Bayesian orientation estimate and structure information from sparse single-molecule x-ray diffraction images

Michał Walczak and Helmut Grubmüller*

Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany

(Received 17 October 2013; published 20 August 2014)

We developed a Bayesian method to extract macromolecular structure information from sparse single-molecule x-ray free-electron laser diffraction images. The method addresses two possible scenarios. First, using a "seed" structural model, the molecular orientation is determined for each of the provided diffraction images, which are then averaged in three-dimensional reciprocal space. Subsequently, the real space electron density is determined using a relaxed averaged alternating reflections algorithm. In the second approach, the probability that the "seed" model fits to the given set of diffraction images as a whole is determined and used to distinguish between proposed structures. We show that for a given x-ray intensity, unexpectedly, the achievable resolution *increases* with molecular mass such that structure determination should be more challenging for small molecules than for larger ones. For a sufficiently large number of recorded photons (>200) per diffraction image an $M^{1/6}$ scaling is seen. Using synthetic diffraction data for a small glutathione molecule as a challenging test case, successful determination of electron density was demonstrated for 20 000 diffraction patterns with random orientations and an average of 82 elastically scattered and recorded photons per image, also in the presence of up to 50% background noise. The second scenario is exemplified and assessed for three biomolecules of different sizes. In all cases, determining the probability of a structure given set of diffraction patterns allowed successful discrimination between different conformations of the test molecules. A structure model of the glutathione tripeptide was refined in a Monte Carlo simulation from a random starting conformation. Further, effective distinguishing between three differently arranged immunoglobulin domains of a titin molecule and also different states of a ribosome in a tRNA translocation process was demonstrated. These results show that the proposed method is robust and enables structure determination from sparse and noisy x-ray diffraction images of single molecules spanning a wide range of molecular masses.

DOI: 10.1103/PhysRevE.90.022714 PACS number(s): 87.64.Bx, 87.15.B−

I. INTRODUCTION

X-ray crystallography is a powerful tool to obtain high-resolution structural information of macromolecules. However, this technique requires a crystalline specimen; hence, many proteins that cannot be crystallized are inaccessible. Further, the crystal environment may sterically hinder conformational motions, such that structural heterogeneity, often crucial for protein function, may be altered or suppressed. Another limitation of x-ray crystallography is the phase problem. Because only the intensities of discrete Bragg reflections are measured, the phases need to be retrieved by other means [1–3].

A step towards single-molecule x-ray scattering experiments is nanocrystallography. While this technique still requires crystalline specimen, it offers a clear advantage in cases where all attempts to grow larger crystals, as required for traditional crystallography, have failed, but nanocrystals have been grown successfully. In 2011, Chapman *et al.* [4] reported a 8.5 Å resolution structure from Photosystem I nanocrystals using a hard x-ray free-electron laser (XFEL), and, recently, a 2.1 Å resolution lysozyme structure was determined *de novo* from microcrystals [5].

Scattering experiments on single molecules, using ultrashort XFEL pulses, hold the promise to overcome the above limitations of traditional x-ray crystallography [6]. In such experiments, a stream of hydrated particles enters the x-ray beam at a high rate, ideally of one molecule per pulse [7], and one diffraction image per molecule is recorded. Many ($10^4 \ldots 10^6$) diffraction images can thus be recorded, but only relatively few elastically scattered photons (of the order of $10 \ldots 1000$, depending on molecular mass) per diffraction image are expected, including substantial noise [8]. Similarly to single-molecule cryoelectron microscopy, the images need to be sufficiently accurately aligned and averaged to obtain sufficient signal-to-noise ratios; here it is for the Fourier transform of the electron density in reciprocal space. Because it is challenging to experimentally control the orientation of each molecule during scattering for proper classification prior to averaging, this orientation is here assumed to be random and thus needs to be determined *a posteriori* from the obtained diffraction images.

While in x-ray crystallography low-intensity radiation is distributed over many molecules in the crystal lattice, in XFEL experiments extremely high doses will be absorbed by a single molecule within the very short time of few femtoseconds. For example, after focusing to an ∼100-nm diameter spot, the flux of XFEL pulses will be higher than that of the synchrotron-radiation used in x-ray crystallography by at least a factor of 10^6 [9]. As a result, during the XFEL pulse, each atom of the target molecule will absorb several photons, which will knock out many electrons via the photo effect and subsequent Auger processes. Due to the perturbed chemistry, and in particular due to the substantial excess charge of the nuclei, the molecule will undergo a rapid Coulomb explosion [6]. Hence, if the

*hgrubmu@gwdg.de

pulse is too long, the recorded diffraction image will be compromised [9,10]. Therefore, femtosecond exposure times are essential to enable recording of the diffraction pattern before the initial structure suffers severely from ionization effects due to the very high radiation dose. Achieving pulse lengths in the femtosecond regime will therefore also allow us to obtain time-resolved structural information, which is crucial to elucidate functional conformational dynamics, e.g., during enzymatic catalysis or protein folding dynamics [11].

In the absence of crystal symmetries, the diffraction pattern is continuous, which enables one to oversample the molecular transform [12]. Recently developed iterative phasing algorithms allow one to exploit this additional information to generate the phases required for back-transforming the absolute-squared Fourier amplitudes and thus to retrieve the real space electron density from the available data [8,12–16].

Here, we will address the question regarding what resolution can be achieved in these experiments and whether the high intensity delivered by the XFEL pulse yields sufficiently many elastically scattered photons from a single molecule to be able to obtain the molecule's electron density from averaging the diffraction images. Indeed, only very few photons are expected for one scattering event. For example, a 500-kDa molecule is expected to scatter only about 4×10^{-2} per Shannon pixel for mean photon count in the high-resolution part [8], assuming the XFEL beam is focused to a 100-nm diameter spot. Obtaining such a small focal spot is challenging; in fact, for recently conducted experiments with ultraintense x rays a focal spot close to 100 nm was achieved [17]. The resulting very low signal-to-noise ratio implies that only very little information on the molecular structure is contained within each single diffraction pattern. Second, even at higher signal-to-noise ratios, each single diffraction pattern provides only partial information about the molecule, as the detector plane covers only part of a certain Ewald sphere in reciprocal space. As in x-ray crystallography, many different orientations of the molecule need to be recorded to fully sample three-dimensional (3D) reciprocal space. Whether this is indeed possible under the given conditions, and whether atomistic resolution can be achieved, critically depends on the ability to determine the unknown molecular orientation for each of the many recorded diffraction images with sufficient accuracy or to circumvent this problem.

To this end, Huldt *et al.* [18] proposed the "common line" orientation determination method. This approach rests on the fact that two Ewald spheres, corresponding to diffraction patterns recorded on the detector plane, intersect in reciprocal space, creating a common curve. Locating the common line in any three diffraction patterns suffices to determine the relative orientations of the Ewald spheres. However, because of the expected low photon count the images have to be averaged first. Huldt *et al.* have proposed to group the diffraction patterns by evaluating the cross-correlation function between any two of them, which limits the applicability of this method to mean photon counts of 10 per pixel or more [8], i.e., three orders of magnitude higher than the expected XFEL values.

As an alternative, a method suggested by Fung *et al.* [19] uses generative topographic mapping to determine a maximum likelihood manifold in the orientational space, which serves to arrange the diffraction patterns into orientation classes. A clear advantage of this approach is the fact that the only input required, apart from the diffraction patterns, is the dimensionality of the orientational space. However, averaging of the diffraction patterns within determined orientation classes might lead to information loss, resulting from insufficient sampling of 3D reciprocal space as we will show in the Results section. Also, the number of elastically scattered photons required per image, about 100 (excluding the innermost central pixels), is rather high for small molecules. A similar method exploiting symmetries resulting from image formation was proposed recently [20].

Loh and Elser [21] proposed an expansion–expectation maximization–compression (EMC) method to iteratively maximize the likelihood of an intensity model of an irradiated molecule in reciprocal space with respect to a set of diffraction images. In their approach, the orientation of the images was estimated using an intensity model, which was further updated by averaging the images in 3D reciprocal space. It was demonstrated that the method is capable of determining the structure of a GroEL molecule at 2-nm resolution from simulated diffraction images. A similar approach was developed by Tegze and Bortel [22].

A complementary route was taken recently by eliminating the need to determine the individual orientations altogether [23,24]. Rather, the molecular shape is represented by a spherical harmonic expansion of the diffraction intensity in reciprocal space, determined from cross-correlations between diffraction images. However, it is unclear how much detail can be extracted, and similarly to the correlation based "common line" method, realistic signal-to-noise ratios will probably remain inaccessible also for this method.

Recently, Liu *et al.* demonstrated a Monte Carlo refinement of a low-resolution electron-density model by the angular correlation function of many diffraction images [25]. Similarly to approaches used in small-angle x-ray scattering experiments, the authors use a grid representation of the electron density, which they locally perturb by performing random dilations or erosions and compare the resulting correlations with the ones rising from the experimental data. It is worth noting that this method is applied to diffraction images obtained from many identical, randomly oriented molecules per exposure. In contrast, Oroguchi and Nakasako have suggested reconstructing the 3D electron density from two-dimensional (2D) projections from diffraction images of multiple copies of a molecule [26].

The correlation approach was also pursued by Starodub *et al.* [27]. By using partial triple correlation of scattered intensity distribution an electron-density map of two polystyrene spheres with a diameter of 91 nm was obtained with a 20-nm resolution. Although the demonstrated procedure was restricted to an object with cylindrical symmetry, which reduces the complexity of correlation analysis, a more general treatment utilizing full triple correlation analysis should enable one to obtain high-resolution structure information in the absence of any symmetry [24].

