Bayesian orientation estimate from single molecule x-ray scattering data

Diplomarbeit

vorgelegt von
Michal Walczak
aus
Warschau

angefertigt am
Max Planck Institut für biophysikalische Chemie Göttingen
in der
Abteilung für theoretische und computergestützte Biophysik

30.08.2010

Contents

Chapter 1

Introduction

For a complete description of biological systems it is essential to know the function of their compounds. Thus, to understand ongoing processes in living organisms one has to determine the role of involved molecules in the first place. The behaviour of a biomolecule in question depends on its structure. Therefore, in theoretical biophysics atomic structure determination is of great importance.

Furthermore, without the information about atomic positions, it is impossible to perform molecular dynamics simulations [32], which are a powerful tool helping to understand the processes present in living organisms.

The main experimental method used for obtaining structural information about biomolecules is x-ray crystallography, which provides high resolution data. However, there are also some limitations of this technique. The most important obstacle, which has not been overcome yet, is the need of a crystalline specimen. In fact, only about 60% of all biomolecules can be crystallized [27]. Others, like some membrane proteins, do not form crystals because of the repulsive interactions between the hydrophobic and hydrophilic residues. And yet, it is important to understand the functions of the proteins embedded in the cell membranes of living organisms. Furthermore, crystallization and purification processes are sometimes laborious, and have low efficiency. The other limitation of the x-ray crystallography is the phase problem. Only the amplitude of the Fourier transform of the electron density function, also called the molecular transform, is measured. Additionally, the recorded diffraction pattern consists of discrete Bragg peaks resulting in an insufficient sampling in the reciprocal space. Thus, to calculate the electron density map, one has to deal with an underdetermined system due to the lack of the phase information. This problem, however, can be solved for instance by replacing some of the atoms with heavy ones and calculating the Patterson function [6]. Furthermore, tracing chemical reactions requires determining atomic positions within short time intervals. In x-ray crystallography a resolution in picoseconds range has been achieved for monitoring biomolecules [28].

Application of x-ray free electron (XFEL) lasers in single molecule experiments is expected to overcome the aforementioned problems of x-ray crystallography. With the XFEL it is possible to achieve ultra short pulses of high intensity in the hard x-rays regime. Recently, a hard-x-ray FEL generating a femtoseconds long pulse at Linac Coherent Light Source facility has been reported operational [33]. Another similar facility is under construction in Hamburg and should be completed by the end of 2013. In the planned single molecule diffraction experiments, a stream of hydrated particles will enter the x-ray beam at a rate of one molecule per pulse [12]. The molecules will be injected by applying electrospraying techniques. Since molecule orientation will be random, it will be necessary to determine the orientation corresponding to each obtained diffraction image. While in x-ray crystallography low intensity radiation (the intensity of XFEL pulses is approximately 10^6 higher than the synchrotron-radiation pulses) is distributed among many molecules located upon a lattice, in XFEL experiments extremely high doses will be absorbed by a single molecule. This means, each atom of the target molecule will absorb multiple photons within the duration of a single pulse. Thus, due to ionization effects, the electron density of the irradiated molecule will undergo changes and result in Coulomb explosion. Hence the pulse width should be possibly short, such that the explosion of the molecule will take place after the exposure. Because the electron density sample undergoes damage within the duration of the pulse, it influences the recorded diffraction image [15, 33]. Therefore, exposure times in femtoseconds range are essential to record the diffraction pattern before the initial structure suffers severely from ionization effects due to very high dose. The ultra high intensity per XFEL pulse is believed to result in a sufficient number of elastically scattered photons on a single molecule to reconstruct the electron density function. Further, the diffraction pattern is continuous due to the lack of translational symmetry in single molecules, in contrast to crystals, and thus it will be possible to oversample the reconstructed molecular transform. Hence, phases can be determined from the measured intensities by using iterative phasing algorithms [8, 20, 21, 26, 30]. Achieving pulse lengths in the femtoseconds regime is also expected to enable acquisition of time-resolved structural information, which can be applied to tracing enzymatic reactions [24].

Yet it is impossible to perform the reconstruction of the molecular transform from a single diffraction pattern. Firstly, despite the high intensity of the XFEL beam, a very low number of photons will be registered in a single diffraction pattern. In literature a prediction for a 500 kD molecule gives a value of about $4 \cdot 10^{-2}$ per pixel for mean photon count in the high resolution part [30]. As a result, diffraction patterns of molecules with small scattering cross sections will be affected by a low signal to noise ratio. Secondly, a single diffraction pattern provides only partial information about the molecule, as the detector plane corresponds only to a selected area on a certain Ewald sphere. Therefore, it is necessary to record a series of patterns from differently oriented molecules to obtain complete information about the object in the 3D reciprocal space. Since particles can rotate freely, their random orientation has to be extracted from the gathered scattering data. A method for achieving that is studied in this work. Estimation of the orientation in the reconstruction process results in additional error. Hence, to reduce the noise, averaging over many diffraction patterns is required.

A suitable method for the reconstruction purposes should thus be able to handle both sparse and noisy data. In their paper, Huldt et al. [16], have proposed the 'common line' orientation determination method based on the fact that two Ewald spheres, corresponding to diffraction patterns recorded on the detector plane, intersect in the reciprocal space creating a common curve. Locating the common line in any three diffraction patterns is sufficient to determine the relative orientations of the Ewald spheres. Because of the 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. However, Shneerson et al. in their numerical study of a scattering experiment for a chignolin molecule [30] have shown that the 'common line' method fails already at mean photon count of about 10 per pixel, which is three orders of magnitude higher than the predicted values. At very low numbers of registered photons, as expected in the XFEL experiments, it will be impossible to locate the common curves in the obtained diffraction patterns.

