Computer Physics Communications 201 (2016) 159–166

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

BASDet: Bayesian approach(es) for structure determination from single molecule X-ray diffraction images☆

Michał Walczak, Helmut Grubmüller*

Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form
16 December 2015
Accepted 19 December 2015
Available online 29 December 2015

Keywords:
Single molecule
X-ray free electron laser diffraction
Structure determination
Bayesian analysis
Monte Carlo

ABSTRACT

X-ray free electron lasers (XFEL) are expected to enable molecular structure determination in single molecule diffraction experiments. In this paper, we describe an implementation of two orthogonal Bayesian approaches, previously introduced in Walczak and Grubmüller (2014), capable of extracting structure information from sparse and noisy diffraction images obtained in these experiments. In the 'Orientational Bayes' approach, a 'seed' model is used to determine for every recorded diffraction image the underlying molecular orientation. The molecular transform of the irradiated molecule is obtained by aligning and averaging those images in three-dimensional reciprocal space. By contrast, in the 'Structural Bayes' approach, a real space structure model is optimized to fit best to an *entire* set of diffraction images. This approach is used in a Monte Carlo structure refinement procedure.

Both presented approaches were implemented in C; previous tests (Walczak and Grubmüller, 2014) suggest that the algorithms are robust against low signal to noise ratios and can deliver high resolution structural information.

Program summary

Program title: BASDet

Catalogue identifier: AEZH_v1_0

Program summary URL: http://cpc.cs.qub.ac.uk/summaries/AEZH_v1_0.html

Program obtainable from: CPC Program Library, Queen's University, Belfast, N. Ireland

Licensing provisions: GPL version 3

No. of lines in distributed program, including test data, etc.: 1881590

No. of bytes in distributed program, including test data, etc.: 49039580

Distribution format: tar.gz

Programming language: C (ANSI 99), Perl.

Computer: Workstation (8 CPUs).

Operating system: Linux.

Classification: 3, 4.13, 16.1.

External routines: GNU Scientific Library (GSL), Message Passing Interface (MPI) library

Nature of problem: Extracting structural information from sparse and noisy single molecule XFEL diffraction images.

Solution method: Bayes' formalism is used to calculate either molecular orientation probability distribution with the aim to align individual images; or, alternatively, to calculate directly structure probability given all collected images.

Running time: The examples given:

Orientation_Bayes—50 h on Ivy Bridge Cores Xeon E3-1270v2 ($2 \times 4 \times 3,5$ GHz)

☆ This paper and its associated computer program are available via the Computer Physics Communication homepage on ScienceDirect (http://www.sciencedirect.com/science/journal/00104655).

* Corresponding author.
E-mail address: hgrubmu@gwdg.de (H. Grubmüller).

http://dx.doi.org/10.1016/j.cpc.2015.12.014

Structural_Bayes—These take longer than Orientation_Bayes runs, but can be restarted from checkpoint files.

1. Introduction

Structure determination techniques providing high resolution information are important to understand how biological systems function. X-ray crystallography, as the most widely used high resolution technique, faces certain limitations, however. Many proteins that cannot be crystallized are inaccessible to this technique. For those samples that do form crystals, only intensities at discrete Bragg peaks are measured, hence the missing phase information needs to be retrieved by other means.

Single molecule diffraction experiments with ultra short X-ray free electron laser (XFEL) pulses hold the promise to overcome these limitations [1–3]. In XFEL experiments, every collected image is created by a single molecule, randomly oriented during its exposure to a laser pulse. Despite the very high beam intensity, single molecules scatter only few photons that carry structural information (10–10^4 per picture, depending on molecular mass and beam intensity) [4]. The task to extract structural information from this limited data is further complicated by the presence of considerable background noise.

Recently developed structure determination methods from single molecule XFEL diffraction images focus either at accurate orientation determination for individual diffraction images and averaging those in 3-dimensional (3D) reciprocal space [4–9], or omit the orientation determination altogether by calculating intensity correlations [10–13].

Previously, we developed two alternative Bayesian approaches for structure determination with the intent to retrieve structural information at atomic resolution from sparse and noisy images [14], as depicted in Fig. 1. These approaches are referred to as 'Orientational Bayes' and 'Structural Bayes', respectively. In the Orientational Bayes approach, for every *individual* diffraction pattern $\mathbf{X}$, the probability of the underlying molecular orientation $\mathbf{\Theta}$, $\pi(\mathbf{\Theta}|\mathbf{X})$, is calculated and subsequently used to align the collected images in 3D reciprocal space. By contrast, the Structural Bayes approach does not rely on the molecular orientation determination for individual images; instead, the probability that a model structure S gave rise to an *entire* recorded set of diffraction patterns $\{\mathbf{X}\}$, $\pi\left(S|\{\mathbf{X}\}\right)$, is computed. The Structural Bayes approach is applied in a Monte Carlo (MC) structure refinement scheme. In this paper, we describe the implementation of both approaches.

2. Theory

Before we dwell on the implementation and use of the program, we briefly outline the theoretical foundations of the algorithms. For a derivation of formulas presented here and their in depth discussion, please refer to our previous work [14].

2.1. Orientational Bayes

The aim of the Orientational Bayes approach is to determine the underlying molecular orientation for each of the recorded diffraction patterns. The likelihood $f(\mathbf{X}|\mathbf{\Theta})$ that a particular diffraction pattern $\mathbf{X}$ is observed for an orientation $\mathbf{\Theta}$ is directly calculated from the recorded image, given a 'seed' structure model. However, to extract the hidden information about the underlying orientation, a posterior probability $\pi(\mathbf{\Theta}|\mathbf{X})$ is calculated via Bayes' theorem, $\pi(\mathbf{\Theta}|\mathbf{X}) \propto p(\mathbf{\Theta})f(\mathbf{X}|\mathbf{\Theta})$, under an *a priori* assumption about the molecular orientation distribution $p(\mathbf{\Theta})$.