To overcome the limitations of low photon counts and to be able to address realistic signal-to-noise ratios, while still achieving atomic resolution, we here discuss two complementary approaches (Fig. 1). The first is similar to the EMC algorithm [21] in that for each picture its orientation is determined via a rigorous Bayesian framework from very sparse

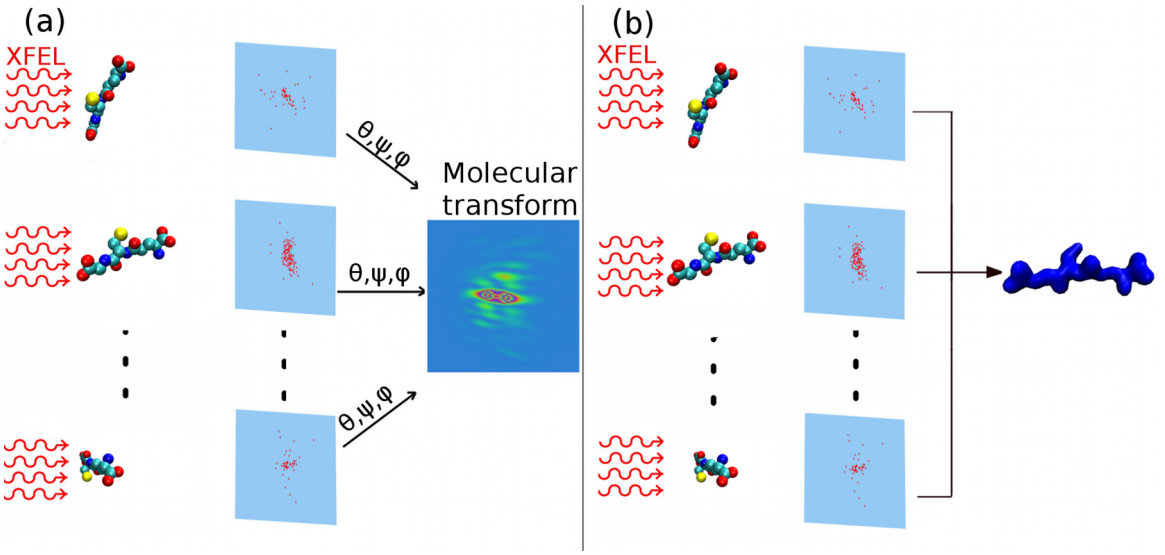

FIG. 1. (Color online) Two approaches to structure determination from single-molecule diffraction experiments. For each recorded image, photon arrival positions are shown in red (dark gray) dots. (a) The spatial orientation (θ,ψ,φ) of the irradiated molecule, here a glutathione, is determined for each of the images separately, and the molecular transform is derived by averaging the images in reciprocal space. (b) A structure that fits best to the entire set of photon arrival positions is determined.

data using a "seed" model. Our approach differs from the EMC algorithm [21] in that it considers probabilities of all individual photons instead of describing the diffraction images in terms of photon counts per pixel and using a noise model such as the Poisson approximation. Thereby our approach captures photon counts for which the Poisson approximation does not hold. We note that for the large number of incident photons expected in XFEL experiments, our more general formulation and the Poisson approximation should yield similar results. We apply this statistical approach to accurately determine the molecular orientation for single XFEL diffraction images individually [Fig. 1(a)] for subsequent averaging. The achievable accuracy will then be assessed as a function of molecular mass, incident beam intensity, and background noise level. Using simulated XFEL data, it will be shown that the averaged intensity in reciprocal space is accurate enough to allow the calculation of electron densities at atomistic resolution and at up to 50% background noise levels, which so far has not been demonstrated.

In the second approach, we propose to use the Bayesian framework to find, among several given candidate structures, the one which fits best to the given set of diffraction patterns as a whole. This approach is an instance of Bayesian model comparison and aims at obtaining ratios of evidence among different structures as models in order to distinguish among them [28]. In that case, instead of determining the orientation for every single image separately, the probability of the model structure is calculated from the entirety of all possible orientations of the model and from the whole set of given diffraction patterns [Fig. 1(b)]. For the small tripeptide as a test case, it is demonstrated that this approach allows for the *de novo* structure determination via a Monte Carlo (MC) approach. For larger molecules, for which the search space becomes too large, we will demonstrate that it is still possible to distinguish among different conformations of three Ig domains of a titin molecule as well as among different states of a ribosome during the tRNA translocation process. In all of these cases, by calculating the probability of a structure given set of images for each of selected structures, the most probable structure is determined.

II. RESULTS AND DISCUSSION

Our approach is similar in spirit to the one developed recently for single-molecule Förster resonance electron transfer (FRET) experiments, which, by applying Bayes's theorem, $\pi(\mathbf{\Theta}|\mathbf{X}) \propto f(\mathbf{X}|\mathbf{\Theta})\, p(\mathbf{\Theta})$, allows for high-resolution reconstruction of distance trajectories from very few recorded FRET photons [29]. Here, the posterior probability distribution $\pi(\mathbf{\Theta}|\mathbf{X})$ of orientations $\mathbf{\Theta}$ given a single diffraction pattern $\mathbf{X}$ is computed from the conditional probability density (or likelihood) $f(\mathbf{X}|\mathbf{\Theta})$ that a particular diffraction pattern is observed *given a particular orientation* and the *a priori* orientation distribution $p(\mathbf{\Theta})$. This information is then used to assemble the 3D reciprocal space density from many 2D Ewald projections, each from one of the recorded diffraction images. In the second approach, again using Bayes's formula, the posterior probability $\pi(\mathbf{S}|\{\mathbf{X}\})$ of a whole structure $\mathbf{S}$ given a complete set of diffraction patterns $\{\mathbf{X}\}$ is calculated and used for structure discrimination.

A. Posterior probability distribution

We assume a total number of incident photons $N_{\text{total}} = I_0 F_{\text{A}}$, resulting from focusing an XFEL beam with an intensity I_0 to a focal spot area F_{A}. Of those, for each recorded diffraction image i, n_i photons are recorded at positions $\mathbf{X}_i = \{(x_i^{(l)}, y_i^{(l)})\}_{l=1\ldots n_i}$ on the image plane; all other photons are not recorded. We denote the orientation of the molecule entering the beam by $\mathbf{\Theta}_i = (\theta_i, \psi_i, \varphi_i)$. The probability of registering a particular configuration of recorded photons $\mathbf{X}_i$, given a certain molecular orientation, is thus expressed by a product of independent probabilities of detecting a photon at

position $(x_i^{(l)}, y_i^{(l)})$ and the probability of $N_{\text{total}} - n_i$ photons not being recorded

$$f(\mathbf{X}_i|\mathbf{\Theta}_i) \propto \left(1 - \frac{A_{\mathbf{\Theta}_i}}{N_{\text{total}}}\right)^{N_{\text{total}}-n_i} \prod_{l=1}^{n_i} \frac{I_{\mathbf{\Theta}_i}\left[\Delta\mathbf{k}\left(x_i^{(l)}, y_i^{(l)}\right)\right]}{N_{\text{total}}}$$

$$\propto \left(1 - \frac{A_{\mathbf{\Theta}_i}}{N_{\text{total}}}\right)^{N_{\text{total}}-n_i} \prod_{l=1}^{n_i} I_{\mathbf{\Theta}_i}\left[\Delta\mathbf{k}\left(x_i^{(l)}, y_i^{(l)}\right)\right]. \quad (1)$$

Here $I_{\mathbf{\Theta}_i}[\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)})]$ is the intensity value (see Sec. III) in a detector pixel corresponding to the recorded photon position $(x_i^{(l)}, y_i^{(l)})$ and $A_{\mathbf{\Theta}_i} = \sum_{l=1}^{N_{\text{pixel}}} I_{\mathbf{\Theta}_i}[\Delta\mathbf{k}(x^{(l)}, y^{(l)})]$ is the expected amount of elastic scattering for orientation $\mathbf{\Theta}_i$ registered by the detector with N_{pixel} pixels.

Equation (1) describes a multinomial distribution, up to a constant combinatorial factor that cancels in Bayes's formula. Note that the distribution in Eq. (1) by construction accounts for shot noise with a mean equal to the expected number of photons per pixel. Correspondingly, in the limit of small probabilities of detecting a photon and $N_{\text{total}} \gg n_i$, this distribution converges towards a Poisson distribution with respect to photon counts per pixel. Additional background noise, e.g., due to inelastic scattering, is included by an appropriate noise model as described in Sec. III.

Because the *a priori* probability distribution for the (unknown) orientation of the scattering molecule can be assumed isotropic, the probability distribution for the molecular orientation *given the recorded photon positions*, $\pi(\mathbf{\Theta}_i|\mathbf{X}_i)$ is—via Bayes's formula—also given by Eq. (1), up to an irrelevant normalization factor.

Note that this approach requires a "seed model" for the respective probability calculations. As a first step, to explore the achievable resolution, we will use the reference structure as the "seed model," which in our synthetic setting is, of course, known. Subsequently, in a second step this somewhat circular requirement will be dropped within the context of finding a structure best fitting to a given set of diffraction images.

As a test case, we simulated XFEL diffraction images for the tripeptide glutathione and subsequently derived the posterior orientation distribution as described above. To this end, we have computed the intensity distribution on the detector plane by Fourier transforming the electron density of the test molecule (as described in Sec. III) for a chosen orientation. From the obtained intensity distribution, Poisson distributed random numbers were drawn to determine the number of photons for each detector pixel. An average photon count of about 82 (shot noise only) was chosen, including the innermost central pixels that are usually protected by the beam stop. Excluding this central peak, 54 photons per diffraction image were considered on average. A total of 20 000 diffraction pictures was calculated and used for the electron-density calculation. Background noise was simulated by including additional Gaussian distributed photons corresponding to 10% and 50% of the average photon count per picture.