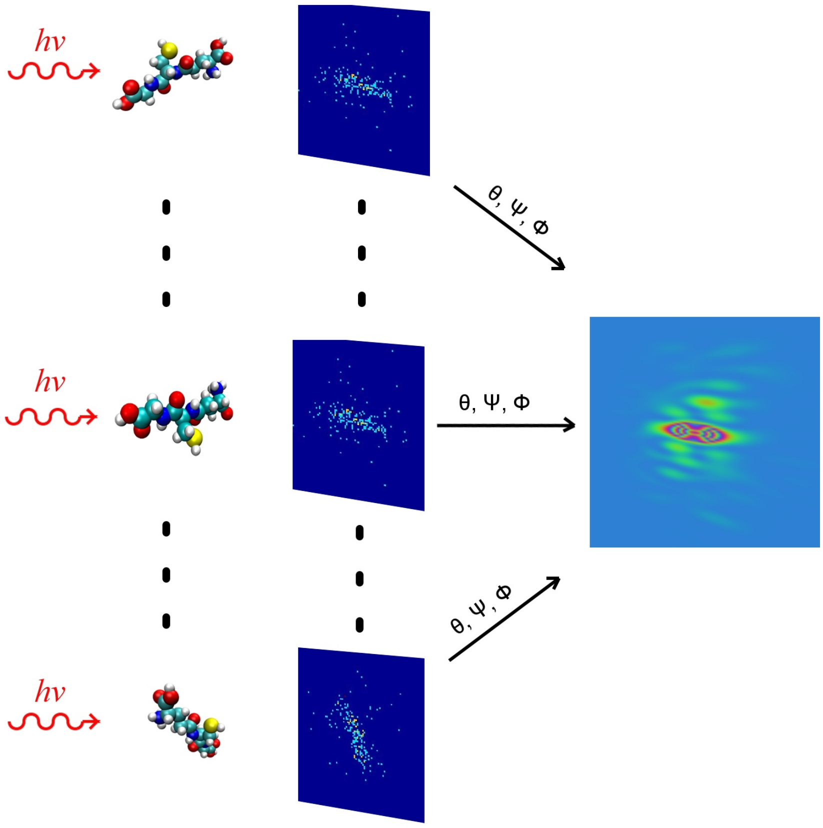

Figure 1.1: To reconstruct the molecular transform of a molecule its orientation has to be extracted from diffraction patterns consisting of very few photons.

A Bayesian based method for determining the orientations has been introduced by Fung et al. [11], in which generative topographic mapping is used to determine a maximum likelihood manifold in the orientational space. A clear advantage of that 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 be considered as a drawback. Assigning a weight to orientations allows better sampling of the reconstructed object in the reciprocal space.

Another approach to structure determination from the single molecule XFEL scattering experiments has been proposed by Saldin et al. [27]. The authors suggest it is not necessary to determine the orientations of single diffraction patterns to perform the reconstruction of the irradiated molecule. They have retrieved a molecular shape by computing a spherical harmonic expansion of a 3D object in the reciprocal space from cross-correlations between scattering images. More structural details might be obtained by applying that method for two molecules with similar structures, one of them known. However, the authors do not specify how much more detailed the extracted information is, as compared to the low resolution general molecular shape.

In this work I will introduce a rigorous statistical approach to reconstruction of the structure of single molecules in the XFEL scattering experiments. This approach is similar in spirit to the one developed for single molecule FRET experiments [29], which by applying Bayes' theorem enables high resolution reconstruction of distance trajectories from very few recorded photons. In this project three methods for structure determination from single molecule x-ray scattering experiments are introduced. In two of those methods Bayes' theorem is applied to extract the orientation information from diffraction images, and is used to locate the corresponding Ewald spheres so as to perform the reconstruction of the 3D object in the reciprocal space. Those methods, further referred to as 'maximum likelihood' and 'Bayes' methods, require a model structure as an input. The third method, derived from the other two Bayesian methods mentioned before, is a Monte Carlo approach to determine the tertiary structure of polypeptides knowing their primary structure. In fact, the two methods requiring a model structure serve with some modifications as a subroutine for the MC based method. Additionally, the influence of the noise on the performance of the methods, and values of input parameters, such as number of incident photons, required for a reliable reconstruction, are investigated in a numerical experiment.

Chapter 2

Theory

This section covers theoretical concepts of my work. In particular, I want to emphasize those elements of both x-ray scattering theory and Bayesian analysis applied in this project. Further, I introduce two methods for orientation determination and a Monte Carlo approach to structure determination, all of them based on Bayes' theorem.

2.1 X-ray free electron laser

Free electron lasers emit a beam of coherent electromagnetic radiation in a broad spectrum of wavelengths and of high energy. In an FEL the lasing medium is a relativistic electron beam passing through the periodic magnetic field of a magnetic structure called an undulator.

In the simplest case, an FEL uses a very long undulator generating a sinusoidal magnetic field, with a wavelength λ_0, perpendicular to the beam axis. The Lorentz transform of the magnetic field of the undulator from the observers to the electron frame of reference yields a plane electromagnetic wave with frequency $\gamma c/\lambda_0$, where γ is the electron energy expressed in units of electron rest mass [2]. Thus, the electron oscillations enforced by the magnetic field generate radiation. In the so called low-gain regime, in which the radiation field is almost constant, the energy exchange between the electrons and the radiation field $\mathbf{E}$ is described by

$$\frac{d\epsilon}{dt} = -e\mathbf{v_e} \cdot \mathbf{E}, \tag{2.1}$$

where ϵ is the electron energy and $\mathbf{v_e}$ is the electron velocity [9]. The energy exchange reaches optimum value for wavelengths close to undulator resonance

$$\lambda = \frac{\lambda_0}{2\gamma^2}(1 + 0.5K^2), \tag{2.2}$$

where $K = eB\lambda_0/2\pi mc$ is the undulator parameter [2, 9].

The acceleration and deceleration of the electron by the electromagnetic field results in a periodic velocity modulation, with a period corresponding to the wavelength λ. The velocity modulation further leads to the density modulation of the beam. This gives rise to more coherent electron radiation and exponential growth of the intensity (up to a saturation level).

2.2 X-ray scattering

As for any electromagnetic wave, the propagation of x-rays can be described with the general formula

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E_0}e^{i(2\pi\hat{\mathbf{k}}\cdot\mathbf{r}/\lambda-\omega t+\varphi)} \tag{2.3}$$

where $\mathbf{E}(\mathbf{r}$,t) is the electric field, $\hat{\mathbf{k}}$ is the unit wavevector in the propagation direction, ω is the angular frequency and φ is the phase. For x-rays, typical wavelengths range from 0.1 Å to 100 Å.

X-ray photons interact with matter in several ways. In the photoelectric effect an absorbed photon causes ejection of an electron, causing unstable electronic configuration. At 1 Å wavelength, the photoelectric cross section of a carbon atom is approximately 10 times larger than the elastic scattering cross section. Hence, biomolecules exposed to the XFEL beam undergo Coulomb explosion as the result of knocking off the electrons in the photoelectric effect [25]. Another possible event is the inelastic (or Compton) scattering, during which an x-ray photon transfers some of its momentum to a bound electron, and thus, the photon energy decreases while its wavelength increases. The third type of interaction, which is the relevant one in this work, is the elastic scattering. In that case, the photon energy is maintained, and only the direction of the momentum is changed.