To calculate the likelihood, we assume that every incident XFEL pulse contains a constant total number of photons N_{total}. However, for a diffraction image i, only n_i of those photons are recorded on a detector plane, whereas the remaining $N_{\text{total}} - n_i$ photons are not. The arrival positions of the n_i photons form a diffraction pattern $\mathbf{X}_i = \left\{(x_i^{(l)}, y_i^{(l)})\right\}_{l=1\ldots n_i}$ on the detector plane. The likelihood of observing a particular pattern $\mathbf{X}_i$, given a molecular orientation defined in terms of Euler angles $\mathbf{\Theta}_i = (\theta_i, \psi_i, \varphi_i)$, is thereby obtained by multiplying independent probabilities of detecting individual photons at positions $(x_i^{(l)}, y_i^{(l)})$ and the probability of the remaining $N_{\text{total}} - n_i$ photons not being recorded

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

Here, $I_{\mathbf{\Theta}_i}[\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)})]$ is the expected scattering intensity value registered by a detector pixel at the 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 scattering for orientation $\mathbf{\Theta}_i$ registered by an N_{pixel} pixel detector. Note that in actual experiments, incident beam intensity, and the resulting number of photons N_{total}, will vary from shot to shot. Hence, to account for these fluctuations, the incident beam intensity has to be treated as an additional parameter to be optimized, similarly as described in [15].

Assuming unpolarized XFEL pulses, the expected scattering intensity value registered by a detector pixel is given by

$$I(\Delta\mathbf{k}) = r_e^2 \frac{1 + \cos^2 2\gamma}{2} \Delta\Omega \int_{-\infty}^{\infty} \mathrm{d}t\, I_0(t) \left| \int \mathrm{d}^3\mathbf{r}\, \rho(\mathbf{r}, t) \mathrm{e}^{\mathrm{i}\Delta\mathbf{k}\cdot\mathbf{r}} \right|^2$$
$$+ \frac{A_{\text{BN}}}{2\pi\sigma} \mathrm{e}^{-\Delta k^2/(2\sigma^2)}, \quad (2)$$

where I_0 is the incident beam intensity, r_e is the classical electron radius, γ is the scattering angle, $\Delta\Omega$ is a solid angle subtended by the detector pixel, $\rho(\mathbf{r}, t)$ is the time dependent electron density [1], A_{BN} is the expected amount of background noise due to inelastic scattering, and σ is the anticipated width of the background noise distribution. In our calculations, we assume sufficiently short pulses (<10 fs) with low temporal coherence, and thus scattering amplitudes of a time-independent electron density are incoherently summed to compute intensity distributions. If required, pulse polarization and partial coherence can be accounted for in Eq. (2); however, this generalization of Eq. (2) is beyond the scope of this paper.

Note that to account for inelastic scattering noise, in Eq. (2), a normal distribution is added incoherently to the intensity distribution of the target molecule. The parameters A_{BN} and σ are set such as to model anticipated noise that mainly affects the central region of the detector, but also extends to high resolution regimes. In the actual XFEL experiments further background noise sources will affect the signal photon counts. Thus the noise model in Eq. (2) will

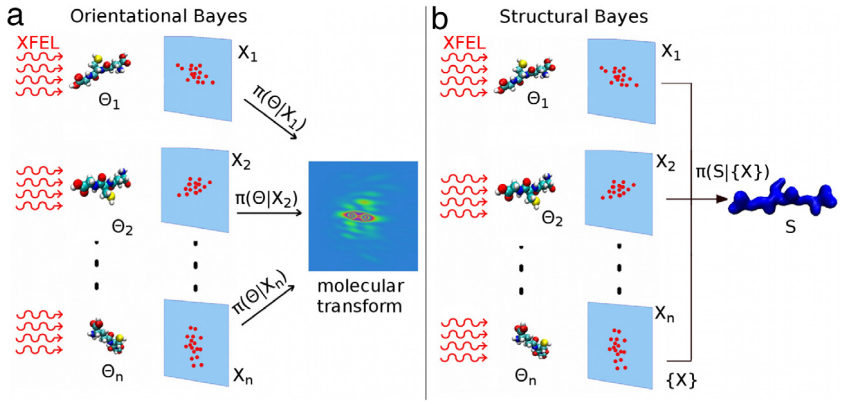

Fig. 1. Two alternative Bayesian approaches addressing the following questions: (a) Given a diffraction image $\mathbf{X}$, what is the probability of a particular orientation $\boldsymbol{\Theta}$ of the target molecule $\pi(\boldsymbol{\Theta}|\mathbf{X})$? The underlying molecular transform is obtained by aligning and averaging the images in reciprocal space. (b) Given an *entire* set of diffraction images $\{\mathbf{X}\}$, how probable is it that they all stem from a particular real space structure S, $\pi\left(S|\{\mathbf{X}\}\right)$?.

have to be modified to account for, e.g., electronic detector noise, irregular photon background, scattering on hydration shells, etc. Once single molecule scattering data is available, a suitable background noise model can be devised and implemented in the algorithm in a similar fashion to XFEL scattering on nanoparticles [15].

We further assume that the orientation of a single molecule entering the XFEL pulse follows a uniform distribution. Hence, according to Bayes' theorem, the posterior probability $\pi(\boldsymbol{\Theta}_i|\mathbf{X}_i)$ of the molecular orientation given the diffraction pattern is proportional to the likelihood $f(\mathbf{X}_i|\boldsymbol{\Theta}_i)$, defined in Eq. (1), with an irrelevant proportionality constant. If the user anticipates that target molecules follow a particular orientation distribution, different from a uniform one, both the prior and posterior distributions in the algorithm need to be modified such that they reflect this initial knowledge.