Figure 2 shows an example of a cut (ψ φ plane) through the three-dimensional posterior probability surface thus obtained for one particular diffraction image. A clear maximum can be

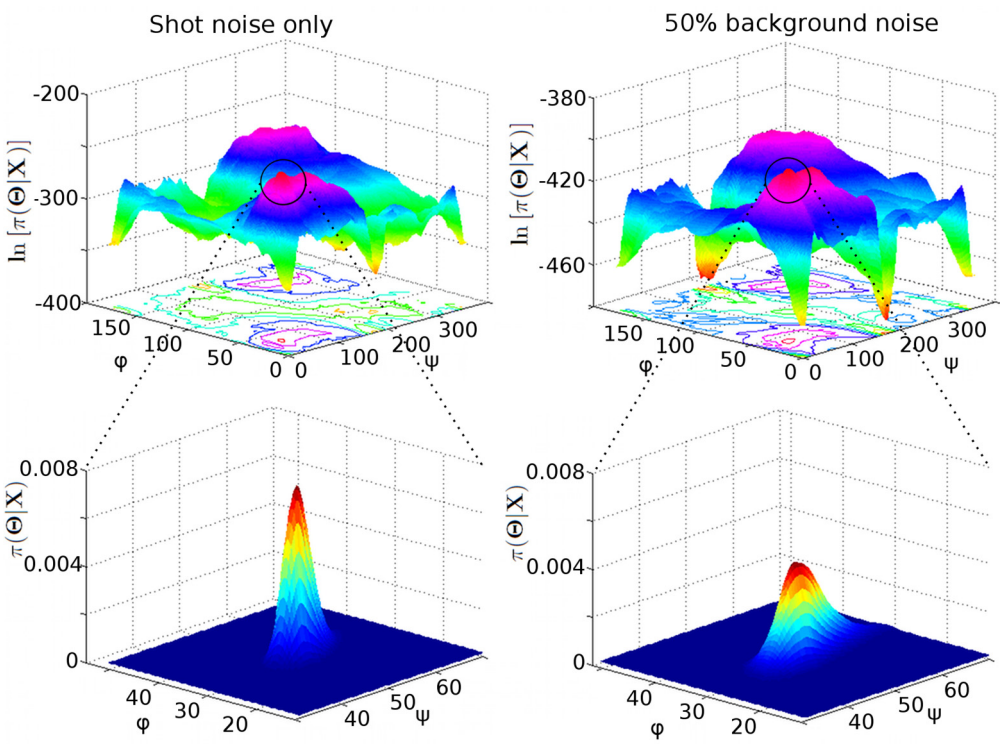

FIG. 2. (Color online) Examples of posterior probability surfaces $\pi(\mathbf{\Theta}|\mathbf{X})$. Shown are cuts through the three-dimensional angular probability landscapes obtained for a glutathione molecule with actual orientation $\theta = 73^\circ, \psi = 52^\circ, \varphi = 34^\circ$ using diffraction images with shot noise only (left) and with additional 50% background noise (right). Slices at the Θ value corresponding to the posterior maximum ($\theta_{\text{max}} = 75^\circ$ for shot noise only, $\theta_{\text{max}} = 63^\circ$ for background noise) are depicted at a logarithmic scale (top row). The insets (bottom row) show the maximum peaks at a linear scale.

seen in the logarithmic plot (top row). The zoom below, using a linear probability scale, shows that this peak is indeed quite pronounced. Its maximum is shifted, due to the shot noise, by ca. 2.2° ($\theta = 75°, \psi = 52°, \varphi = 33°$) with respect to the actual orientation ($\theta = 73°, \psi = 52°, \varphi = 34°$), which is well within the half width of the peak of about 3.2°. Given the fact that only 65 scattered photons have been used, the obtained accuracy is remarkable.

Including additional 50% background noise (e.g., nonelastically scattered photons) slightly changes the posterior probability landscape (right two plots in Fig. 2). In particular, the maximum shifts to $\theta = 63°, \psi = 50°, \varphi = 38°$ and is somewhat broader (half width of 4.1°) than for shot noise only.

B. Orientation determination and electron-density calculation

As the detector plane corresponds to a fragment of an Ewald sphere, whose position in reciprocal space in turn depends on the orientation of the molecule, we calculated the 3D molecular transform from the obtained diffraction patterns by appropriately mapping the detector plane onto the respective section of the Ewald sphere, as determined from the obtained orientation. By accumulating the photons from the Ewald sphere into corresponding 3D voxels of a Cartesian grid, an average 3D reciprocal space density is obtained.

For orientation determination, we compared two approaches, which we will refer to as "maximum likelihood" and "Bayesian," respectively. The maximum likelihood method uses the position of the maximum of the posterior probability distribution, computed from Eq. (1), as a point estimate of the orientation that most likely gives rise to the observed diffraction pattern. The Bayesian approach, in contrast, uses the entire posterior probability distribution as a weighting function, and thus each orientation is represented with an appropriate weight. Accordingly, the Bayesian method is expected to be less sensitive to information loss due to insufficient sampling as well as to the necessary discretization of reciprocal space.

Figure 3 compares cuts through the obtained 3D molecular transform along the k_x axis. The plots in the upper row serve to compare the two above approaches; the bottom row shows the influence of the background noise on the accuracy of the Bayesian approach. Below each of the four profiles, the difference between the reference and calculated molecular transform is shown. The molecular transform was calculated from 20 000 simulated diffraction images, containing on average 82 elastically scattered photons. Additional 10% and 50% photons, relative to the mean scattered photon count, were considered to simulate different background noise levels.

As can be seen, both the maximum likelihood as well as the Bayesian approach allow us to determine the molecular

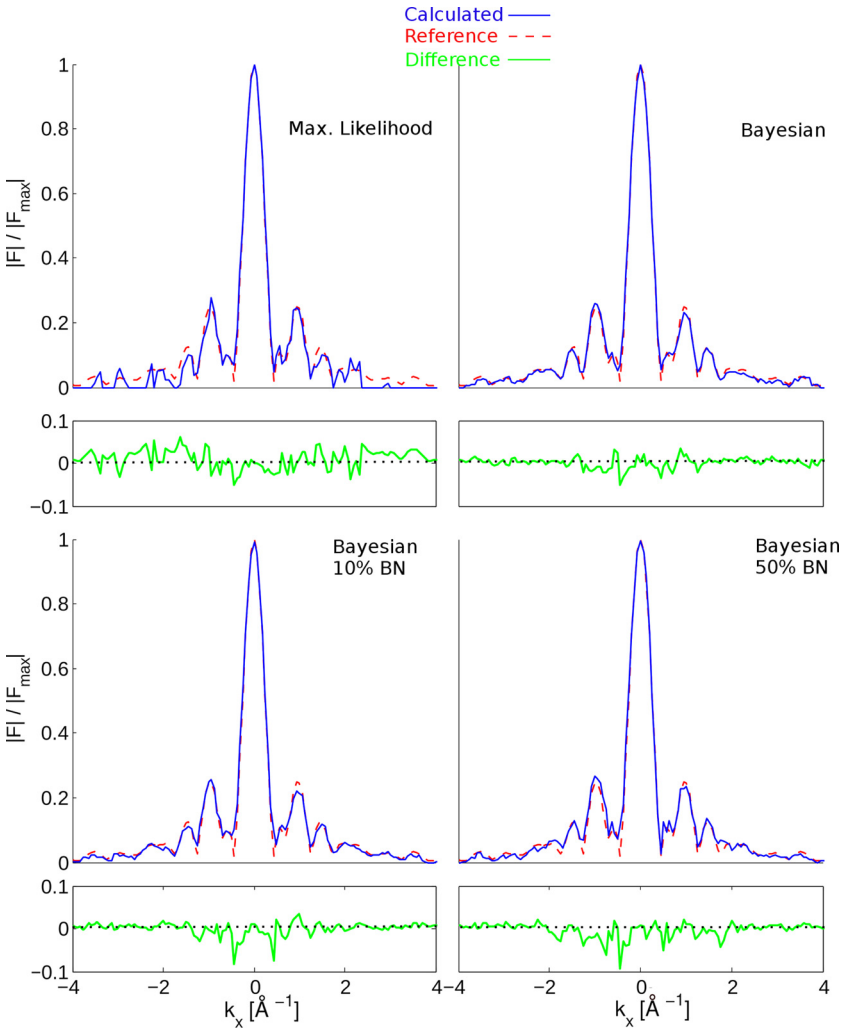

FIG. 3. (Color online) Quality of determined molecular transforms. Shown are cuts along the k_x axis of the calculated molecular transform (blue solid lines) compared to the reference (red dashed). Below each panel, the difference [green (light gray)] between the calculated and the reference molecular transform is shown. The top row shows differences in the high-resolution regime between the two tested methods, the maximum likelihood and Bayesian; the bottom row shows the effect of different levels of background noise for the Bayesian method.

TABLE I. *R* factors obtained for the maximum likelihood and Bayesian structure determination methods at two different noise levels: shot noise only (SN) and with additional 50% background noise (BN). The upper three rows contain *R* factors showing the accuracy of the determined molecular transforms. In the lower three rows, *R* factors reflect the resemblance in reciprocal space of the retrieved electron densities to the reference densities. *R* factors were computed using intensity values within a sphere with a 4.4 Å^{-1} radius (last column).

	Method	Noise level	*R* factor ($\|\Delta\mathbf{k}\| \leqslant 4.4$ Å^{-1})
Molecular transform determination	Max. likelihood	SN	0.48
	Bayesian	SN	0.21
	Bayesian	50% BN	0.23
Electron-density determination	Max. likelihood	SN	0.54
	Bayesian	SN	0.27
	Bayesian	50% BN	0.28

transform quite accurately, particularly in the small wave-vector regime. In the large k-vector regime, the Bayesian approach is clearly superior and captures details that are missed by the maximum likelihood approach. Apparently, this improvement is due to the additional information contained within the posterior probability distribution, which also allows for a better coverage of reciprocal space. The enhanced accuracy of the Bayesian approach is also reflected in the respective *R* factors (upper three rows of Table I).

Background noise (bottom row in Fig. 3) in the diffraction images was modeled by adding Gaussian-distributed random photon positions to the intensity distribution model in Eq. (1) used for the posterior probability calculations. After averaging the diffraction images in 3D k space, the background noise was subtracted from the calculated molecular transform. This approach enabled us to retrieve accurate profiles for background noise levels of 10% and 50%, respectively. As is also reflected in the calculated *R* factors (the second and the third row of Table I), no significantly deteriorated density compared to the shot noise only case is observed.