In the elastic scattering electrons of the irradiated molecule become the source of secondary waves. Differences in relative electron positions lead to differences in optical path length, hence an interference pattern is recorded on a detector.

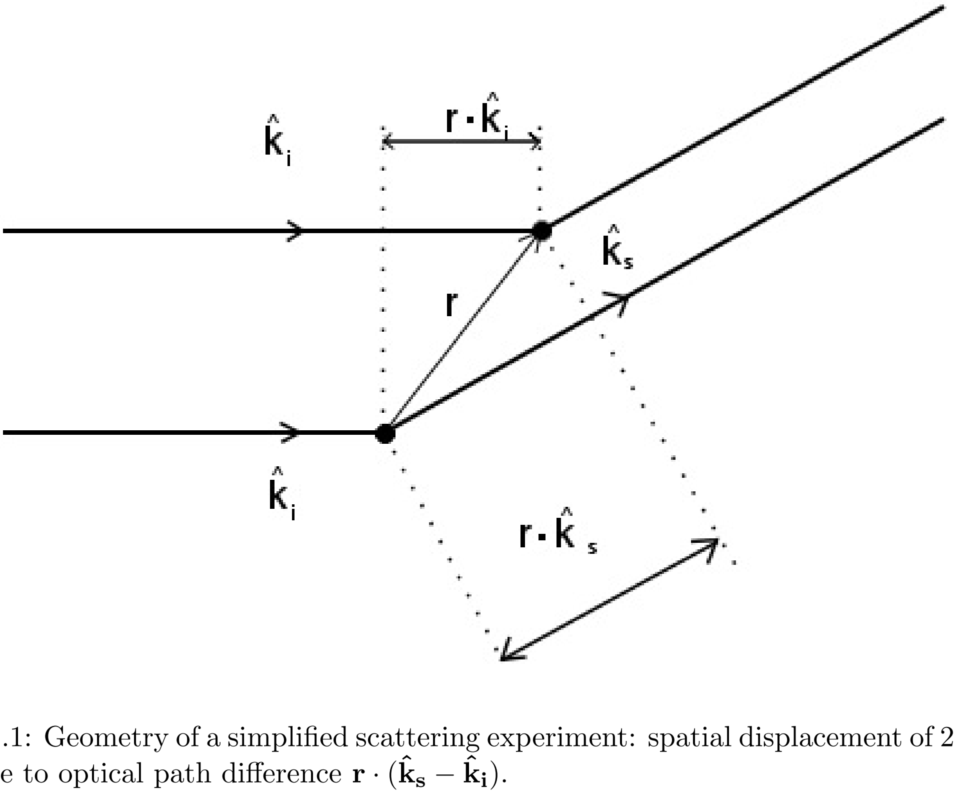

Figure 2.1: Geometry of a simplified scattering experiment: spatial displacement of 2 electrons gives rise to optical path difference $\mathbf{r}\cdot(\hat{\mathbf{k}}_\mathbf{s} - \hat{\mathbf{k}}_\mathbf{i})$.

Because the electrons in a sample are delocalized, it is essential to make use of the concept of electron density function. Structure factor is then given by

$$F(\Delta\mathbf{k}) = \iiint \rho(\mathbf{r}) e^{2\pi i \Delta\mathbf{k}\cdot\mathbf{r}} \mathrm{d}V, \tag{2.4}$$

which for N atoms, located at positions $\mathbf{r_n}$, is equivalent to

$$F(\Delta\mathbf{k}) = \sum_{n=1}^{N} f_n(\Delta\mathbf{k}) e^{2\pi i \Delta\mathbf{k}\cdot\mathbf{r_n}}, \tag{2.5}$$

where f_n are atomic scattering factors [6]. The scattered intensity recorded on the detector is defined as $I(\Delta\mathbf{k}) = F(\Delta\mathbf{k})F^*(\Delta\mathbf{k})$. For a real valued electron density function $F(\Delta\mathbf{k}) = F^*(-\Delta\mathbf{k})$ holds, thus the intensity distribution in the k-space reveals central symmetry, which is an important feature of x-ray imaging. In order to obtain the total scattered intensity, one has to include additional factor given by the Thomson formula, which relates the secondary (scattered) waves intensity to the incident wave

$$I(\theta) = I_0 r_e^2 \frac{1 + cos^2 2\theta}{2a^2}, \tag{2.6}$$

where I_0 is the incident beam intensity, a is the distance from the object to a point on the detector, r_e is the classical electron radius and θ is the scattering angle.

In the XFEL experiments, both the intensity of the incident pulse and the electron density function of the specimen are time dependent, the latter one as a result of radiation damage done by the incident beam. Thus, the registered intensity in such an experiment is given by

$$I(\Delta\mathbf{k}) = r_e^2 \frac{1 + cos^2 2\theta}{2a^2} \int_{-\infty}^{\infty} I_0(t) \left| \iiint \rho(\mathbf{r}, t) e^{2\pi i \Delta\mathbf{k}\cdot\mathbf{r}} \mathrm{d}V \right|^2 \mathrm{d}t. \tag{2.7}$$

The time evolution of the electron density was simulated for a carbon atom [15] assuming different pulse lengths. While average number of electrons in the K shell decreases only slightly with time, the loss of L shell electrons happens abruptly especially for longer pulses. Therefore, it was suggested [15] to compute the structure factor as

$$F(\Delta\mathbf{k}) = \sum_{n=1}^{N} \sum_{j} f_n^{(j)}(\Delta\mathbf{k}) e^{2\pi i \Delta\mathbf{k}\cdot\mathbf{r_n}}, \tag{2.8}$$

where $f_n^{(j)}$ is the atomic scattering factor of the n-th atom in the j-th electron configuration. For analysis of real diffraction images such a model would be the crudest possible approximation. In this project, however, for both generation and analysis of diffraction patterns, the electron density changes within the duration of the XFEL pulse were neglected, i.e. the intensity distribution was computed as squared modulus of the Fourier transform of time independent electron density.

2.3 Ewald spheres

An Ewald sphere is a geometrical construction from the incident and scattered wave vectors ($\mathbf{k_i}$ and $\mathbf{k_s}$). In case of the elastic scattering both vectors have a length of $2\pi/\lambda$, where λ is the wavelength, thus their difference, i.e the scattering vector $\Delta\mathbf{k}$, has to be located on a sphere with a radius of $2\pi/\lambda$.

To each pixel on the detector plane corresponds a scattered wave vector, hence by knowing the incident wave vector, it is possible to calculate the scattered vector, and map the pixels to points on an Ewald sphere. Because of limited surface of the detector a recorded diffraction pattern contains only partial information about the corresponding Ewald sphere.