2.2. Structural Bayes

Note that the Orientational Bayes approach requires a 'seed model' to calculate the intensity distribution [Eq. (2)] and consequently the posterior probability distribution [Eq. (1)]. To avoid this prerequisite, we explicitly include the molecular structure in the posterior probability distribution as an additional parameter. In the most general case, the initial structure is modelled utilizing stereochemical knowledge about the target molecule, i.e., the sequence and the internal structure of the building blocks, e.g., amino acids, constituting the molecule. This way, starting from a random conformation, the structure is optimized to simultaneously fit best to *all* collected diffraction images.

A molecular structure is described by N atomic positions $S = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$. The likelihood of observing a diffraction pattern $\mathbf{X}_i = \{(x_i^{(l)}, y_i^{(l)})\}_{l=1,\ldots,n_i}$ obtained from photons scattered by a structure S_j oriented according to $\boldsymbol{\Theta}_i^{(j)} = (\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$ is

$$f(\mathbf{X}_i|S_j, \boldsymbol{\Theta}_i^{(j)}) \propto \left[1 - \frac{A(\boldsymbol{\Theta}_i^{(j)}, S_j)}{N_{\text{total}}}\right]^{N_{\text{total}}-n_i} \times \prod_{l=1}^{n_i} I\left[R(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)}), S_j\right]. \quad (3)$$

Here, $I(\Delta\mathbf{k}, S_j)$ is the expected scattering intensity value registered by a detector pixel at the photon position $(x_i^{(l)}, y_i^{(l)})$, the scattering vector $\Delta\mathbf{k}$ pointing to that pixel is rotated by a rotation matrix $R(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$ to account for the orientation $\boldsymbol{\Theta}_i^{(j)} = (\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$ of the structure S_j, $A(\boldsymbol{\Theta}_i^{(j)}, S_j) = \sum_{l=1}^{N_{\text{pixels}}} I\left[R(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})\Delta\mathbf{k}(x^{(l)}, y^{(l)}), S_j\right]$ is the expected amount of scattering for this molecular orientation registered by an N_{pixel} pixel detector, and N_{total} is the total number of incident photons.

Because individual images are recorded independently, the likelihood of recording an entire set of diffraction patterns $\{\mathbf{X}_i\}$ is given by the product of individual likelihoods

$$f\left(\{\mathbf{X}_i\}|S_j, \{\boldsymbol{\Theta}_i^{(j)}\}\right) = \prod_i f\left(\mathbf{X}_i|S_j, \boldsymbol{\Theta}_i^{(j)}\right). \quad (4)$$

The *a priori* distribution of structure coordinates $p(S_j)$ is assumed uniform, hence, according to Bayes' theorem, the posterior probability reads

$$\pi\left(S_j, \{\boldsymbol{\Theta}_i^{(j)}\}|\{\mathbf{X}_i\}\right) \propto \prod_i f\left(\mathbf{X}_i|S_j, \boldsymbol{\Theta}_i^{(j)}\right). \quad (5)$$

Finally, by integrating the above expression with respect to $\boldsymbol{\Theta}_i^{(j)}$, the posterior probability that structure S_j gave rise to the entire set of recorded diffraction images $\{\mathbf{X}_i\}$ is calculated

$$\pi\left(S_j|\{\mathbf{X}_i\}\right) \propto \prod_i \iiint f\left(\mathbf{X}_i|S_j, \theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)}\right) \times \sin\theta_i^{(j)} \mathrm{d}\theta_i^{(j)} \mathrm{d}\psi_i^{(j)} \mathrm{d}\varphi_i^{(j)}. \quad (6)$$

This expression is used to refine a structure model such that it fits best to the entire set of images.

3. Program structure

In this section, we describe the implementation of the two above described Bayesian approaches in computer programs for structure determination.

3.1. Orientational Bayes

As depicted in Fig. 2, the program based on the Orientational Bayes approach is divided into two stages. The first stage begins with the program parsing command line arguments; then it reads in input files containing k-space parameters, a seed model molecular transform, a set of N_{dp} diffraction patterns, and a pixel mask. After this preparatory stage, the Orientational Bayes approach is used to align individual diffraction patterns in 3D reciprocal space. Subsequently, the underlying molecular transform is retrieved from averaged photon counts in the aligned images.

For tests, we generated a set of synthetic diffraction patterns from the reference molecular transform, as shown in Fig. 3. To generate a diffraction pattern, first, a random molecular orientation

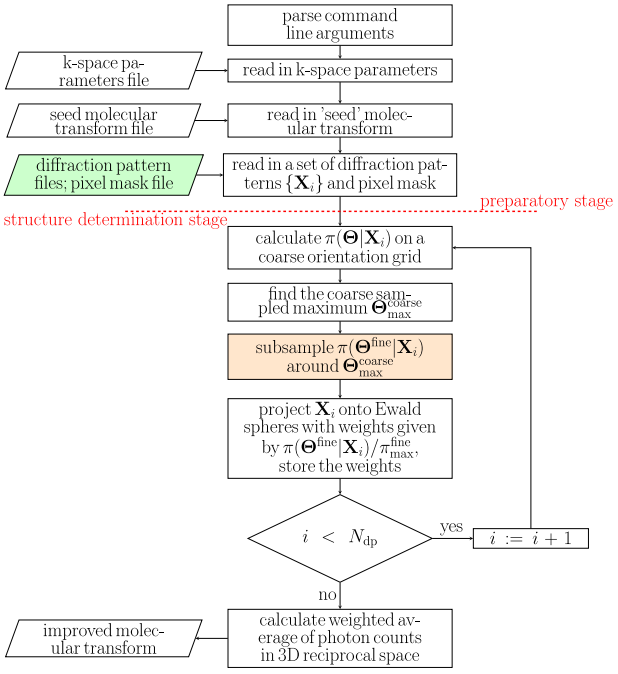

Fig. 2. Flowchart of the Orientational Bayes structure determination program. The subsampling procedure is detailed in Fig. 4.