To test if the above robustness is also reflected in the quality and the level of resolved detail of the calculated electron-density map in real space, we calculated and compared the electron density from the respective molecular transforms using relaxed averaged alternating reflections algorithm [16] (Fig. 4). Indeed, as can be seen by comparing the middle row with the top reference density, the maximum likelihood approach (left) yields a less accurate electron density in the shot noise only case than the Bayesian approach. Corresponding *R*-factor values are listed in the fourth and fifth rows of Table I.

The loss of detail was expected from the respective lack of high-resolution information in reciprocal space. The bottom row in Fig. 4 demonstrates the effect of the background noise on the quality of the calculated electron density using the Bayesian approach. No significant difference is seen between the maps restored from shot noise only images and those with additional 10% or even 50% background noise. A similar quality of the structures derived from images with 50% background noise and with shot noise only is also reflected in the *R*-factor values (bottom two rows of Table I). These results allow us to conclude that the Bayesian approach is robust against substantial noise levels and, unlike the maximum likelihood approach, does not suffer from high-resolution structural information loss.

C. Resolution dependence on molecular mass

Having established that even the structure of relatively small biomolecules can be solved from quite sparse single-molecule diffraction data, we next asked how, under these conditions, the expected resolution scales with molecular mass over a larger range, given different beam intensities and background noise levels.

To this aim, we estimated the spatial resolution Δx via the product of angular resolution $\Delta\Theta$, i.e., the estimated orientation uncertainty, and radius of gyration R_g of the irradiated molecule. The angular resolution $\Delta\Theta$, in turn, was estimated via the respective posterior probability distribution as the mean distance to the correct orientation. Distances between two orientations were computed using Riemannian metrics [30].

Two opposing effects are expected. On the one hand, according to the law of large numbers, the accuracy $\Delta\Theta$ of the orientation estimate should increase with the number N_{phot} of recorded photons, $\Delta\Theta \propto N_{\text{phot}}^{-1/2}$ [31]. This can be seen, e.g., by considering a diffraction image with N_{phot} of recorded photons that gives rise to a well-pronounced maximum at $\boldsymbol{\Theta}_{\max}$ in the

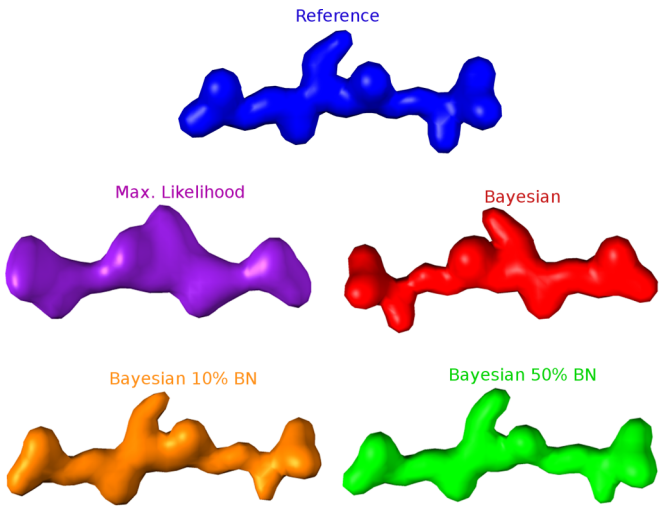

FIG. 4. (Color online) Quality of calculated electron densities. Top: Reference density; middle: densities obtained by maximum likelihood (left) and Bayesian (right) methods. The bottom row demonstrates the robustness of the Bayesian method to the background noise (BN) levels of 10% (left) and 50% (right).

likelihood distribution and in the resulting posterior probability landscape. We further assume that this diffraction image is a superposition of m independent photon positions subsets, each of them containing $N_{\mathrm{sub}} = N_{\mathrm{phot}}/m$ photons that are drawn from the same likelihood distribution. By Taylor expansion of the logarithm of the likelihood for the superimposed image around $\mathbf{\Theta}_{\max}$ it can be seen that the width of the maximum scales with $m^{-1/2}$. Because N_{phot} is proportional to the molecular mass M and to the beam intensity I_0, we expect $\Delta\Theta \propto (I_0 M)^{-1/2}$. On the other hand, the achievable spatial resolution $\Delta x \propto R_g \Delta\Theta$ for given orientational accuracy decreases with the size of the molecule, given, e.g., by the radius of gyration $R_g \propto M^{1/3}$. Combining these two effects yields the somewhat counterintuitive result $\Delta x \propto I_0^{-1/2} M^{-1/6}$, i.e., the achievable spatial resolution *increases* with molecular mass or size. The above assumption that it is the *small* molecules which represent the most challenging test cases, rests on this scaling argument.

To assess the validity of the expected scaling, we repeated the above synthetic glutathione scattering experiments with varying beam intensity and, hence, varying average number of scattered photons ranging from $N_{\mathrm{phot}} = 24$ to 3724. For each synthetic experiment, 500 diffraction images were generated and used to determine the average orientational error $\Delta\Theta$. In Fig. 5, the intersection of the vertical dashed line labeled "GTT" with the colored curves shows the achieved resolutions (vertical axis), which provide an estimate for the expected resolution for molecular mass $M = 307$ Da.

Further, assuming $N_{\mathrm{phot}} \propto I_0 M$, these data can be used to estimate the expected resolution for *any* molecular mass by considering a "scaled" glutathione molecule of α-fold size and mass $\alpha^3 M$ (horizontal axis in Fig. 5). For large photon numbers (>200), the obtained colored curves for different beam intensities I_0 indeed show the expected resolution increase $\propto M^{-1/6}$ and also $\propto I_0^{-1/2}$. For small photon counts, a stronger variation of resolution with molecular mass is seen because the $m^{-1/2}$ scaling of the orientation determination error breaks down. Specifically, in addition to orientations close to the correct one, pronounced posterior probabilities are also obtained for misaligned orientations (by typically 180°). For very few photons (left ends of the curves), the achievable resolution saturates at a maximal average orientational error of 90° for that reason.

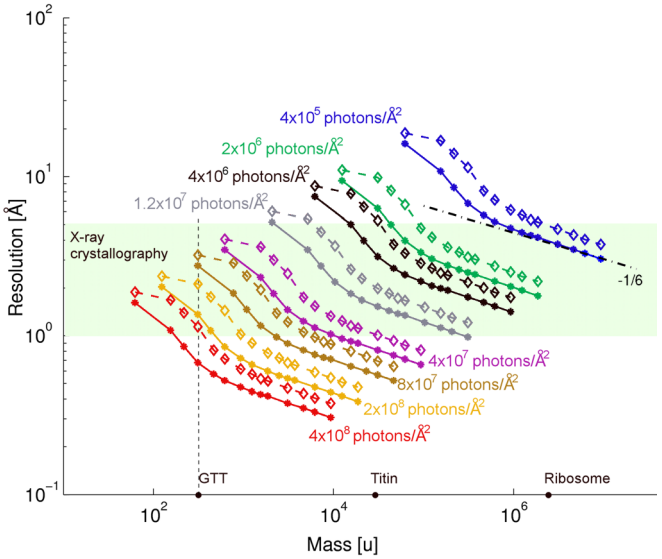

FIG. 5. (Color online) Achievable resolution for different molecular masses and incident beam intensities. Solid lines (dots) represent a signal with shot noise only and dashed lines (diamonds) a signal with additional 50% background noise. Line colors encode incident beam intensities (photons/Å^2). The resolution achievable with x-ray crystallography (~1–5 Å) is highlighted in green (light gray). The molecular masses of the test molecules used in this study are indicated on the x axis; glutathione (GTT), titin, and ribosome. The dot-dash line shows the expected slope of $-1/6$.

The black lines show the achievable spatial resolutions for an assumed beam intensity of $I_0 = 4.0\times10^6$ photons/Å^2 for 12-keV photons focused to a 100-nm spot [6]. Currently available beams offer comparable intensities (approximately 10^5 photons/Å^2 photons in a 1-μm focal spot), however, for energies up to 2 keV [9]. Remarkably, already this assumed intensity level would suffice to resolve structures such as three Ig domains of a titin molecule or a ribosome with a resolution similar to that typically achieved by x-ray crystallography [green (light gray) region]. Smaller molecules require higher intensities; e.g., to achieve atomic resolution for glutathione (GTT), a 50-fold intensity would be required, corresponding to a 10-nm focal spot, which also seems within reach for 6-keV pulses [32].

D. Structure optimization

So far we have shown that the first Bayesian approach is capable of accurately determining the molecular orientation for each single diffraction image, given a "seed structure." We will now introduce the second approach that considers the fit to the whole registered set of diffraction images rather than to each single diffraction image. We will then embed this approach within a refinement procedure, in which the "seed structure" is iteratively optimized in real space. This approach will be particularly useful when studying conformational motions by XFEL single-molecule scattering, in which case an already-known conformation can serve as the seed structure. To that end, the expression for posterior probability needs to be modified such as to calculate the probability of a structure, given a complete set of diffraction patterns. This probability will then serve to guide structure optimization in a Monte Carlo scheme.