Different orientations of the irradiated molecule in the real space are equivalent to rotations of the wave vectors, and of the Ewald spheres accordingly, hence, to reconstruct the 3D Fourier transform of the electron density one has to map the diffraction images, obtained for different orientation of the molecule, to the corresponding Ewald spheres and superimpose them in the reciprocal space.

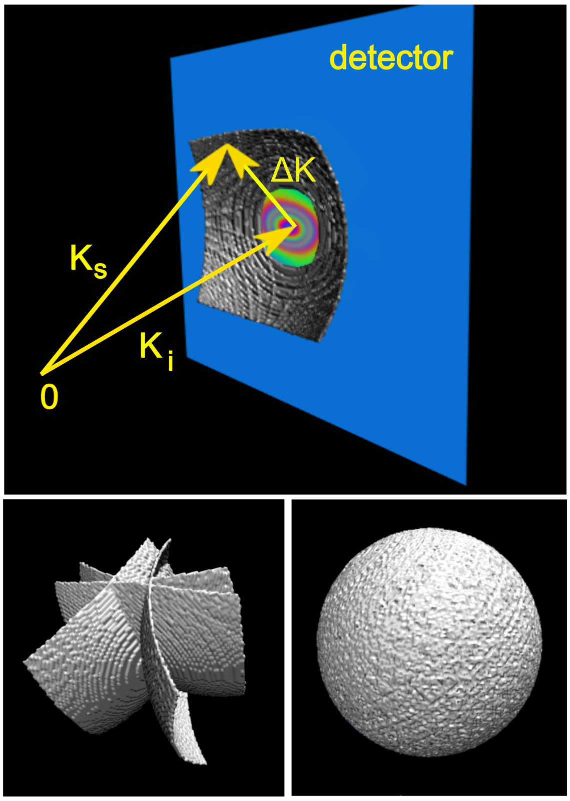

Figure 2.2: Pixels on the detector plane correspond to a part of an Ewald sphere determined by the incident ($\mathbf{k_i}$) and the scattered ($\mathbf{k_s}$) wave vectors. Superposition of many differently oriented Ewald spheres is essential to reconstruct a 3D object in the reciprocal space.

2.4 Bayesian analysis

The presented here derivation of Bayes theorem was taken from W. M. Bolstad's book [3].

Assuming a set of n disjoint events $B_1, \ldots, B_n$ whose union has a probability of one, and an observable event A, the total probability of A is given by

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i). \tag{2.9}$$

Further, assuming that both prior $P(B_i)$ and conditional probabilities $P(A|B_i)$ are known, the expression (2.9) can be rewritten as follows:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i). \tag{2.10}$$

By combining the usual formula for conditional probability of B_i given A $P(B_i|A) = \frac{P(A \cap B_i)}{P(A)}$ with (2.10), one obtains the Bayes' theorem, which states that

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}. \tag{2.11}$$

In other words, this theorem is used to revise prior beliefs about an event basing on the evidence gained from an experiment. The events $B_1, \ldots, B_n$ are not observed directly in a given experiment, yet one can assign prior probabilities, which reflect beliefs about the occurrence of those events. The conditional probability $P(A|B_i)$ is also called the likelihood of the event B_i, which is equivalent to a weighting function of B_i for the event A to be observed in the experiment. According to Bayes' theorem, the posterior probability $P(B_i|A)$ combines the beliefs prior to the experiment with updated beliefs after registering event A, thus containing entire information about the sample.

In the continuous case for observable x and a parameter θ belonging to the parameter space Θ, equation (2.11) becomes

$$\pi(\theta|x) = \frac{f(x|\theta) \cdot \pi(\theta)}{\int_{\Theta} f(x|\theta) \cdot \pi(\theta) \mathrm{d}\theta}. \tag{2.12}$$

For a given x, the integral in the denominator is a constant, thus the Bayes' theorem can be written as a proportionality

$$\pi(\theta|x) \propto f(x|\theta) \cdot \pi(\theta). \tag{2.13}$$

This can be further simplified by assuming $\pi(\theta)$ to be uniformly distributed i.e. using the noninformative prior

$$\pi(\theta|x) \propto f(x|\theta). \tag{2.14}$$

In this project I assume uniform distribution of the orientation, being the estimated parameter, for there is no reason why some orientations of the molecule entering the beam would be preferred to the others.

An advantage of Bayesian analysis is that it can be efficiently applied to small samples [1].

The main objective of many experiments is to find an estimate of the parameter θ basing on the posterior distribution. The easiest possible solution is finding, per analogy to the classical maximal likelihood method, the so called generalized maximum likelihood estimate of θ, being the largest mode of $\pi(\theta|x)$. This should yield satisfactory results for single mode distributions with narrow peaks, but in general, one should consider using other methods that make use of the entire information contained in the posterior probability distribution. In this work I compared the point estimate approach with use of the posterior probability distribution as a weighting function.

2.5 Posterior probability calculation

To estimate the orientation of a molecule that produced certain diffraction pattern, one has to compute the posterior probability distribution first. Assuming a uniform distribution of the orientations the simplified version of Bayes' theorem applies

$$\pi(\mathbf{\Theta_i}|\mathbf{X_i}) \propto f(\mathbf{X_i}|\mathbf{\Theta_i}), \tag{2.15}$$

where $\mathbf{\Theta_i} = (\theta_i, \psi_i, \varphi_i)$ is the orientation of the i-th molecule entering the beam, $\mathbf{X_i} = \{(x_i^{(l)}, y_i^{(l)})\}, l = l(i)$ denotes the i-th diffraction pattern, i.e. a set of positions of l recorded photons. An intuitive way to calculate the likelihood $f(\mathbf{X_i}|\mathbf{\Theta_i})$ is to treat intensity distributions corresponding to different orientations as probability distributions and express the probability of detecting photons at a certain pixel by taking the value of the intensity respectively

$$f(\mathbf{X_i}|\mathbf{\Theta_i}) \propto \prod_{l=1}^{l(i)} I_{\mathbf{\Theta}}(\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)})), \tag{2.16}$$

where by $I_{\mathbf{\Theta}}$ is denoted the intensity distribution from a molecule oriented according to $\mathbf{\Theta_i}$. The likelihood function, in this case, is a measure how well a given diffraction pattern overlaps with an intensity distribution for a certain orientation.

