yunbeom Lee, Seonggon Lee, Minseo Choi and Jae Hyuk Lee for their support during the experiment at PAL-XFEL and Sungjun Park, Jun Heo and Deokbeom Suh for providing their MD simulation results.

Funding information

The following funding is acknowledged: Institute for Basic Science (grant No. IBS-R004).

References

Ahn, C. W., Ki, H., Kim, J., Kim, J., Park, S., Lee, Y., Kim, K. H., Kong, Q., Moon, J., Pedersen, M. N., Wulff, M. & Ihee, H. (2018). *J. Phys. Chem. Lett.* **9**, 647–653.

Andersson, M., Malmerberg, E., Westenhoff, S., Katona, G., Cammarata, M., Wöhri, A. B., Johansson, L. C., Ewald, F., Eklund, M., Wulff, M., Davidsson, J. & Neutze, R. (2009). *Structure*, **17**, 1265–1275.

Andersson, M., Vincent, J., van der Spoel, D., Davidsson, J. & Neutze, R. (2008). *Structure*, **16**, 21–28.

Arnlund, D., Johansson, L. C., Wickstrand, C., Barty, A., Williams, G. J., Malmerberg, E., Davidsson, J., Milathianaki, D., DePonte, D. P., Shoeman, R. L., Wang, D. J., James, D., Katona, G., Westenhoff, S., White, T. A., Aquila, A., Bari, S., Berntsen, P., Bogan, M., van Driel, T. B., Doak, R. B., Kjaer, K. S., Frank, M., Fromme, R., Grotjohann, I., Henning, R., Hunter, M. S., Kirian, R. A., Kosheleva, I., Kupitz, C., Liang, M. N., Martin, A. V., Nielsen, M. M., Messerschmidt, M., Seibert, M. M., Sjöhamn, J., Stellato, F., Weierstall, U., Zatsepin, N. A., Spence, J. C. H., Fromme, P., Schlichting, I., Boutet, S., Groenhof, G., Chapman, H. N. & Neutze, R. (2014). *Nat. Methods*, **11**, 923–926.

Bandara, H. M. & Burdette, S. C. (2012). *Chem. Soc. Rev.* **41**, 1809–1825.

Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. (2007). *J. Am. Chem. Soc.* **129**, 5656–5664.

Berntsson, O., Diensthuber, R. P., Panman, M. R., Björling, A., Hughes, A. J., Henry, L., Niebling, S., Newby, G., Liebi, M., Menzel, A., Henning, R., Kosheleva, I., Möglich, A. & Westenhoff, S. (2017). *Structure*, **25**, 933–938.

Biasin, E., van Driel, T. B., Levi, G., Laursen, M. G., Dohn, A. O., Moltke, A., Vester, P., Hansen, F. B. K., Kjaer, K. S., Harlang, T., Hartsock, R., Christensen, M., Gaffney, K. J., Henriksen, N. E., Møller, K. B., Haldrup, K. & Nielsen, M. M. (2018). *J. Synchrotron Rad.* **25**, 306–315.

Bortolus, P. & Monti, S. (1979). *J. Phys. Chem.* **83**, 648–652.

Brandt van Driel, T., Kjaer, K. S., Biasin, E., Haldrup, K., Lemke, H. T. & Nielsen, M. M. (2015). *Faraday Discuss.* **177**, 443–465.

Cammarata, M., Levantino, M., Schotte, F., Anfinrud, P. A., Ewald, F., Choi, J., Cupane, A., Wulff, M. & Ihee, H. (2008). *Nat. Methods*, **5**, 988.

Cammarata, M., Lorenc, M., Kim, T. K., Lee, J. H., Kong, Q. Y., Pontecorvo, E., Lo Russo, M., Schiró, G., Cupane, A., Wulff, M. & Ihee, H. (2006). *J. Chem. Phys.* **124**, 124504.

Canton, S. E., Kjaer, K. S., Vanko, G., van Driel, T. B., Adachi, S. I., Bordage, A., Bressler, C., Chabera, P., Christensen, M., Dohn, A. O., Galler, A., Gawelda, W., Gosztola, D., Haldrup, K., Harlang, T., Liu, Y. Z., Moller, K. B., Nemeth, Z., Nozawa, S., Papai, M., Sato, T., Sato, T., Suarez-Alcantara, K., Togashi, T., Tono, K., Uhlig, J., Vithanage, D. A., Warnmark, K., Yabashi, M., Zhang, J. X., Sundstrom, V. & Nielsen, M. M. (2015). *Nat. Commun.* **6**, 1–10.