In Eq. (1) the structure was assumed to be known; now it will be treated as an unknown parameter to be determined form the posterior probability distribution. The likelihood of observing the diffraction pattern $\mathbf{X}_i = \{(x_i^{(l)}, y_i^{(l)})\}_{l=1,\ldots,n_i}$ given the structure $S_j = \{\mathbf{r}_1^{(j)}, \ldots, \mathbf{r}_N^{(j)}\}$, defined as a set of N atomic positions, and the orientation $\mathbf{\Theta}_i^{(j)} = (\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$ of the j-th structure for diffraction pattern i is

$$f\left(\mathbf{X}_i | S_j, \mathbf{\Theta}_i^{(j)}\right) \propto \left[1 - \frac{A\left(\mathbf{\Theta}_i^{(j)}, S_j\right)}{N_{\mathrm{total}}}\right]^{N_{\mathrm{total}} - n_i} \times \prod_{l=1}^{n_i} I\left[R\left(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)}\right)\Delta\mathbf{k}\left(x_i^{(l)}, y_i^{(l)}\right), S_j\right], \quad (2)$$

where $I(\Delta\mathbf{k},S_j)$ is the intensity in a detector pixel corresponding to a scattering vector $\Delta\mathbf{k}$ and structure S_j, $R(\theta_i^{(j)},\psi_i^{(j)},\varphi_i^{(j)})$ is a rotation matrix corresponding to the orientation $\mathbf{\Theta}_i^{(j)}=(\theta_i^{(j)},\psi_i^{(j)},\varphi_i^{(j)})$, $A(\mathbf{\Theta}_i^{(j)},S_j)=\sum_{l=1}^{N_{\text{pixel}}} I[R(\theta_i^{(j)},\psi_i^{(j)},\varphi_i^{(j)})\Delta\mathbf{k}(x^{(l)},y^{(l)}),S_j]$ is the expected amount of elastic scattering for the orientation $\mathbf{\Theta}_i^{(j)}$ of structure S_j registered by a detector with N_{pixel} pixels, and N_{total} is the total number of incident photons. Because all probabilities $f(\mathbf{X}_i|S_j,\mathbf{\Theta}_i^{(j)})$ are independent, the total probability of obtaining a whole set of diffraction patterns $\{\mathbf{X}_i\}$ is given by the product

$$f\left(\{\mathbf{X}_i\}|S_j,\left\{\mathbf{\Theta}_i^{(j)}\right\}\right)=\prod_i f\left(\mathbf{X}_i|S_j,\mathbf{\Theta}_i^{(j)}\right). \tag{3}$$

Assuming a uniform distribution of structural coordinates, Bayes's theorem yields the posterior probability

$$\pi\left(S_j,\left\{\mathbf{\Theta}_i^{(j)}\right\}|\{\mathbf{X}_i\}\right)\propto\prod_i f\left(\mathbf{X}_i|S_j,\mathbf{\Theta}_i^{(j)}\right). \tag{4}$$

By integrating this expression with respect to $\mathbf{\Theta}_i^{(j)}$, one obtains the posterior probability distribution of structure S_j,

$$\pi(S_j|\{\mathbf{X}_i\})\propto\prod_i\iiint f\left(\mathbf{X}_i|S_j,\theta_i^{(j)},\psi_i^{(j)},\varphi_i^{(j)}\right)\times\sin\theta_i^{(j)}d\theta_i^{(j)}d\psi_i^{(j)}d\varphi_i^{(j)}, \tag{5}$$

which will serve to assess how well structure S_j, during refinement, fits to the recorded set of diffraction patterns.

As a proof of principle, this approach is demonstrated for the above glutathione structure. Starting from a randomly chosen structure (i.e., randomly chosen dihedral angles), each dihedral angle is changed by a Gauss-distributed random angle. For each MC step, the posterior probability of the new structure, $\pi_j=\pi(S_j|\{\mathbf{X}_i\})$, was calculated from Eq. (5) and compared to that of the structure obtained in the previously accepted step. As an acceptance criterion, the Metropolis criterion [33] was used, with energies $E_j=-k_BT\ln\pi_j$. Accordingly, the new structure was accepted only if $\xi<\exp(-\Delta E/k_BT)=\pi_j/\pi_{j-1}$, where ξ is a random number between [0,1).

For the glutathione tripeptide, only 200 synthetic diffraction images were used, with a mean photon count of ca. 76 photons per picture, assuming an incident beam intensity of 2×10^8 photons/Å^2. No background noise was included. Figure 6 shows the optimization and convergence of the (normalized) posterior probability $\pi(S|\{\mathbf{X}\})$ for 12 random starting structures during subsequent MC runs. Two sample random starting structures are shown in the pink and green boxes. As can be seen, the reference structure, from which the synthetic diffraction images were calculated, was approached already after few hundred accepted MC steps, and, after about 1000 accepted steps, a typical root-mean-square deviation (RMSD) of 0.02 Å between the most probable structure [lower right, orange (dark gray)] to the target structure [yellow (light gray)] was reached.

In contrast to the small peptide glutathione, *de novo* MC refinement of proteins is currently complicated by the huge conformational space that has to be sampled. We therefore

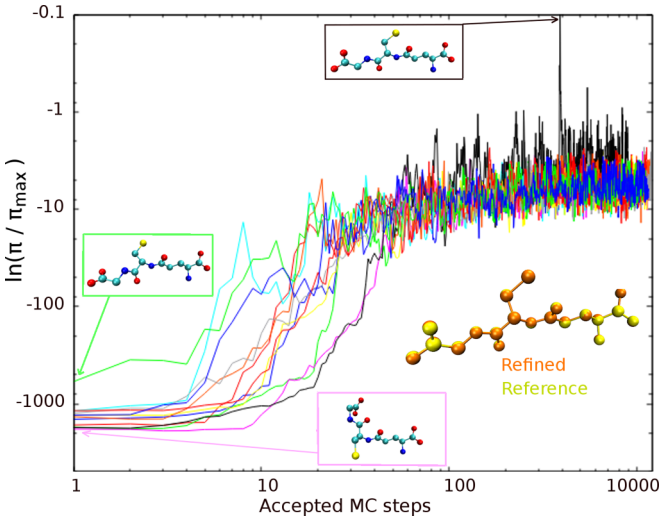

FIG. 6. (Color online) Monte Carlo structure refinement of glutathione. The logarithm of normalized probability that the accepted structure simultaneously agrees with 200 synthetic diffraction images containing 76 photons was plotted for 12 independent MC runs (color lines) starting from different random structures. Two examples of starting structures are shown in green (mid-left) and pink (bottom left) boxes. After about 500 steps, the most probable structure is found. A comparison of the refined [orange (dark gray)], i.e., the most probable, and the reference structure [yellow (light gray)] is shown in the bottom right corner.

asked if, for larger molecules, our method is capable of distinguishing between correct and incorrect conformations. To this aim, we considered a 283 residues titin construct consisting of three Ig domains (Ig67–Ig69) that are internally rigid but loosely coupled to each other by flexible PEVK linkers (PDB entry 2RIK [34]).

A set of conformations that differ in their mutual domain orientations was generated from a 2.81-ns molecular dynamics (MD) simulation of the titin molecule in vacuum with intradomain distance restrains, leaving all linkers flexible. The structure at 2800 ps was chosen as the reference, from which a set of 200 diffraction images was generated with an average of 376 photons per picture, assuming an incident beam intensity of 4×10^6 photons/Å^2. Only shot noise was considered, i.e., no background noise was included. For each of the proposed structures the posterior probability $\pi_j=\pi(S_j|\{\mathbf{X}_i\})$ and RMSD to the reference structure were calculated. For each of the 290 sampled conformations, Fig. 7 shows the respective posterior probability (symbols) as a function of its RMSD from the target structure. As an illustration of the structural differences, the target structure [Fig. 7(a)] is shown in blue, along with three sample structures [Figs. 7(b), 7(c), and 7(d)].

Indeed, the largest probability is obtained for the target structure, and any deviation from this structure results in a lower probability. This is true even for an RMSD as small as 0.6 Å (note the large log-scale), which suggests that the target structure would be correctly identified among all 290 trial structures from the diffraction data to better than 0.6 Å RMSD.

According to the estimated dependency of achievable resolution on molecular mass (Fig. 5), our method should also work for very large biomolecular complexes. To test the limits

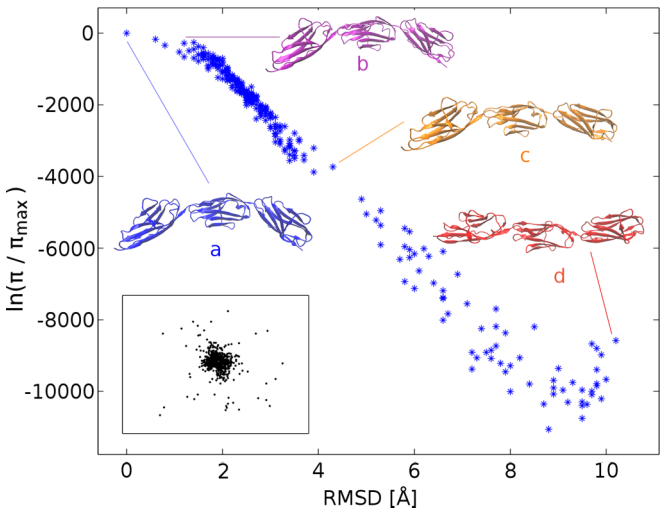

FIG. 7. (Color online) Ability to identify the correct titin structure. Two hundred synthetic images with 376 photons per picture (inset: sample image) were generated for the reference structure (a). For 290 different conformations, the normalized posterior probability (blue symbols) is plotted as a function of the RMSD with respect to the reference structure. Insets show three intermediate structures, (b), (c), and (d).

of the Bayesian approach, we have chosen near-atomistic structural models of the ribosome (molecular mass: about 2.5 MDa) as a test case, also because of its high internal dynamics at multiple length scales during the translation cycle, which has been characterized recently [35,36]. Seven structures at different stages during translation were used for posterior probability calculations, kindly provided by Ref. [36]. The pre1 state was chosen as the reference structure, from which 200 diffraction images with on average 1.075×10^5 photons per picture were generated, assuming an incident beam intensity of 4×10^6 photons/Å^2. Of particular functional importance are the pronounced structural changes of the bound tRNA during its translocation. We have therefore studied the ability of our approach to distinguish among different structures at three different, increasingly challenging, levels.