2.6 Reconstruction methods

The posterior probability distribution determined for each diffraction pattern is further used to perform the reconstruction of the Fourier transformed electron density in the 3D k-space. In the reconstruction process the wavevectors are rotated (corresponding to the orientation of the molecule that produced certain diffraction pattern), so the result is the molecular transform of the electron density of the molecule in the reference frame. The rotated fragments of the Ewald spheres cover the 3D k-space, if the sampling of the Euler angles is fine enough. I have studied two possibilities of using the information contained in the posterior probability distribution for purposes of the reconstruction.

2.6.1 'Maximum likelihood' method

The simplest reconstruction procedure is conducted as follows. The first step is to calculate for each diffraction pattern the posterior probability distribution using eq. (2.16) and (2.15), and estimate the Euler angles by locating the maximum of the posterior probability distribution. Having done that, for each detector pixel the scattering vector is computed and rotated according to the estimated angles. It is equivalent to locating the corresponding Ewald sphere in the 3D k-space. To reduce the noise in the reconstructed object, histogram averaging is performed. This is done by updating the reconstructed intensity value, corresponding to the rotated scattering vector, by the number of photons registered at the pixel, to which the scattering vector points, and incrementing the counter of the intensity entries.

The disadvantage of this method is the fact that it only uses part of the information contained in the posterior probability distribution, thus it is vulnerable to errors resulting from the posterior distribution maxima dislocation. A high level of both shot and background noise causes a shift of the position of the maximum with respect to the true orientation, hence the information about correct orientation is lost in this reconstruction method.

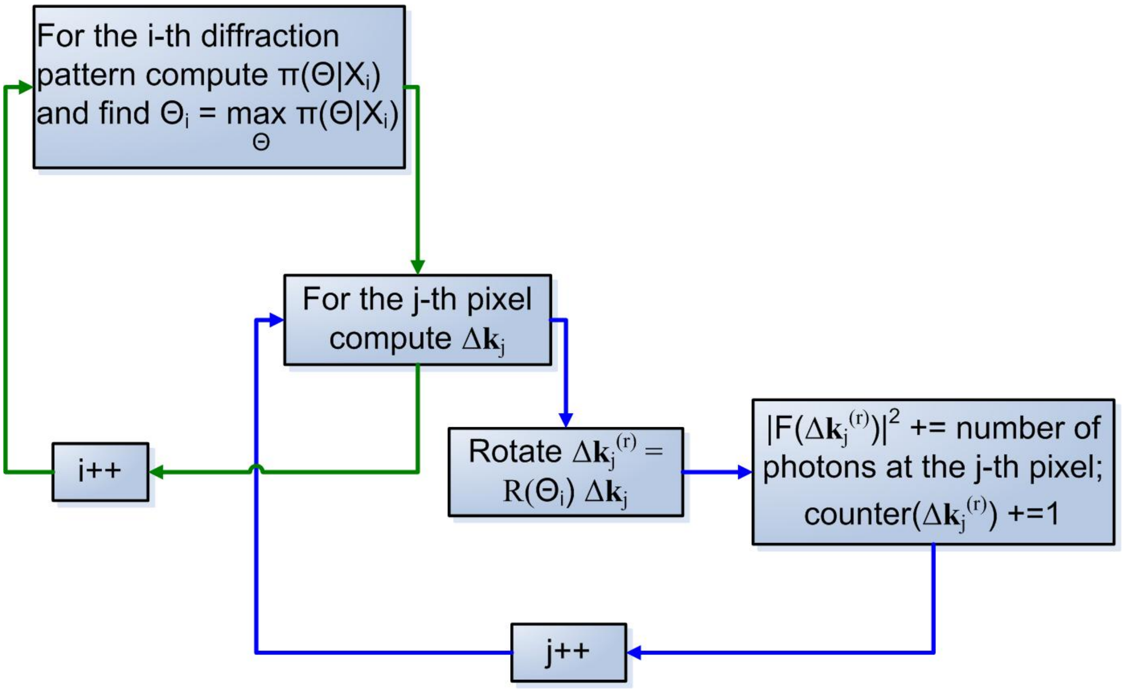
For the i-th diffraction pattern compute π(Θ|X_i) and find Θ_i = max_Θ π(Θ|X_i)
For the j-th pixel compute Δk_j
Rotate Δk_j^(r) = R(Θ_i) Δk_j
|F(Δk_j^(r))|^2 += number of photons at the j-th pixel; counter(Δk_j^(r)) +=1
i++
j++

Figure 2.3: Flow chart of the 'maximum likelihood' method

2.6.2 'Bayes' method

An improvement to the afore described reconstruction approach is achieved by using whole information contained in the posterior distribution function. The first step, calculating the posterior distribution for each diffraction pattern, is the same as in the 'maximum likelihood' method. The resulting posterior probability distribution is then used as a weighting function

$$W(\mathbf{\Theta}) = \pi(\mathbf{\Theta}|\mathbf{X_i})]/\pi_{max}, \tag{2.17}$$

which assigns the value 1 to the maximum of the posterior probability distribution. Then, for each detector pixel the scattering vector is computed and rotated according to each possible combination of Euler angles, so that all orientations, for which the posterior probability distribution was computed, are sampled. The reconstructed intensity values corresponding to the rotated scattering vector is updated by the number of photons multiplied by the orientation weight, and the counter of intensity entries is incremented by the weighting function value.

The weighted histogram averaging based method is less vulnerable to the disturbances in the posterior probability distribution caused by the shot and background noise, compared to the simple histogram averaging ('maximum likelihood') method. Even if not with the maximum probability, the true orientation is still taken into account during the reconstruction.

2.7 Phase retrieval

The reconstructed object in the 3D k-space carries only information about the amplitude of the Fourier transform of the electron density function (i.e. $I(\Delta\mathbf{k}) = |F(\Delta\mathbf{k})|^2$). Even though, it is possible to retrive the phase set iteratively from given intensity distribution.