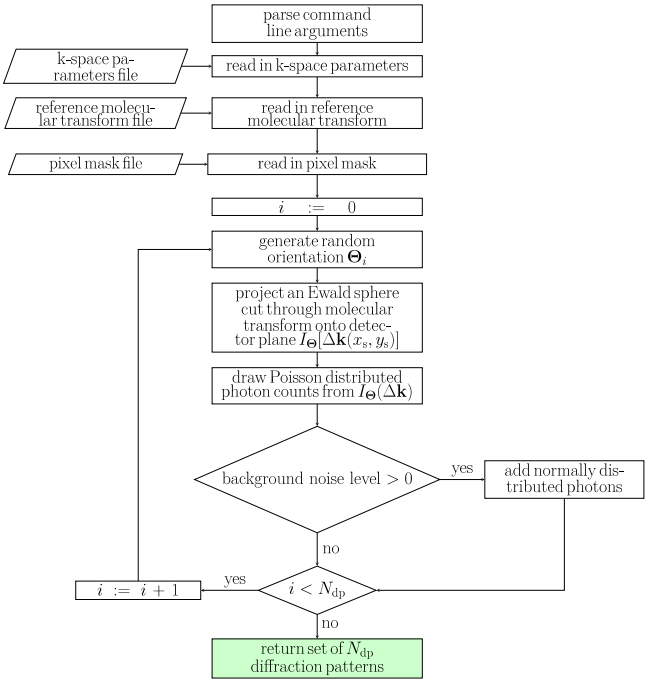

Fig. 3. Flowchart of diffraction pattern generation.

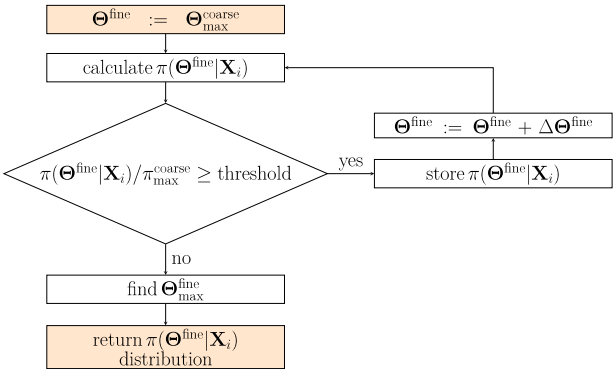

Fig. 4. Flowchart of subsampling around coarse sampled maximum $\boldsymbol{\Theta}^{\text{coarse}}_{\text{max}}$.

$\boldsymbol{\Theta}_i = (\theta_i, \psi_i, \varphi_i)$ is drawn from a uniform distribution. To this end, Euler angles are sampled from the following probability density $g(\theta, \psi, \varphi) = (8\pi)^{-1} \sin\theta$ [16], i.e., $\psi \in I[0, 2\pi)$, $\varphi \in I[0, \pi)$, and $\theta = \arccos z$, where $z \in I[-1, 1]$. Next, to calculate an intensity distribution recorded by the detector for that particular orientation $I_{\boldsymbol{\Theta}}[\Delta\mathbf{k}(x_s, y_s)]$, molecular transform values lying on appropriately oriented Ewald sphere are projected on the detector plane. Photon positions constituting the diffraction pattern are drawn at random from the calculated intensity distribution. For efficiency reasons, photon count fluctuations in every detector pixel are approximated by a Poisson distribution,

$$p(n, \Delta\mathbf{k}) = \frac{[I_\Theta(\Delta\mathbf{k})]^n}{n!} \mathrm{e}^{-I_\Theta(\Delta\mathbf{k})}, \tag{7}$$

where n is the photon count at a pixel corresponding to the scattering vector $\Delta\mathbf{k}$. We use the GNU Scientific Library [17] implementation of the 'Mersenne twister' algorithm [18] as pseudo random number generator to generate the diffraction patterns.

In XFEL scattering experiments, elastically scattered photons that carry the structural information are recorded along with others that do not contribute to the signal but form background noise. To account for the inelastic scattering noise, normally distributed photons are added incoherently to the diffraction patterns described above. The width and amplitude of the distribution are provided in the input parameters. This noise model is also included in posterior probability calculation by adding an appropriate Gaussian function to the intensity distribution within the right side of Eq. (2), as described in Section 2.1.

In the actual structure determination stage, for each diffraction pattern $\mathbf{X}_i$, the posterior probability $\pi(\boldsymbol{\Theta}|\mathbf{X}_i)$ is first calculated on a coarsely sampled orientational grid using Eq. (1). To achieve sufficient orientational resolution at a low computational cost, high probability regions are subsampled with a fine step. Thus, after the maximum $\boldsymbol{\Theta}^{\text{coarse}}_{\text{max}}$ of the coarse posterior probability landscape is located, surrounding regions, where the ratio of the fine sampled probability to the maximum of coarse sampled probability exceeds a given threshold t, $\pi(\boldsymbol{\Theta}^{\text{fine}}|\mathbf{X}_i)/\pi^{\text{coarse}}_{\text{max}} \geq t$, are subsampled. The subsampling procedure is illustrated by the flowchart in Fig. 4. Note that to avoid numerical underflows, in the source code, all posterior probability values are represented as logarithms and exponentiated when necessary.

Subsequently, photon positions are mapped onto differently oriented Ewald spheres with weights assigned by the posterior probability $\pi(\boldsymbol{\Theta}^{\text{fine}}|\mathbf{X}_i)/\pi^{\text{fine}}_{\text{max}}$. The weights are stored to calculate the weighted average of all accumulated photon counts in the final step. The molecular transform of the irradiated molecule is calculated by histogramming the weighted photon counts in corresponding 3D voxels of a Cartesian grid. The result is exported to an output cube file.