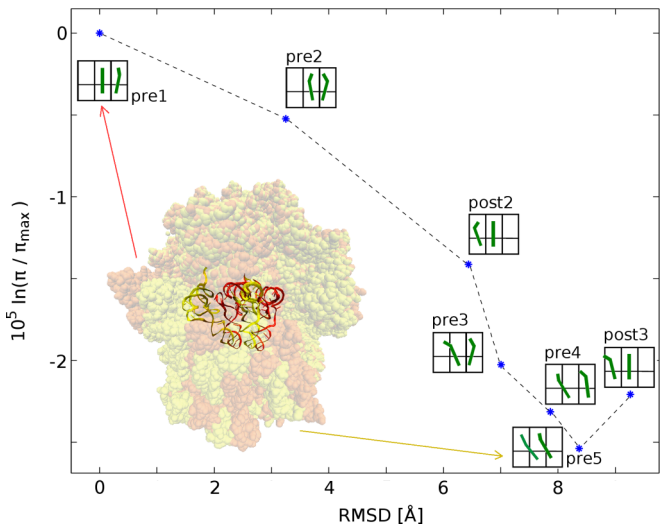

FIG. 8. (Color online) Ability to identify the correct ribosomal translocation state. Two hundred synthetic diffraction images with 1.075×10^5 photons per picture were generated from the reference structure (pre1 state). For seven different translocation states, the logarithm of the normalized posterior probability and the RMSD with respect to the reference (pre1 state) was calculated. The inset shows an overlay of the pre1 [red (dark gray)] and the pre5 [yellow (light gray)] states, in particular, structural differences within subunits (surface) and the tRNA chains (cartoon) are highlighted. Boxes next to each of the points depict the location of the tRNA chains in the ribosome at corresponding translocation states.

First, differences involving the complete ribosome structure were considered. Figure 8 shows the posterior probability versus the RMSD for all seven states (symbols). The schematic representations at each point indicate the position of the tRNA chains at the three binding sites A, P, and E (as defined in, e.g., Ref. [35]) in those states. An overlay of pre1 [red (dark gray)] and pre5 [yellow (light gray)] structures shows the extent of structural differences caused by both rotation of subunits (represented as a surface) and displacement of tRNA chains (represented as a cartoon). As expected, the reference structure yields the largest probability and can therefore be reliably identified by an enormous probability ratio against the other six structures. Overall, structural deviations from the reference (expressed in terms of RMSD) result in lower posterior probabilities. The rightmost structure is an exception, presumably due to a partial compensation between subunits rotation and tRNA translocation.

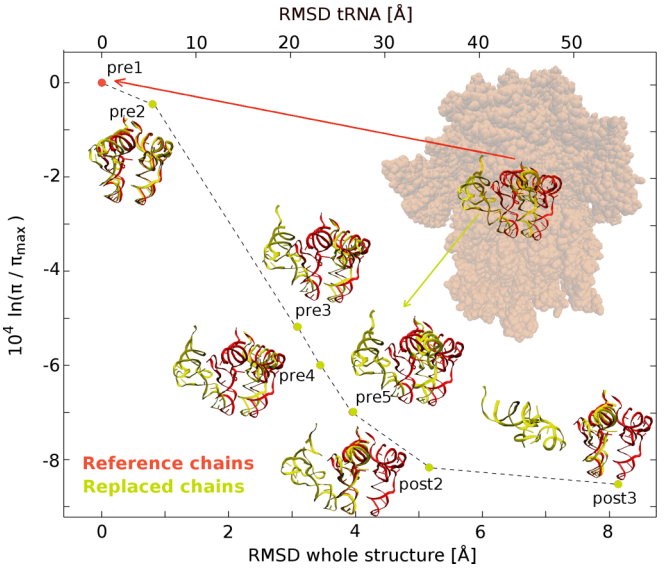

FIG. 9. (Color online) Tracing local structural changes during tRNA translocation. For each translocation state model, obtained by embedding the native tRNA chains conformation within the subunit structure of the pre1 state, normalized posterior probability was calculated (y axis) using 200 synthetic diffraction images with 1.075×10^5 photons per picture generated from the pre1 state. Bottom x axis corresponds to the RMSDs of whole structures, while the upper axis shows the RMSDs of the tRNA chains alone. The difference between the translocated [yellow (light gray)] and the reference chains [red (dark gray)] is shown for each of the states. The inset demonstrates the differences between the tRNA chains locations from the pre1 and the pre5 states, embedded in the ribosomal structure of the pre1 state (shaded surface).

Next, we focused at the tRNA, which in itself is only a very small part of the entire ribosome: Can our method also pinpoint structural changes of the tRNA only, against a large background of (unchanged) ribosome residues? To address this question, the tRNA chains from the seven different ribosome state structures were inserted into the pre1 ribosome structure, such that the only difference among the resulting set of seven structures is the position and the internal structure of the bound tRNA. Again, Fig. 9 shows a pronounced decrease of the posterior probability with increasing structural deviation, expressed in RMSD, of the tRNA from the reference structure. The lower x axis corresponds to the RMSD value of the entire complex with respect to the pre1 state, while the upper x axis marks the RMSD value of the unaligned tRNA chains alone. The actual difference between the reference chains [red (dark gray)] and the replaced ones [yellow (light gray)] is illustrated by the sample overlays of tRNAs for states other than the pre1. Apparently, and despite the large ribosome background that in this case does not yield any structural information, it was possible to detect quite local changes of the tRNA conformation along the translocation process. We therefore assume that, in general, ligand binding will be accessible to single-molecule x-ray diffraction.

In the third step, we finally asked if this sensitivity is retained even against the background of an inaccurate ribosome seed structure. To this aim, the seven tRNA configurations were embedded into the ribosome structure of the pre2 state, while the diffraction images were generated using the pre1 state. In a similar way as above, Fig. 10 shows the obtained posterior probabilities for the seven chimera structures. Again, the reference structure turns out to be the most probable one with respect to the set of recorded diffraction images and therefore would be readily identified. Still, a clear decrease of the structure probability with increasing displacement of the tRNA chains is seen, despite the inaccuracy of the seed model, albeit with a somewhat smaller but still significant ratio between the target structure and the second-best candidate $\ln(\pi_{\text{target}}/\pi_{\text{2nd}}) \approx 4.04\times10^3$.

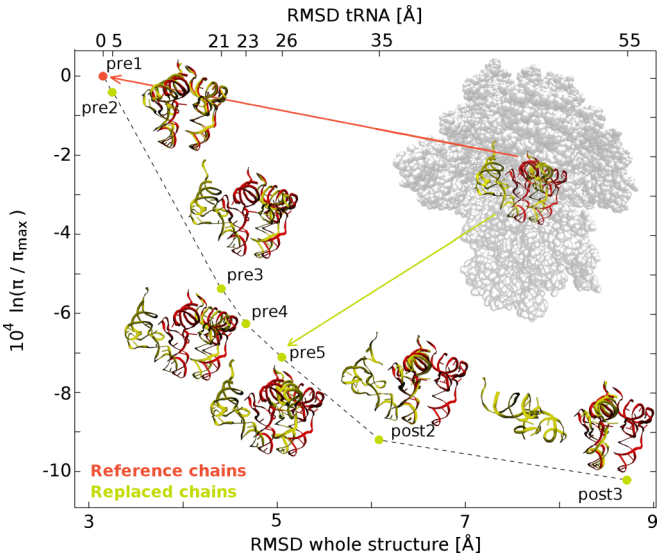

FIG. 10. (Color online) Tracing local structural changes during tRNA translocation using an inaccurate structure model. For each translocation state model, obtained by embedding the native tRNA chains conformation within the subunit structure of the pre2 state (inaccuracy of the model), normalized posterior probability was calculated (y axis) using 200 synthetic diffraction images with 1.075×10^5 photons per picture generated from the pre1 state. The bottom x axis corresponds to the RMSDs of whole structures, while the upper axis shows the RMSDs of the tRNA chains alone. The difference between the translocated [yellow (light gray)] and the reference chains [red (dark gray)] is shown for each of the states. The inset demonstrates the difference between the tRNA chains locations from the pre1 and the pre5 state embedded in the ribosomal structure of the pre2 state (shaded surface).

III. METHODS

A. Intensity distribution model and diffraction images

In ultra-short-pulse single-molecule XFEL experiments, both the intensity of the incident pulse and the electron density of the specimen are time dependent, the latter as a result of radiation damage by the incident beam. For unpolarized x-ray pulses, the registered intensity is given by

$$I(\Delta\mathbf{k}) = r_e^2 \frac{1+\cos^2 2\theta}{2}\Delta\Omega \times \int_{-\infty}^{\infty} I_0(t)\left|\iiint \rho(\mathbf{r},t)e^{2\pi i\Delta\mathbf{k}\cdot\mathbf{r}}dV\right|^2 dt, \quad (6)$$

where I_0 is the incident beam intensity, r_e is the classical electron radius, θ is the scattering angle, $\Delta\Omega$ is a solid angle subtended by a pixel on the detector plane, and $\rho(\mathbf{r},t)$ is the time depended electron density [6]. Assuming sufficiently short pulses with low temporal coherence, intensity distributions were computed as incoherently summed scattering amplitudes of a time-independent electron density. If required, Eq. (6) can be generalized to account for polarized pulses and partial coherence between time slices should be possible, which is, however, beyond the scope of this paper.

The electron density was modeled as a sum of Gaussian functions centered at positions of nonhydrogen atoms, with amplitudes given by the number of electrons and standard deviations equal to the atomic radius of particular elements. All intensity distributions were calculated for a 1 Å wavelength (photon energy of 12 keV) on a $200\times200\times200$ grid with a 6.3×10^{-2} Å^{-1} spacing for the glutathione, a $300\times300\times300$ grid with a 6.3×10^{-3} Å^{-1} spacing for the titin, and a $300\times300\times300$ grid with a 1.3×10^{-3} Å^{-1} spacing for the ribosome. A beam intensity I_0 of 2×10^8 photons/Å^2, a result of focusing approximately 1.57×10^{12} photons to a 10-nm diameter spot, was assumed for the glutathione and $I_0 = 4\times10^6$ photons/Å^2 (approx. 3.14×10^{12} photons in a 100 nm diameter spot) for the titin and the ribosome.

From the intensity $I(\Delta\mathbf{k})$, sample images with n_i photons were generated to test our method. To that aim, photon positions were chosen at random, following the intensity distribution defined in Eq. (6). For efficiency reasons, here the stochastic fluctuations in registered photon counts were approximated by a Poisson distribution,

$$p(n,\Delta\mathbf{k}) = \frac{[I(\Delta\mathbf{k})]^n}{n!}e^{-I(\Delta\mathbf{k})}, \quad (7)$$

where n is the number of photons registered at a pixel to which $\Delta\mathbf{k}$ points to; photons at $\Delta\mathbf{k}=0$ were used in the orientation determination approach but not for calculating the structure probability in the second approach.