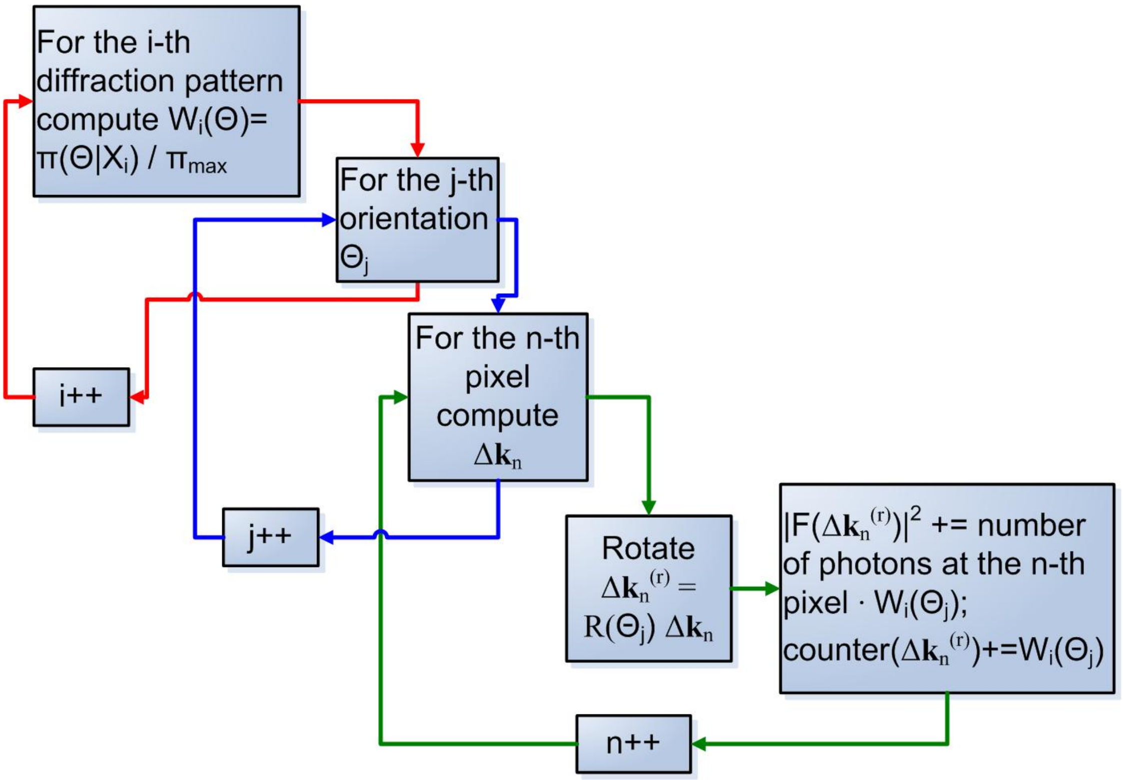
For the i-th diffraction pattern compute $W_i(\Theta)$= $\pi(\Theta|X_i)$ / π_{max}
For the j-th orientation Θ_j
For the n-th pixel compute $\Delta\mathbf{k}_n$
i++
j++
Rotate $\Delta\mathbf{k}_n^{(r)} = R(\Theta_j)\,\Delta\mathbf{k}_n$
$|F(\Delta\mathbf{k}_n^{(r)})|^2$ += number of photons at the n-th pixel · $W_i(\Theta_j)$; counter($\Delta\mathbf{k}_n^{(r)}$)+=$W_i(\Theta_j)$
n++

Figure 2.4: Flow chart of the 'Bayes' method

Several approaches basing on Fineup's algorithm [10] have been proposed for solving the phase problem in the x-ray scattering experiments. All of them [8, 20, 21, 30] are based on applying constraints to the real and reciprocal space and switch between them by means of the Fourier transform. One starts with a randomly guessed phase set, which is iteratively updated, such that the resulting electron density is positive within a finite support and zero otherwise. It is important to choose the finite support size that allows oversampling. The oversampling condition is fulfilled by choosing the volume with nonnegative electron density, such that its ratio to the total volume is greater than 2 [20].

2.8 De novo structure determination

The afore described methods require a model of the molecule as input data, whilst the ultimate goal of the x-ray structure determination is to reconstruct the electron density map without prior knowledge about the molecule structure. Here, I reformulate the structure determination problem. If one wants to determine relative orientation of protein subunits, whose structure is known, then the possible course of action is to generate an ensemble of potential conformations and apply Bayes' formulas, to find the most probable conformation with respect to given diffraction images.

The likelihood of observing diffraction pattern $\mathbf{X_i} = \left\{(x_i^{(l)}, y_i^{(l)})\right\}, l = l(i)$, being a set of positions of l recorded photons, given structure $S_j = \left\{(\mathbf{r}_1^{(j)}, \ldots, \mathbf{r}_N^{(j)}\right\}$, being a set of N atomic positions, and orientation of the j-th structure corresponding to the i-th diffraction pattern

$\mathbf{\Theta}_i^{(j)} = (\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$ is expressed as

$$f\big(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}\big) = \prod_{l=1}^{l(i)} I\Big(R(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)}), S_j\Big), \tag{2.18}$$

where $I(\Delta\mathbf{k}, S_j) = \Big|\iiint \sum_{m=1}^{N} a_m e^{-(\mathbf{r}-\mathbf{r}_m^{(j)})^2/(2\sigma_m^2)} e^{2\pi i\Delta\mathbf{k}\cdot\mathbf{r}} \mathrm{d}V\Big|^2$ is the intensity value for scattering vector $\Delta\mathbf{k}$ and structure S_j, similar as in eq. (2.7). Probabilities $f\big(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}\big)$ are independent, thus the probability of registering a set of diffraction patterns $\{\mathbf{X}_i\}$ is given by a product of those

$$f\big(\{\mathbf{X}_i\}|S_j, \{\mathbf{\Theta}_i^{(j)}\}\big) = \prod_i f\big(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}\big). \tag{2.19}$$

Assuming uniform distributions of orientations and structures, Bayesian theorem yields following formula for the posterior probability

$$\pi\big(S_j, \{\mathbf{\Theta}_i^{(j)}\}|\{\mathbf{X}_i\}\big) \propto \prod_i f\big(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}\big). \tag{2.20}$$

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

$$\pi\big(S_j|\{\mathbf{X}_i\}\big) \propto \prod_i \iiint f\big(\mathbf{X}_i|S_j, \theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)}\big) \sin\theta_i^{(j)} \mathrm{d}\theta_i^{(j)} \mathrm{d}\psi_i^{(j)} \mathrm{d}\varphi_i^{(j)}. \tag{2.21}$$

Evaluation of the probability distribution for different structures is thus a way to choose the most probable (with respect to the provided diffraction patterns) structure from a given set of conformers.

By applying aforementioned approach in a Monte Carlo simulation, it is possible to determine the tertiary structure of polypeptides basing on the primary structure and the diffraction patterns obtained from the experiment. Starting structure, i.e. one with a random set of dihedral angles, is modified each step according to new set of dihedral angles. For each step the posterior probability of the structure $\pi_j = \pi\big(S_j|\{\mathbf{X}_i\}\big)$ is computed from formula 2.21 and used in the Metropolis criterion [19]. Appointing energy to the probability of the j-th structure $E_j = -k_B T \ln \pi_j$ by introducing virtual temperature T, and plugging the energy term in the Metropolis criterion, the proposed structure is accepted if $\xi < e^{-\frac{\Delta E}{k_B T}} = \frac{\pi_j}{\pi_{j-1}}$ holds, where ξ is a random number from I(0,1). The outcome of this procedure is a canonical ensemble of structures, from which the average structure is further obtained.