Cho, H. S., Dashdorj, N., Schotte, F., Graber, T., Henning, R. & Anfinrud, P. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 7281–7286.

Christensen, M., Haldrup, K., Bechgaard, K., Feidenhans'l, R., Kong, Q. Y., Cammarata, M., Russo, M. L., Wulff, M., Harrit, N. & Nielsen, M. M. (2009). *J. Am. Chem. Soc.* **131**, 502–508.

Davidsson, J., Poulsen, J., Cammarata, M., Georgiou, P., Wouts, R., Katona, G., Jacobson, F., Plech, A., Wulff, M., Nyman, G. & Neutze, R. (2005). *Phys. Rev. Lett.* **94**, 245503.

Dohn, A. O., Biasin, E., Haldrup, K., Nielsen, M. M., Henriksen, N. E. & Møller, K. B. (2015). *J. Phys. B At. Mol. Opt. Phys.* **48**, 244010.

Förster, F., Webb, B., Krukenberg, K. A., Tsuruta, H., Agard, D. A. & Sali, A. (2008). *J. Mol. Biol.* **382**, 1089–1106.

Haldrup, K. (2014). *Philos. Trans. R. Soc. B*, **369**, 20130336.

Haldrup, K., Christensen, M., Cammarata, M., Kong, Q. Y., Wulff, M., Mariager, S. O., Bechgaard, K., Feidenhans'l, R., Harrit, N. & Nielsen, M. M. (2009). *Angew. Chem. Int. Ed.* **48**, 4180–4184.

Haldrup, K., Christensen, M. & Meedom Nielsen, M. (2010). *Acta Cryst.* A**66**, 261–269.

Haldrup, K., Levi, G., Biasin, E., Vester, P., Laursen, M. G., Beyer, F., Kjaer, K. S., Brandt van Driel, T., Harlang, T., Dohn, A. O., Hartsock, R. J., Nelson, S., Glownia, J. M., Lemke, H. T., Christensen, M., Gaffney, K. J., Henriksen, N. E., Møller, K. B. & Nielsen, M. M. (2019). *Phys. Rev. Lett.* **122**, 063001.

Haldrup, K., Vankó, G., Gawelda, W., Galler, A., Doumy, G., March, A. M., Kanter, E. P., Bordage, A., Dohn, A., van Driel, T. B., Kjaer, K. S., Lemke, H. T., Canton, S. E., Uhlig, J., Sundström, V., Young, L., Southworth, S. H., Nielsen, M. M. & Bressler, C. (2012). *J. Phys. Chem. A*, **116**, 9878–9887.

Ihee, H. (2009). *Acc. Chem. Res.* **42**, 356–366.

Ihee, H., Lorenc, M., Kim, T. K., Kong, Q. Y., Cammarata, M., Lee, J. H., Bratos, S. & Wulff, M. (2005). *Science*, **309**, 1223–1227.

Ihee, H., Wulff, M., Kim, J. & Adachi, S. (2010). *Int. Rev. Phys. Chem.* **29**, 453–520.

Josts, I., Niebling, S., Gao, Y., Levantino, M., Tidow, H. & Monteiro, D. (2018). *IUCrJ*, **5**, 667–672.

Kang, H. S., Min, C. K., Heo, H., Kim, C., Yang, H., Kim, G., Nam, I., Baek, S. Y., Choi, H. J., Mun, G., Park, B. R., Suh, Y. J., Shin, D. C., Hu, J., Hong, J., Jung, S., Kim, S. H., Kim, K., Na, D., Park, S. S., Park, Y. J., Han, J. H., Jung, Y. G., Jeong, S. H., Lee, H. G., Lee, S., Lee, S., Lee, W. W., Oh, B., Suh, H. S., Parc, Y. W., Park, S. J., Kim, M. H., Jung, N. S., Kim, Y. C., Lee, M. S., Lee, B. H., Sung, C. W., Mok, I. S., Yang, J. M., Lee, C. S., Shin, H., Kim, J. H., Kim, Y., Lee, J. H., Park, S. Y., Kim, J., Park, J., Eom, I., Rah, S., Kim, S., Nam, K. H., Park, J., Park, J., Kim, S., Kwon, S., Park, S. H., Kim, K. S., Hyun, H., Kim, S. N., Kim, S., Hwang, S. M., Kim, M. J., Lim, C. Y., Yu, C. J., Kim, B. S., Kang, T. H., Kim, K. W., Kim, S. H., Lee, H. S., Lee, H. S., Park, K. H., Koo, T. Y., Kim, D. E. & Ko, I. S. (2017). *Nat. Photon.* **11**, 708–713.