From all recorded photons, only those elastically scattered by the target molecule contribute to the signal, others form background noise. To mimic experiments, this noise was described by adding normal distributed photons to the diffraction images defined above. The width of the distribution, 1/10 of the detector size, was chosen such that the background noise mainly affected the center of the images, as in recent experiments [37]; a small amount of background noise within the high-resolution regions was tolerated. Accordingly, for the calculation of posterior probabilities, an appropriate Gaussian function was included within the right side of Eq. (1). Assuming a reduced number of inelastically scattered photons due to energy filtering, 10% and 50% ratios of background noise photons to the mean signal photons per picture were considered.

For the glutathione we assumed a 121×121 pixel (6 cm×6 cm) detector, a 241×241 pixel (1.2 cm×1.2 cm) one for the titin, and 241×241 pixel (2.4 mm×2.4 mm) one for the ribosome, in each case placed 10 cm from the molecule.

B. Random orientation generation

Single molecules entering the XFEL beam were assumed to be oriented randomly, following a uniform distribution. To generate uniformly distributed orientations, Euler angles were sampled from a probability density given by $g(\theta,\psi,\varphi)=(8\pi)^{-1}\sin\theta$ [38], i.e., $\psi\in I[0,2\pi)$, $\varphi\in I[0,\pi)$, and $\theta=\arccos z$, where $z\in I[-1,1]$.

We used the Gnu Scientific Library [39] implementation of the "Mersenne twister" algorithm [40] as the pseudo-random-number generator for simulating the diffraction images.

C. Computing posterior distributions

To obtain the posterior probability distribution in both the maximum likelihood and Bayesian method, for a diffraction pattern the probability $\pi(\mathbf{\Theta}_i|\mathbf{X}_i)$ was calculated from Eq. (1) for orientations sampled on a grid. Given a particular orientation, the intensity distribution on the detector plane $I_{\mathbf{\Theta}}[\Delta\mathbf{k}(x,y)]$ was computed by projecting the corresponding Ewald sphere onto the detector plane. Intensity values on the Ewald sphere were trilinearly interpolated from the 3D grid representation of the molecular transform calculated beforehand from the model structure. All posterior probability values were expressed as logarithms to avoid numerical underflows, and the results were exponentiated when necessary.

To obtain sufficient orientational resolution at affordable computational cost, relevant regions with large posterior probability were sampled at increased resolution. To that aim, after all probability maxima were located on a coarse grid, the relevant regions around these maxima were subsequently fine sampled. Euler angles $\theta=(0,\pi)$, $\psi=[0,2\pi)$, and $\phi=[0,\pi)$ were discretized on the coarse grid with a 10° step. The subsequent subsampling was performed with a 2° step, with the relevant regions defined by the ratio of the fine sampled probability to the maximum of coarse sampled probability, exceeding a given threshold $\pi^{\text{fine}}(\mathbf{\Theta}_i|\mathbf{X}_i)/\pi^{\text{coarse}}_{\max}\geqslant10^{-3}$ for the glutathione and $\pi^{\text{fine}}(\mathbf{\Theta}_i|\mathbf{X}_i)/\pi^{\text{coarse}}_{\max}\geqslant5\times10^{-4}$ for the titin and the ribosome.

In the maximum likelihood approach, the position of the fine sampled maximum was used as the orientation estimate, whereas in the Bayesian method all probability values above the threshold were used as a weighting function $W_i^{\text{fine}}(\mathbf{\Theta}_i)=\pi^{\text{fine}}(\mathbf{\Theta}_i|\mathbf{X}_i)/\pi^{\text{fine}}_{\max}$.

All posterior probability distributions for estimating the accuracy $\Delta\Theta$ of orientation determination were calculated using a 1° step. Diffraction images were generated from the glutathione molecule rotated with respect to the reference by $\theta=58°$, $\psi=74°$, and $\varphi=136°$.

To calculate the posterior probability of a structure given a set of diffraction pictures $\pi(S_j|\{\mathbf{X}_i\})$, the likelihood for each picture in the set $f(\mathbf{X}_i|S_j,\theta_i^{(j)},\psi_i^{(j)},\varphi_i^{(j)})$ was integrated over all orientations using the rectangle rule. Source code will be available at http://www.mpibpc.mpg.de/9660732/28-SM-Ultrafast-XRay-Diffraction.

D. Electron-density determination

To calculate the electron density in real space from the molecular transform, the relaxed averaged alternating reflections algorithm (RAAR) [16] was used. Before applying the algorithm to the calculated 3D molecular transform, a random phase was assigned to each amplitude $|F(\Delta\mathbf{k})|=\sqrt{I(\Delta\mathbf{k})}$. The amplitude $|F(0)|$ was included in the calculations.

The finite support in the real space was defined as a cube with an edge twice the length of the radius of gyration (for the glutathione $R_g=4.5$ Å) centered at the origin and was kept constant. Phases were retrieved in 300 iterations of the RAAR algorithm. The β parameter was updated according to a smooth approximation of a step function from the initial value $\beta_0=0.75$ to the final value $\beta_{\max}=0.99$ centered at the seventh iteration [16]. We used the fast Fourier transform implementation from the FFTW library [41] to project the data from the real to reciprocal space and vice versa.

The quality of determined molecular transforms and retrieved electron densities was assessed in terms of R factors,

$$R=\frac{\sum\big\|F_{\text{ref}}(\Delta\mathbf{k})|-|F_{\text{det}}(\Delta\mathbf{k})\big\|}{\sum|F_{\text{ref}}(\Delta\mathbf{k})|},\tag{8}$$

where $|F_{\text{ref}}(\Delta\mathbf{k})|$ is the amplitude of the Fourier transformed reference electron density and $|F_{\text{det}}(\Delta\mathbf{k})|$ is the result of the performed structure determination. All R factors were computed within a 0.22 Å resolution sphere ($|\Delta\mathbf{k}|\leqslant4.4$ Å^{-1}).

E. MC simulation for *de novo* structure determination

To generate random structures of the reference glutathione tripeptide, we changed the dihedral angles in the glycine and cysteine residues. Starting structures for the MC were generated by assigning random dihedral angles from a uniform distribution. At each MC step, new dihedral angles were obtained by varying previous angles using a normal-distributed random step. For each MC run, an initial average step size of 10° was used. The step size was halved when the acceptance ratio dropped below 0.2 and doubled when this threshold was exceeded.

To prevent the simulation from being trapped in a local minimum of the energy landscape, simulated annealing [42] was applied. By introducing a dimensionless temperature ratio $T_\mathrm{r} = T/T_\mathrm{a}$, the Metropolis criterion reads

$$\xi < \exp\left[\frac{(\ln \pi_j - \ln \pi_{j-1})T}{T_\mathrm{a}}\right] = \left(\frac{\pi_j}{\pi_{j-1}}\right)^{T_\mathrm{r}}, \qquad (9)$$

where T_a is the annealing temperature and T is a pseudotemperature ensuring nondimensionality of the argument of the exponent function. For annealing, the temperature ratio was changed at each accepted MC step according to $T_\mathrm{r}(n) = T_\mathrm{r}^\mathrm{f} + (T_\mathrm{r}^0 - T_\mathrm{r}^\mathrm{f})\mathrm{e}^{-n\tau}$, where n denotes the number of previously accepted MC steps, $T_\mathrm{r}^0 = 0.002$ is the starting temperature ratio, $T_\mathrm{r}^\mathrm{f} = 1.2$ is the final temperature ratio, and $\tau = 0.005$ is a time constant. Values of these parameters were chosen heuristically.

F. MD simulation of the titin molecule in vacuum

To create a set of structures, the simulation was carried out using the GROMACS 4.5 simulation package [43] with the OPLS-AA force field [44]. Long-range electrostatic interactions (beyond a cutoff radius of 1.0 nm) were computed with the particle mesh Ewald method [45]. Lennard-Jones interactions were calculated up to a cutoff of 1.4 nm. The temperature of the protein was kept at 300 K by coupling it to a temperature bath using the velocity rescale algorithm [46] with a time constant of 0.2 ps. All bonds were constrained with the LINCS algorithm [47], additional distance restrains were put on atoms within same Ig domains. An integration time step of 2 fs was used. The total length of the trajectory was 2.81 ns, and snapshots were taken every 100 ps. During the last 10 ps of the simulation, snapshots were taken every 1 ps to obtain conformations with small structural changes from the reference structure and thus to also fill the small RMSD gap in Fig. 7.

IV. SUMMARY AND CONCLUSION

We have discussed and assessed two Bayesian approaches for structure determination and discrimination from sparse and noisy single-molecule x-ray diffraction data. The first approach requires a "seed model," which serves to determine the orientation of the irradiated molecule for each of the collected images separately. For each obtained orientation, the registered photons are mapped onto the appropriate Ewald sphere, and the molecular transform is accumulated from many diffraction images in 3D reciprocal space. Two orientation determination variants were compared. The maximum likelihood approach uses the position of the maximum of the posterior distribution function as the orientation estimate, whereas the Bayesian approach uses the complete information contained within the posterior probability to derive orientation estimates. As expected, the Bayesian approach yielded a more accurate sampling of 3D reciprocal space.

The structure of the posterior probability landscape depends both on the number of registered photons and on the level of background noise. As one should expect, with increasing number of photons, the maximum of the probability distribution becomes narrower, thus improving the accuracy of the orientation estimate. Inclusion of additional background noise photons broadens the probability distribution maximum. Probably contrary to first intuition, the obtained orientational accuracy as a function of average photon count per diffraction image showed that, in terms of achievable resolution for a given beam intensity, it is the small molecules that are most challenging. The spatial resolution improves with molecular mass to the power of $-\frac{1}{6}$, hence better resolution is expected for larger molecules both in the absence and presence of additional background noise. Currently available beam intensity of about 4×10^6 photons/Å^2, focused to a 100-nm focal spot, will not yield 80 photons per diffraction image, which were assumed in this work for the test tripeptide; however, such a photon count could be achieved with a 10-nm focal spot, which seems not unrealistic within the near future. Determining larger biomolecular structures using presented methods should already be within reach with the present advances at the Stanford Linear Accelerator Center.