Chapter 3

Materials and methods

Because no single biomolecule XFEL scattering data has been available up to now, the proposed methods had to be tested on synthetic data. Thus, in this section I first describe how the future XFEL scattering experiments were simulated. Consequently, I applied the three methods introduced in the previous section to the simulated data, and determined the accuracy of the obtained reconstruction as a function of number of incident photons, level of background noise, and number of provided diffraction patterns.

3.1 Molecules used in the numerical experiment

In order to additionally test the impact of symmetry of a molecule on results of the proposed methods, I decided to use benzene molecule. However, for the utmost part of my project, I used a glutathione molecule, which reveals no symmetry at all. This tripeptide, being a ligand, is a step towards proteins that will be used in the single molecule XFEL scattering experiments. Because of its small scattering cross section, the glutathione is a suitable candidate for investigating the influence of the background noise on the reconstruction methods performance. Also the fact that it consists of three amino acids makes it a good choice for the test molecule as a Monte Carlo based structure prediction approach.

3.2 Simulation of the x-ray scattering

The goal of the x-ray scattering experiments is localization of atoms of the irradiated molecule, thus at that level of details it is sufficient to model the electron density as a sum of Gaussian functions centered at the atomic positions

$$\rho(\mathbf{r}) = \sum_i a_i e^{-\frac{(\mathbf{r}-\mathbf{r_i})^2}{\sigma_i^2}} \tag{3.1}$$

where σ_i is the atom radius and coefficients a_i are chosen so that the integration of the Gaussian function yields the number of electrons. With the electron density function given by a sum of Gaussians, the Fourier transform is computed analytically

$$F(\Delta\mathbf{k}) = \mathcal{F}(\rho(\mathbf{r})) = \sum_i a_i \sigma_i^3 e^{j(\Delta\mathbf{k}\cdot\mathbf{r_i})} \cdot e^{-\Delta k^2 \sigma_i^2/2} \tag{3.2}$$

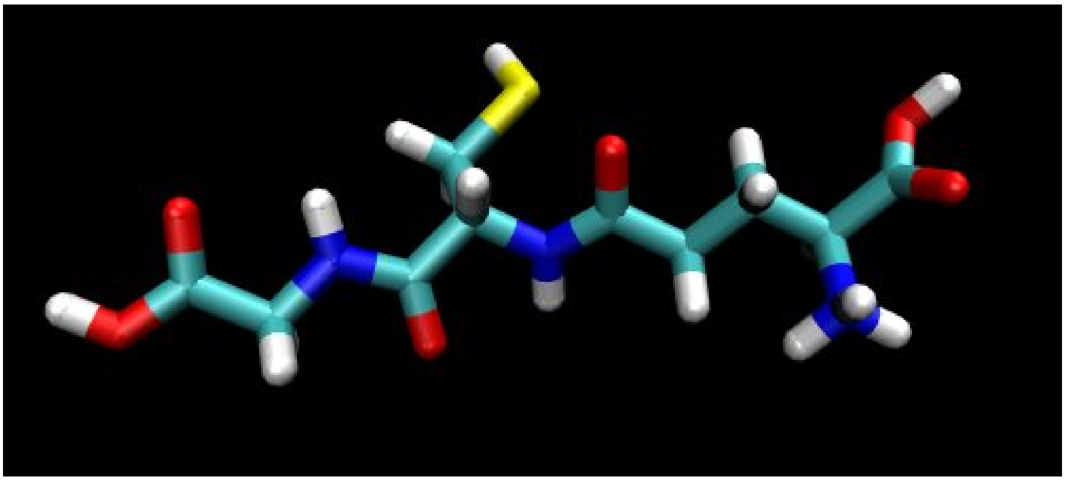

Figure 3.1: Rod shaped glutathione molecule. Following colour encoding is used: blue - nitrogen, green - carbon, red - oxygen, white - hydrogen, yellow - sulfur

To create an intensity distribution ensemble needed for further use, the orientational space is sampled with a 10 degrees step for each of the Euler angles. For any given orientation of the molecule, the intensity value for each pixel is calculated by determining the corresponding scattering vector from the geometric relations (see fig 3.2).

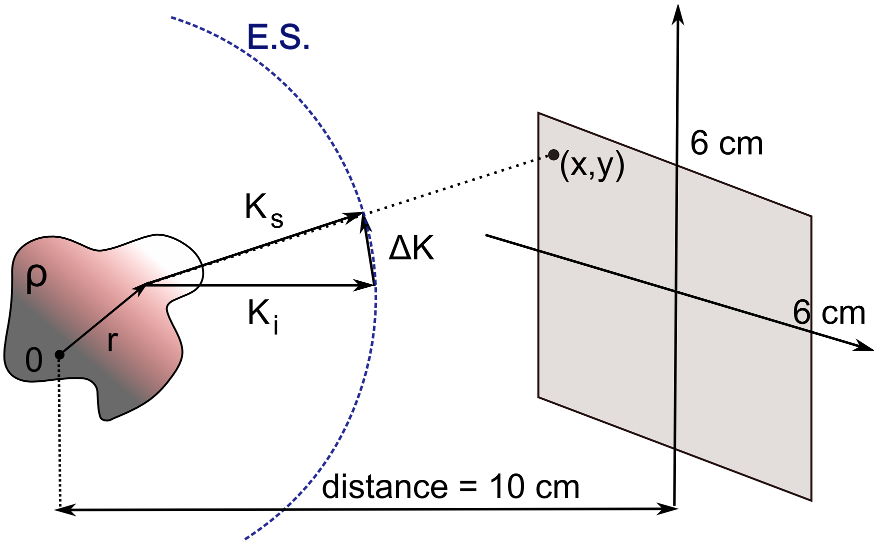

Figure 3.2: Geometry of the scattering experiment. $\Delta\mathbf{k}$ is calculated from geometric relations given the pixel coordinates on the detector plane and the distance between the sample and the detector.

The intensity value is computed from equation (2.7) using the time independent electron density function and treating the integrated intensity of the incident beam as a parameter. I have chosen 1 Å wavelength for the incident radiation. In order to take the orientation of the molecule into account, one has rotate the Ewald sphere corresponding to a given intensity distribution, i.e. to calculate $\Delta\mathbf{k}' = R\Delta\mathbf{k}$, where R is the rotation matrix corresponding to the orientation of the molecule given by the Euler angles.