3.2. Structural Bayes

The Structural Bayes approach was implemented within a Monte Carlo refinement scheme. Given a sequence of amino acids, the refinement is carried out by randomly changing dihedral angles between residues of a peptide and calculating the probability of a particular conformation, given a recorded set of diffraction images. In every MC step, new dihedral angles are generated by varying previously accepted angles using a normally distributed step. The step size is updated during the simulation to maintain a desired acceptance rate.

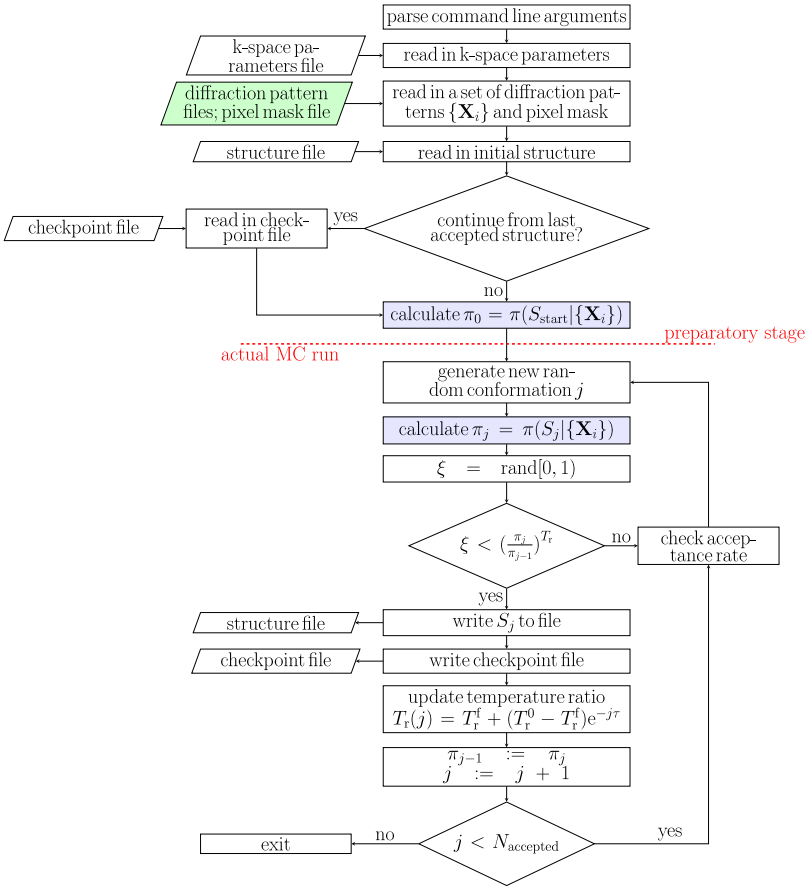

Fig. 5. Flowchart of the Monte Carlo refinement program using the Structural Bayes approach to calculate structure probability. Probability calculation procedure is detailed in Fig. 6.

Simulated annealing is used to prevent entrapment of MC runs in local energy minima. To this end, we introduce a dimensionless temperature ratio $T_r = T/T_a$ in the Metropolis criterion

$$\xi < \exp\left[\frac{(\ln \pi_j - \ln \pi_{j-1})T}{T_a}\right] = \left(\frac{\pi_j}{\pi_{j-1}}\right)^{T_r}, \tag{8}$$

where ξ is a random number between [0,1), $\pi_j = \pi\left(S_j|\{\mathbf{X}_i\}\right)$ is the probability of jth structure, T_a is the annealing temperature, and T is a pseudo temperature that ensures nondimensionality of the argument of the exponent function.

As depicted in the flow chart in Fig. 5, the program is divided into two stages: a preparatory stage and the actual MC run. In preparation for the actual MC run, the program first parses command line arguments, then it reads in input files containing k-space parameters, an initial random structure, a set of diffraction patterns, and a pixel mask. In case of a continuation run, the program additionally reads in a checkpoint file containing the dihedral angles combination from the last accepted step and the temperature ratio. In the last preparatory step, the starting structure probability $\pi_0 = \pi\left(S_{\text{start}}|\{\mathbf{X}_i\}\right)$ is calculated.

The flowchart in Fig. 6 illustrates the structure probability calculation. First, the molecular transform is calculated for structure S_j, using Eq. (2). To calculate scattering intensities, the electron density of the irradiated molecule is modelled as a sum of Gaussians centred at non-hydrogen atoms in the molecule:

$$\rho(\mathbf{r}) = \sum_{i=0}^{N} \frac{N_i^{\text{el}} e^{-(\mathbf{r}-\mathbf{r}_i)^2/(2\sigma_i^2)}}{(\sqrt{2\pi}\sigma_i)^3}, \tag{9}$$

where N_i^{el} is the number of electrons in the ith atom, $\mathbf{r}_i$ is its position, and σ_i is its radius. Next, for each diffraction pattern

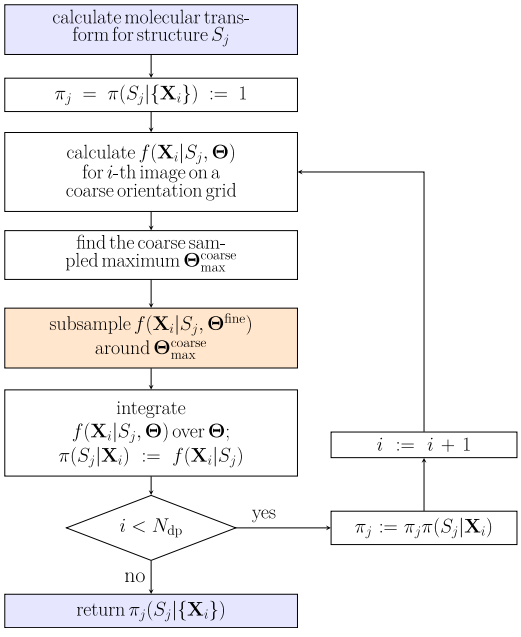

Fig. 6. Flowchart of the structure probability calculation procedure. The subsampling procedure is detailed in Fig. 4.