In the second approach, instead of assigning an orientation to *each* of the diffraction patterns separately, the probability that a particular structure gives rise to *all* recorded images is calculated. As a result, it is possible to identify a structure that simultaneously fits best to all collected diffraction images. Further, because the structures are defined in real space, no phase retrieval is required. Using the posterior probability of a structure given a set of diffraction images as a Monte Carlo acceptance criterion, *de novo* structure determination of a tripeptide was demonstrated. Structure refinement of longer polypeptide chains or proteins will be challenging due to the severe sampling problem. Several strategies to address this problem have, e.g., been developed for structure determination from small angle scattering x-ray experiments (SAXS) [48], which should be applicable to the case at hand. The size and resolution limits, to which these methods extend, however, remain to be established.

For the three Ig domain titin construct as well as for the ribosome, our approach proved to be capable of discriminating between slightly different candidate conformations, even for RMSD values as small as 0.6 Å. Additionally, in the case of the ribosome, very local small structural changes were correctly identified against a large structural background, thus demonstrating how, e.g., translocation of tRNA chains within a ribosomal complex could be tracked. Obtaining insight into specific regions of interest in complex biological systems should also be possible, even with slightly inaccurate model structures used for probability calculations. Taken together, these results strongly suggest that the structure and slight structural changes of even small molecules should be accessible through single-particle femtosecond XFEL diffraction experiments.

ACKNOWLEDGMENTS

We thank Russell Luke for helpful discussions and help with implementing his relaxed averaged alternating reflections algorithm and Ilme Schlichting, Tim Salditt, and Simone Techert for helpful discussions. We also thank Lars Bock, Christian Blau, and Andrea Vaiana for providing structural models of the ribosome. This work was supported by the Deutsche Forschungsgemeinschaft, Grant No. SFB 755/B4.

APPENDIX: DERIVATION OF EQ. (1)

To derive an expression for the likelihood $f(\mathbf{X}_i|\mathbf{\Theta}_i)$ of registering a configuration of n_i recorded photons positions $\mathbf{X}_i = \{(x_i^{(l)}, y_i^{(l)})\}_{l=1\ldots n_i}$ for a given molecular orientation $\mathbf{\Theta}_i$, we assume a uniform beam intensity I_0 over an area F_A, resulting in N_total incident photons. A diffraction pattern is formed by registering $n_1, n_2, \ldots, n_m$ photons in m detector pixels. The remaining $N_\mathrm{total} - \sum_{j=1}^{m} n_j$ photons are not registered. The probability of observing the diffraction image with $(n_1, n_2, \ldots, n_m)$ photon counts for all m pixels, and $N_\mathrm{total} - \sum_{j=1}^{m} n_j$ outside, follows a multinomial distribution,

$$f(n_1,n_2,\ldots,n_m) = \frac{N_\mathrm{total}!}{\left(N_\mathrm{total} - \sum_{j=1}^{m} n_j\right)! \prod_{j=1}^{m} n_j!} \left(1 - \frac{\sum_{j=1}^{m} I_j}{N_\mathrm{total}}\right)^{N_\mathrm{total} - n_j} \prod_{j=1}^{m} \left(\frac{I_j}{N_\mathrm{total}}\right)^{n_j} \propto \left(1 - \frac{\sum_{j=1}^{m} I_j}{N_\mathrm{total}}\right)^{N_\mathrm{total} - n_j} \prod_{j=1}^{m} I_j^{n_j}, \tag{A1}$$

where I_j is the expected amount of elastic scattering in the j-th pixel calculated using Eq. (6), and thus I_j/N_total is the probability of a single photon being registered at j-th pixel. Considering arrival positions $(x_i^{(l)}, y_i^{(l)})$ of all $n_i = \sum_{j=1}^{m} n_j$ photons separately, the above product reads $\prod_{l=1}^{n_i} I[\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)})]$, which yields Eq. (1).

[1] F. H. C. Crick and B. S. Magdoff, Acta Crystallogr. **9**, 901 (1956).
[2] W. A. Hendrickson, Science **254**, 51 (1991).
[3] M. G. Rossmann, Acta Crystallogr. Sect. A **46**, 73 (1990).
[4] H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall *et al.*, Nature **470**, 73 (2011).
[5] T. R. M. Barends, L. Foucar, S. Botha, R. B. Doak, R. L. Shoeman, K. Nass, J. E. Koglin, G. J. Williams, S. Boutet, M. Messerschmidt *et al.*, Nature **505**, 7482 (2014).
[6] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu, Nature **406**, 752 (2000).
[7] K. J. Gaffney and H. N. Chapman, Science **316**, 1444 (2007).
[8] V. L. Shneerson, A. Ourmazd, and D. K. Saldin, Acta Crystallogr. Sect. A **64**, 303 (2008).
[9] L. Young, E. Kanter, B. Krassig, Y. Li, A. March, and S. Pratt, Nature **466**, 56 (2010).
[10] S. P. Hau-Riege, Phys. Rev. A **76**, 042511 (2007).
[11] R. Neutze, G. Huldt, J. Hajdu, and D. van der Spoel, Radiat. Phys. Chem. **71**, 905 (2004).
[12] J. Miao, K. O. Hodgson, and D. Sayre, Proc. Natl. Acad. Sci. USA **98**, 6641 (2001).
[13] J. Miao, D. Sayre, and H. N. Chapman, J. Opt. Soc. Am. A **15**, 1662 (1998).
[14] G. Oszlanyi and A. Suto, Acta Crystallogr. Sect. A **60**, 134 (2004).
[15] V. Elser, Acta Crystallogr. Sect. A **59**, 201 (2003).
[16] D. R. Luke, Inverse Problems **21**, 37 (2005).
[17] C. David, S. Gorelick, S. Rutishauser, J. Krzywinski, J. Vila-Comamala, V. A. Guzenko, O. Bunk, E. Färm, M. Ritala, M. Cammarata *et al.*, Sci. Rep. **1**, 57 (2011).
[18] G. Huldt, A. Szoke, and J. Hajdu, J. Struct. Biol. **144**, 219 (2003).
[19] R. Fung, V. Shneerson, D. K. Saldin, and A. Ourmazd, Nat. Phys. **5**, 64 (2009).
[20] D. Giannakis, P. Schwander, and A. Ourmazd, Opt. Express **20**, 12799 (2012).
[21] Ne-Te Duane Loh and V. Elser, Phys. Rev. E **80**, 026705 (2009).
[22] M. Tegze and G. Bortel, J. Struct. Biol. **179**, 41 (2012).
[23] D. K. Saldin, V. L. Shneerson, R. Fung, and A. Ourmazd, J. Phys.: Condens. Matter **21**, 134014 (2009).
[24] B. von Ardenne, Master's thesis, Georg-August-Universität Göttingen, 2012.
[25] H. Liu, B. K. Poon, D. K. Saldin, J. C. H. Spence, and P. H. Zwart, Acta Crystallogr. Sect. A **69**, 365 (2013).
[26] T. Oroguchi and M. Nakasako, Phys. Rev. E **87**, 022712 (2013).
[27] D. Starodub, A. Aquila, S. Bajt, M. Barthelmess, A. Barty, C. Bostedt, J. D. Bozek, N. Coppola, R. B. Doak, S. W. Epp *et al.*, Nat. Commun. **3**, 1276 (2012).
[28] U. von Toussaint, Rev. Mod. Phys. **83**, 943 (2011).
[29] G. F. Schröder and H. Grubmüller, J. Chem. Phys. **119**, 9920 (2003).
[30] M. Moakher, SIAM J. Matrix Anal. Appl. **24**, 1 (2002).
[31] L. Schermelleh, R. Heintzmann, and H. Leonhardt, J. Cell Biol. **190**, 165 (2010).
[32] D. Nilsson, F. Uhlén, J. Reinspach, H. M. Hertz, A. Holmberg, H. Sinn, and U. Vogt, New J. Phys. **14**, 043010 (2012).
[33] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller *et al.*, J. Chem. Phys. **21**, 1087 (1953).
[34] E. von Castelmur, M. Marino, D. I. Svergun, L. Kreplak, Z. Ucurum-Fotiadis, P. V. Konarev, A. Urzhumtsev, D. Labeit, S. Labeit, and O. Mayans, Proc. Natl. Acad. Sci. USA **105**, 1186 (2008).
[35] N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina, and H. Stark, Nature **466**, 329 (2010).
[36] L. V. Bock, C. Blau, G. F. Schröder, I. I. Davydov, N. Fischer, H. Stark, M. V. Rodnina, A. C. Vaiana, and H. Grubmüller, Nat. Struct. Mol. Biol. **20**, 1390 (2013).
[37] N. D. Loh, M. J. Bogan, V. Elser, A. Barty, S. Boutet, S. Bajt, J. Hajdu, T. Ekeberg, F. R. N. C. Maia, J. Schulz *et al.*, Phys. Rev. Lett. **104**, 225501 (2010).
[38] R. Miles, Biometrika **52**, 636 (1965).
[39] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual* (Network Theory Ltd., UK, 2009).
[40] M. Matsumoto and T. Nishimura, ACM T. Model. Comput. S. **8**, 3 (1998).
[41] M. Frigo and S. G. Johnson, Proc. IEEE **93**, 216 (2005).
[42] S. P. Brooks and B. J. T. Morgan, The Statistician **44**, 241 (1995).

[43] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[44] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, J. Phys. Chem. B **105**, 6474 (2001).
[45] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).
[46] G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).
[47] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[48] D. I. Svergun, M. V. Petoukhov, and M. H. J. Koch, Biophys. J. **80**, 2946 (2001).