Ki, H., Lee, Y., Choi, E. H., Lee, S. & Ihee, H. (2019). *Struct. Dyn.* **6**, 024303.

Ki, H., Oang, K. Y., Kim, J. & Ihee, H. (2017). *Annu. Rev. Phys. Chem.* **68**, 473–497.

Kim, J., Lee, J. H., Kim, J., Jun, S., Kim, K. H., Kim, T. W., Wulff, M. & Ihee, H. (2012). *J. Phys. Chem. A*, **116**, 2713–2722.

Kim, J. G., Muniyappan, S., Oang, K. Y., Kim, T. W., Yang, C., Kim, K. H., Kim, J. & Ihee, H. (2016). *Struct. Dyn.* **3**, 023610.

Kim, K. H., Kim, J. G., Oang, K. Y., Kim, T. W., Ki, H., Jo, J., Kim, J., Sato, T., Nozawa, S., Adachi, S. & Ihee, H. (2016). *Struct. Dyn.* **3**, 043209.

Kim, S., Kim, S., Kim, M., Hwang, S., Hyun, H., Eom, I., Park, K., Park, J., Park, J., Kim, K., Kim, S., Kim, S., Lee, C. & Nah, S. (2018). *Proceedings of the 10th Mechanical Engineering Design of Synchrotron Radiation Equipment and Instrumentation (MEDSI2018)*, 25–29 June 2018, Paris, France, pp. 397–399. THPH28.

Kim, T. K., Lee, J. H., Wulff, M., Kong, Q. Y. & Ihee, H. (2009). *ChemPhysChem*, **10**, 1958–1980.

Kim, T. W., Yang, C., Kim, Y., Kim, J. G., Kim, J., Jung, Y. O., Jun, S., Lee, S. J., Park, S., Kosheleva, I., Henning, R., van Thor, J. J. & Ihee, H. (2016). *Phys. Chem. Chem. Phys.* **18**, 8911–8919.

Kim, Y., Ganesan, P., Jo, J., Kim, S. O., Thamilselvan, K. & Ihee, H. (2018). *J. Phys. Chem. B*, **122**, 4513–4520.

Kirian, R. A., Schmidt, K. E., Wang, X. Y., Doak, R. B. & Spence, J. C. H. (2011). *Phys. Rev. E*, **84**, 011921.

Kjaer, K. S., Van Driel, T. B., Harlang, T. C. B., Kunnus, K., Biasin, E., Ledbetter, K., Hartsock, R. W., Reinhard, M. E., Koroidov, S., Li, L., Laursen, M. G., Hansen, F. B., Vester, P., Christensen, M., Haldrup, K., Nielsen, M. M., Dohn, A. O., Pápai, M. I., Møller, K. B., Chabera, P., Liu, Y. Z., Tatsuno, H., Timm, C., Jarenmark, M., Uhlig, J., Sundström, V., Wärnmark, K., Persson, P., Németh, Z.,