$\mathbf{X}_i$ in the set, the likelihood of observing that pattern, given the structure and its orientation, $f(\mathbf{X}_i|S_j, \boldsymbol{\Theta})$, is first calculated on a coarse orientational grid using Eq. (3). To increase accuracy, high probability regions around $\boldsymbol{\Theta}_{\text{max}}^{\text{coarse}}$ are subsampled with a fine

step, similar to the Orientational Bayes approach. The likelihood of observing the diffraction pattern given a particular structure $f(\mathbf{X}_i|S_j)$ is then obtained by integrating both coarse and fine sampled likelihood distributions $f(\mathbf{X}_i|S_j, \boldsymbol{\Theta})$ over orientations $\boldsymbol{\Theta}$. The integration is carried out using the rectangle rule; likelihood values are weighted by volume elements corresponding to coarse and fine sampling steps, respectively. In case fine sampled regions overlap with coarse sampled ones, the fine sampled volume is subtracted from appropriate coarse volume element. According to Eq. (6), the structure probability $\pi\left(S_j|\{\mathbf{X}_i\}\right)$ is proportional to the product of structure likelihoods $\prod_i f(\mathbf{X}_i|S_j)$ for individual images, however, the irrelevant proportionality constant cancels out in the structure probability ratio π_j/π_{j-1} used in the Metropolis criterion. Therefore, in the probability calculation, the structure likelihood value is directly assigned to the structure probability, $\pi(S_j|\mathbf{X}_i) := f(\mathbf{X}_i|S_j)$ in the flowchart in Fig. 6.

In the actual MC run, in every step, the structure probability of a proposed random conformation is calculated and used to evaluate the Metropolis criterion [Eq. (8)]. Every accepted structure is written to a file, corresponding dihedral angle configurations and current random step size are written to a checkpoint file. A temperature ratio update follows each accepted step j according to the annealing scheme $T_\mathrm{r}(j) = T_\mathrm{r}^\mathrm{f} + (T_\mathrm{r}^0 - T_\mathrm{r}^\mathrm{f})\mathrm{e}^{-j\tau}$, where T_r^0 is the initial temperature ratio, T_r^f is the final temperature ratio, and τ is a time constant. After each evaluation of the Metropolis criterion, the acceptance rate is checked against the requested threshold. The step size for dihedral angle generation is halved when the acceptance rate drops below the threshold, and doubled when the threshold is exceeded. Upon reaching the requested number of accepted structures, the program terminates.

4. Program usage

In this section, we describe all the in- and output options of both the Orientational Bayes and the Structural Bayes program.

4.1. Orientational Bayes

The Orientational Bayes program requires following input parameters:

- file containing seed model molecular transform (`-i`),
- configuration file describing reciprocal space (`-k`),
- coarse (`-C`) and fine (`-F`) orientation sampling steps,
- subsampling threshold (`-S`),
- number of diffraction patterns (`-n`),
- path to diffraction pattern files (`-P`),
- pixel mask file (`-M`),
- background noise level (`-l`),
- scaling prefactor for background noise model (`-A`),
- background noise distribution width (`-W`).

The volumetric data of molecular transforms is stored in Gaussian cube format [19]. The seed model molecular transform is used to calculate posterior probability distributions. Note that in Ref. [14], to test the structure determination quality, we used the same molecular transform as the reference and seed model.

The configuration file describing reciprocal space contains the following parameters:

- `size` specifies the number of voxels along an edge of a cubic grid that is used to discretize a molecular transform,
- `dk` is the cubic grid spacing,
- `sizep` is the number of detector pixels along one of its edges,
- `step` is the size of detector pixels in m,
- `pref` is a multiplication prefactor of the incident beam intensity $I_0 = \mathrm{pref} \times I_0^\mathrm{a}$, where $I_0^\mathrm{a} = 4 \times 10^6$ photons/Å^2 is an anticipated XFEL beam intensity in single molecule experiments [1].

An exemplary file used to simulate diffraction images and to determine the structure of a glutathione molecule [14] reads:

```
size 200
dk 0.01
sizep 121
step 0.001
pref 50
```

The subsampling procedure depicted in Fig. 4 requires a coarse orientation sampling step to locate regions of high posterior probability, a fine step to subsample those regions, and a threshold value to restrict the subsampling to relevant regions. By default, Euler angles $\theta = (0, \pi)$, $\psi = [0, 2\pi)$, and $\phi = [0, \pi)$ are discretized on the coarse grid with a 10° step, the fine sampling is carried out with a 2° step, and a value of 0.001 is used for the subsampling threshold to ensure an atomic resolution of the retrieved molecular transform [14].

The Orientational Bayes program requires diffraction pattern files generated externally. The first line in a diffraction pattern file contains the total number of entries. Each entry contains the number of photons at position (x_s, y_s) in the first column, followed by the coordinates in the other two columns. The photon position is expressed in pixel indices ranging from 0 to `sizep` − 1; position (0, 0) corresponds to the lower left corner of the detector. An exemplary diffraction pattern for a glutathione molecule reads:

```
76
1       11      54
1       38      40
1       39      46
1       40      41
...
```

Because photons cannot be recorded in certain areas of segmented detectors used in experiments, additional information specifying regions with no readout is required by the program. A pixel mask file contains the total number of entries, and (x_s, y_s) pixel positions as the entries.