Szemes, D. S., Bajnóczi, É., Vankó, G., Alonso-Mori, R., Glownia, J. M., Nelson, S., Sikorski, M., Sokaras, D., Canton, S. E., Lemke, H. T. & Gaffney, K. J. (2019). *Chem. Sci.* **10**, 5749–5760.
Kong, Q. Y., Laursen, M. G., Haldrup, K., Kjaer, K. S., Khakhulin, D., Biasin, E., van Driel, T. B., Wulff, M., Kabanova, V., Vuilleumier, R., Bratos, S., Nielsen, M. M., Gaffney, K. J., Weng, T. C. & Koch, M. H. J. (2019). *Photochem. Photobiol. Sci.* **18**, 319–327.
Kong, Q. Y., Wulff, M., Lee, J. H., Bratos, S. & Ihee, H. (2007). *J. Am. Chem. Soc.* **129**, 13584–13591.
Konuma, T., Kimura, T., Matsumoto, S., Goto, Y., Fujisawa, T., Fersht, A. R. & Takahashi, S. (2011). *J. Mol. Biol.* **405**, 1284–1294.
LCLS (2018). *LCLS Strategic Facility Development Plan*, https://lcls.slac.stanford.edu/sites/lcls.slac.stanford.edu/files/LCLS_Strategic_Development_Plan.pdf.
LCLS-II (2018). *LCLS-II Design and Performance*, https://lcls.slac.stanford.edu/lcls-ii/design-and-performance.
Leshchev, D., Harlang, T. C. B., Fredin, L. A., Khakhulin, D., Liu, Y., Biasin, E., Laursen, M. G., Newby, G. E., Haldrup, K., Nielsen, M. M., Wärnmark, K., Sundström, V., Persson, P., Kjaer, K. S. & Wulff, M. (2018). *Chem. Sci.* **9**, 405–414.
Malmerberg, E., Bovee-Geurts, P. H. M., Katona, G., Deupi, X., Arnlund, D., Wickstrand, C., Johansson, L. C., Westenhoff, S., Nazarenko, E., Schertler, G. F. X., Menzel, A., de Grip, W. J. & Neutze, R. (2015). *Sci. Signal.* **8**, ra26.
Malmerberg, E., Omran, Z., Hub, J. S., Li, X. W., Katona, G., Westenhoff, S., Johansson, L. C., Andersson, M., Cammarata, M., Wulff, M., van der Spoel, D., Davidsson, J., Specht, A. & Neutze, R. (2011). *Biophys. J.* **101**, 1345–1353.
Mathew-Fenn, R. S., Das, R. & Harbury, P. A. B. (2008). *Science*, **322**, 446–449.
Neutze, R., Wouts, R., Techert, S., Davidsson, J., Kocsis, M., Kirrander, A., Schotte, F. & Wulff, N. (2001). *Phys. Rev. Lett.* **87**, 195508.
Pinfield, V. J. & Scott, D. J. (2014). *PLoS One*, **9**, e95664.
Plech, A., Wulff, M., Bratos, S., Mirloup, F., Vuilleumier, R., Schotte, F. & Anfinrud, P. A. (2004). *Phys. Rev. Lett.* **92**, 125505.
Rimmerman, D., Leshchev, D., Hsu, D. J., Hong, J., Kosheleva, I. & Chen, L. X. (2017). *J. Phys. Chem. Lett.* **8**, 4413–4418.
Salassa, L., Borfecchia, E., Ruiu, T., Garino, C., Gianolio, D., Gobetto, R., Sadler, P. J., Cammarata, M., Wulff, M. & Lamberti, C. (2010). *Inorg. Chem.* **49**, 11240–11248.
Schindler, C. E. M., de Vries, S. J., Sasse, A. & Zacharias, M. (2016). *Structure*, **24**, 1387–1397.
Schneidman-Duhovny, D., Hammel, M. & Sali, A. (2010). *Nucleic Acids Res.* **38**, W540–W544.
Schoenlein, R., Boutet, S., Minitti, M. & Dunne, A. (2017). *J. Appl. Sci.* **7**, 850.
Schultz, T., Quenneville, J., Levine, B., Toniolo, A., Martínez, T. J., Lochbrunner, S., Schmitt, M., Shaffer, J. P., Zgierski, M. Z. & Stolow, A. (2003). *J. Am. Chem. Soc.* **125**, 8098–8099.
Sedlak, S. M., Bruetzel, L. K. & Lipfert, J. (2017). *J. Appl. Cryst.* **50**, 621–630.
Stovgaard, K., Andreetta, C., Ferkinghoff-Borg, J. & Hamelryck, T. (2010). *BMC Bioinformatics*, **11**, 429.
Stuart, C. M., Frontiera, R. R. & Mathies, R. A. (2007). *J. Phys. Chem. A*, **111**, 12072–12080.
Westenhoff, S., Malmerberg, E., Arnlund, D., Johansson, L., Nazarenko, E., Cammarata, M., Davidsson, J., Chaptal, V., Abramson, J., Katona, G., Menzel, A. & Neutze, R. (2010). *Nat. Methods*, **7**, 775–776.
Wulff, M., Kong, Q., Cammarata, M., Lo Russo, M., Anfinrud, P., Schotte, F., Lorenc, M., Ihee, H., Kim, T. K. & Plech, A. (2007). *AIP Conf. Proc.* **879**, 1187–1194.
Wullf, M., Plech, A., Eybert, L., Randler, R., Schotte, F. & Anfinrud, P. (2003). *Faraday Discuss.* **122**, 13–26.