To account for inelastic scattering background noise in diffraction patterns, the program requires parameters describing the background noise model. As mentioned in Section 2.1, a normal distribution is added incoherently to the molecular transform model. The standard deviation σ of the distribution is set from the input parameter `W` relative to the detector size `sizep`, $\sigma = W \times \mathrm{sizep}/2$. By default, the standard deviation is set to 1/10 of the detector size such that the background noise mainly affects the centre of the image, but also marginally extends towards high resolution regions. The amount of background noise A_BN expected in the images is calculated from the specified noise level `l` relative to the mean amount of elastic scattering per image `A`, $A_\mathrm{BN} = I_0 \times \mathrm{l} \times \mathrm{A}$. For a glutathione molecule, A = 1.64 Å^2. If no background noise level is provided, the program assumes that images contain shot noise only.

The output of the program is a molecular transform retrieved from the diffraction images. The molecular transform is written to a cube file specified by `-o` switch.

Example command to run the program is:

```
mpiexec -np 8 ./Orientational_Bayes -i moltr_gtt.cub
    -o reconstructed_moltr_BN10_10nmFa.cub -p
    kspace_params_gtt.txt -l 0.1 -s 10.0 -f 2.0 -n
    20000 -t 0.001 -S 1.64 -W 0.2 -P
    diffraction_patterns_dir/ -M pixel_mask.txt
```

here, with 10% background noise. In this input parameters configuration, the program took 12 h to run on 8 Harpertown cores (Intel Xeon Processor E5462, 2.8 GHz).

4.2. Structural Bayes

The Monte Carlo structure refinement program based on the Structural Bayes approach requires the following input parameters:

- initial random structure file number (`-i`),
- configuration file describing reciprocal space (`-k`),
- run number (`-r`),
- random number generator seed (`-B`),
- input checkpoint file for a continuation run (`-K`),
- output checkpoint file (`-V`),
- path to diffraction pattern files (`-P`),
- pixel mask file (`-M`),
- number of diffraction patterns (`-n`),
- background noise level (`-l`),
- scaling prefactor for background noise model (`-A`),
- background noise distribution width (`-W`),
- initial temperature ratio T_r^0 in annealing scheme (`-0`),
- final temperature ratio T_r^f in annealing scheme (`-f`),
- time constant τ in annealing scheme (`-t`),
- coarse (`-C`) and fine (`-F`) orientation sampling steps,
- subsampling threshold (`-S`),
- MC acceptance rate threshold (`-T`),
- output directory path (`-d`),
- number of total accepted MC steps (`-a`).

The following input parameters are the same as in the Orientational Bayes program, and were described in the previous section: configuration file describing reciprocal space, coarse and fine sampling step sizes, subsampling threshold, and the number of diffraction images.

Unlike in the Orientational Bayes, in the Structural Bayes approach, molecular structure is defined in real space. X, y, z coordinates and the type of all non-hydrogen atoms in the molecule are extracted from a Protein Data Bank (PDB) structure file [20]. Those initial random structures are generated prior to the MC run. The index number of an initial structure is passed to the program through the `-i` switch. For the same structure, one might want to initialize several independent runs. To that aim, a run number (`-r`) and a random number generator seed (`-B`) have to be provided to ensure sampling of different regions in configurational space.

In each MC step, dihedral angle configurations between residues are varied. The last accepted configuration is stored in an output checkpoint file. The first line in this file contains the last accepted MC step counter, the second line contains the step size for angle variation (in radians). Remaining lines contain dihedral angle values (in radians). The first column stores the angle of rotation around the $C_\alpha - C'$ bond (ψ) and the second around the $N - C_\alpha$ bond (ϕ). An exemplary checkpoint file for a glutathione molecule, as used in Ref. [14], reads:

```
44
0.087266
-1.979002       0.680588
0.223303        0.967061
```

Note that for this tripeptide, four instead of six dihedral angles were used because of a gamma peptide linkage between glutamate and cysteine residues. A continuation MC run requires a checkpoint file as an input parameter; in that case the simulation is renewed from the last accepted MC step.

For the simulated annealing scheme described in Section 3.2, input values of the initial and final temperature ratios along with the time constant are required. By default, $T_r^0 = 0.002$, $T_r^f = 1.2$, and $\tau = 0.005$ values are used. Values of these parameters were adjusted heuristically to perform simulations described in Ref. [14].

Every accepted structure is written to the directory specified by the `-d` flag. The output file structure is the same as the input file structure described above.

Example command to run the program is:

```
mpiexec -np 8 ./structural_bayes_MC -i 1 -k
    kspace_params_gtt.txt -r 2 -B 352314671 -V
    checkpoint2.txt -K checkpoint1.txt -P
    diffraction_patterns_dir/ -M pixel_mask.txt -n
    200 -0 0.002 -f 1.2 -t 0.005 -C 10.0 -F 2.0 -S
    0.001 -T 0.2 -d output_structure_dir/ -a 1000
```

Here, a run is continued from `-K checkpoint1.txt`. In this input parameters configuration, the program generated on average 24 accepted structures in 24 h when being run on 8 Ivy Bridge cores (Intel Xeon Processor E3-1270v2, 3.5 GHz).

5. Concluding remarks

We described the implementation of two alternative Bayesian approaches for structure determination from single molecule XFEL diffraction images. In the Orientational Bayes approach, *individual* diffraction patterns are aligned and averaged in 3D reciprocal space to retrieve the underlying molecular transform. By contrast, in the Structural Bayes approach the probability that a particular real space structure gave rise to the set of *all* recorded images is used to identify a structure that fits best to the *entire* collected patterns.

In our previous work [14], we demonstrated that the Orientational Bayes approach is capable of determining molecular structures at atomic resolution even from sparse and noisy images of small molecules, such as the glutathione. In fact, with increasing molecular size, a better spatial resolution is to be expected. Thus, small molecules will challenge prospective structure determination methods the most. The Orientational Bayes approach could be implemented in an iterative refinement scheme to retrieve the underlying molecular transform, similarly to an EMC (expansion–expectation maximization–compression) method [8]. In fact, a reader familiar with the XFEL field might notice a similarity between our likelihood formulation in Eq. (1) and that used in the EMC algorithm. However, in our approach, we consider probabilities of all individual photons, whereas the EMC approach regards diffraction patterns in terms of photon counts per pixel and uses a shot noise model such as the Poisson approximation. Note that in the limit of a large number of incident photons, as expected in XFEL experiments, the likelihood in Eq. (1) can be approximated by a Poisson distribution. Yet, it also applies to photon counts for which the Poisson approximation is invalid, and thereby Eq. (1) is more general. For a more detailed discussion on this issue, please refer to our previous work [14].

An iterative structure refinement is possible with the Structural Bayes approach implementation in a Monte Carlo scheme. The structure is then refined by sampling molecular conformations in real space. To construct a structural model, a minimal amount of *a priori* stereochemical knowledge about the target molecule is required: its composition and internal structure of the building blocks. Because a molecular structure is optimized against an entire set of diffraction images, achieving high spatial resolution should be possible even at very low average photon count per image. In our previous study [14], we successfully refined the structure of a glutathione tripeptide with a sub–Å accuracy from synthetic images containing on average 76 elastically scattered photons per image. Even though the algorithm was provided with random peptide conformations (with wrong dihedral angles), MC runs converged to a structure that closely resembled the reference.

Note, however, that for larger molecules such MC refinement based on exhaustive conformational sampling is a formidable task. For that reason, we also applied the Structural Bayes approach to calculate structure probability of a large biological complex, the ribosome, for a limited set of proposed structures, obtained other than by exhaustive conformational sampling. As demonstrated in our previous work [14], even localized minute structural changes, e.g., tRNA chain location, should be traceable despite large and inaccurately modelled regions, e.g., ribosomal subunits.

This computational toolkit is implemented in C and can be easily adapted to other structure determination or imaging problems challenged by low signal to noise ratios.

Acknowledgements

We thank Benjamin von Ardenne and Petra Kellers for reading the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft, grant SFB 755/B4.

References

[1] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, J. Hajdu, Potential for biomolecular imaging with femtosecond X-ray pulses, Nature 406 (6797) (2000) 752–757.
[2] R. Neutze, G. Huldt, J. Hajdu, D. van der Spoel, Potential impact of an X-ray free electron laser on structural biology, Radiat. Phys. Chem. 71 (3–4) (2004) 905–916.
[3] K.J. Gaffney, H.N. Chapman, Imaging atomic structure and dynamics with ultrafast X-ray scattering, Science 316 (5830) (2007) 1444–1448.
[4] V.L. Shneerson, A. Ourmazd, D.K. Saldin, Crystallography without crystals. I. The common-line method for assembling a three-dimensional diffraction volume from single-particle scattering, Acta Crystallogr. Sect. A 64 (2) (2008) 303–315.
[5] G. Huldt, A. Szoke, J. Hajdu, Diffraction imaging of single particles and biomolecules, J. Struct. Biol. 144 (1–2) (2003) 219–227.
[6] R. Fung, V. Shneerson, D.K. Saldin, A. Ourmazd, Structure from fleeting illumination of faint spinning objects in flight, Nat. Phys. 5 (1) (2009) 64–67.
[7] D. Giannakis, P. Schwander, A. Ourmazd, The symmetries of image formation by scattering. I, Theoretical framework, Opt. Express 20 (12) (2012) 12799–12826.
[8] N.-T.D. Loh, V. Elser, Reconstruction algorithm for single-particle diffraction imaging experiments, Phys. Rev. E 80 (2) (2009) 026705.
[9] M. Tegze, G. Bortel, Atomic structure of a single large biomolecule from diffraction patterns of random orientations, J. Struct. Biol. 179 (1) (2012) 41–45.
[10] D.K. Saldin, V.L. Shneerson, R. Fung, A. Ourmazd, Structure of isolated biomolecules obtained from ultrashort X-ray pulses: exploiting the symmetry of random orientations, J. Phys.: Condens. Matter. 21 (13) (2009) 134014.
[11] H. Liu, B.K. Poon, D.K. Saldin, J.C.H. Spence, P.H. Zwart, Three-dimensional single-particle imaging using angular correlations from X-ray laser data, Acta Crystallogr. Sect. A 69 (4) (2013) 365–373.
[12] 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., Single-particle structure determination by correlations of snapshot X-ray diffraction patterns, Nat. Comm. 3 (2012) 1276.
[13] B. von Ardenne, Reconstruction of electron densities from few photon single molecule X-ray scattering experiments (Master's thesis), Georg-August-Universität Göttingen, 2012.
[14] M. Walczak, H. Grubmüller, Bayesian orientation estimate and structure information from sparse single-molecule X-ray diffraction images, Phys. Rev. E 90 (2) (2014) 022714.
[15] 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., Cryptotomography: reconstructing 3D Fourier intensities from randomly oriented single-shot diffraction patterns, Phys. Rev. Lett. 104 (22) (2010) 225501.
[16] R. Miles, On random rotations in R3, Biometrika 52 (1965) 636–639.
[17] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, F. Rossi, GNU Scientific Library Reference Manual, Network Theory Ltd., 2009.
[18] M. Matsumoto, T. Nishimura, Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM T. Model. Comput. S. 8 (1) (1998) 3–30.
[19] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian09 Revision A.02, Gaussian Inc., Wallingford, CT, 2009.
[20] P.W. Rose, A. Prlić, C. Bi, W. Bluhm, C.H. Christie, S. Dutta, R.K. Green, D.S. Goodsell, J.D. Westbrook, J. Woo, et al., The RCSB protein data bank: views of structural biology for basic and applied research and education, Nucleic Acids Res. 43 (D1) (2015) D345–D356.