The actual interaction between the x-ray radiation and the matter is of stochastic nature. Therefore, the number of photons registered by the detector in a single experiment will deviate from the intensity value calculated from (2.7). This is called either photon or shot noise and follows the Poisson distribution, hence, to mimic the experiment, the photon count n for a given pixel is determined from

$$p(n, \Delta\mathbf{k}) = \frac{[I(\Delta\mathbf{k})]^n}{n!} e^{-I(\Delta\mathbf{k})}. \tag{3.3}$$

Such patterns recorded on the detector plane I call the diffraction patterns in distinction from the intensity distributions. It is convenient to think of the diffraction pattern as a set of coordinates of recorded photons on the detector plane.

To mimic the background noise, which is also present in the experiments, an arbitrary number of photons uniformly distributed on the detector is added to the generated diffraction pattern.

3.3 Random numbers

It is essential to use a reliable random number generator in order to generate the diffraction patterns for randomly oriented molecules, so as to mimic the real XFEL experiments. The reconstruction procedure requires a high amount of diffraction patterns to average the noise out, therefore, a pseudo random numbers sequence appropriate for such a simulation should have possibly longest period and exhibit low correlations. Thus, I used the Gnu Scientific Library [13] implementation of the 'Mersenne twister' algorithm [18] which has a period of $2^{19937} - 1$ and is equally distributed in 632 dimensions.

3.4 Random orientations

Single molecules entering the XFEL beam are expected to be oriented randomly, following a uniform distribution. For Euler angles the invariant probability density is given by $g(\theta, \psi, \varphi) = (8\pi)^{-1} sin(\theta)$ [22], where the Euler angles are in xzx notation [4]. Therefore, to generate Euler angles resulting in uniformly distributed orientations, ψ and φ are drawn from uniform distribution $I[0, 2\pi]$, and $\theta = arccos(z)$, where z is a random number from uniform distribution $I[-1, 1]$.

3.5 Reconstruction methods

In posterior probability calculations evaluating the product of intensities for higher numbers of scattered photons causes underflows. To avoid that it is necessary to compute the natural logarithm of posterior probability

$$ln[\pi(\mathbf{\Theta_i}|\mathbf{X_i})] = const \cdot \sum_{l=1}^{l(i)} ln[I_{\mathbf{\Theta}}(\mathbf{\Delta k}(x_i^{(l)}, y_i^{(l)}))]. \tag{3.4}$$

Thus the weighting function in the 'Bayes' reconstruction method becomes

$$ln[W(\mathbf{\Theta})] = ln[\pi(\mathbf{\Theta}|\mathbf{X_i})] - ln[\pi_{max}], \tag{3.5}$$

The posterior probability distribution is initially sampled with a 10° step. To reduce the computational cost of mapping the points from the diffraction pattern to the 3D reciprocal space, instead of doing that for each possible orientation of the molecule, one can introduce a threshold value for the posterior probability, above which the mapping is performed. In that way only significant regions around the maxima in posterior probability region are taken into account. In order to achieve more accurate orientation estimate those regions are subsampled with a finer step. When the global maximum is located, the area around the maximum is subsampled with a step of 2°. The vicinity of the 'coarse sampling' maximum is defined by the probability ratio threshold $\pi^{fine}(\mathbf{\Theta}|\mathbf{X_i})/\pi^{coarse}_{max} \geq 10^{-4}$. In the 'maximum likelihood' approach the position of the 'fine sampling' maximum is used as the orientation estimate, whereas in the 'Bayes' method all probability values above the threshold are used as the weighting function $ln[W^{fine}(\mathbf{\Theta})] = ln[\pi^{fine}(\mathbf{\Theta}|\mathbf{X_i})] - ln[\pi^{fine}_{max}]$.

3.6 Phase retrieval

In this project, to retrieve phases I used a slightly modified version of the algorithm proposed by Miao et al. [20]. Before applying the algorithm to the reconstructed 3D molecular transform, the intensity values at **k** and -**k** are averaged, such that Friedel's law $I(\mathbf{k}) = I(-\mathbf{k})$, is satisfied, further a random phase set is generated. Afterwards, the fast Fourier transform from FFTW library [17] is applied to the data set in order to switch between the real and the reciprocal space. The finite support area in the real space is selected as a cube, with an edge twice the length of the radius of gyration (for glutathione molecule $R_g = 4.5\mathring{A}$), centered at the origin. The negative values of electron density inside the support have their sign changed [26], instead of setting them close to zero [20], because 'charge flipping' was claimed to improve the convergence of the algorithm [27].

3.7 Accuracy measure

One possibility to measure the accuracy of the orientation determination is to calculate the distance between two rotations. Matrix representation of a rotation by Euler angles belongs to the Lie group of orthogonal transformations, denoted by SO(3). The distance function in SO(3) is expressed by Riemannian metrics

$$d_R(R_1, R_2) = \frac{1}{\sqrt{2}} \parallel Log(R_1^T R_2 \parallel_F, \tag{3.6}$$

where $\parallel \cdot \parallel_F$ denotes Frobenius norm. The distance between two rotations can be interpreted as the arc-length of the shortest geodesic curve connecting rotations $\mathbf{R}_1$ and $\mathbf{R}_2$ [23]. The principal logarithm for a matrix R in SO(3) is computed from Rodrigues' formula

$$Log(R) = \begin{cases} 0 & \text{if } \theta = 0 \\ \frac{\theta}{2\sin\theta}(R - R^T) & \text{if } \theta \neq 0 \end{cases} \tag{3.7}$$

where $\theta = \arccos[0.5(trR - 1)]$ and $|\theta| < \pi$ is the rotation angle. In case $\theta = \pi$ the outcome of the Rodrigues' formula is undefined. The Frobenius norm of the logarithm, however, is equal to π.

Rotating an Ewald sphere corresponding to a diffraction pattern according to different orientations in the 'Bayes' method gives rise to a question about its effect on the resolution of the electron density map. As the estimate of the resolution I take the mean distance between the orientation corresponding to the posterior maximum and surrounding orientations, multiplied by the radius of gyration of the molecule. The mean distance between orientations is computed by bootstrap sampling [7] from the posterior distribution for diffraction patterns corresponding to the same orientation of the molecule.

3.8 De novo structure determination

To generate random structures of the reference glutathione tripeptide I change the dihedral angles in the glycine and cysteine residues. The glutamic acid, which is bonded to the cysteine in an unusual way, is left intact.

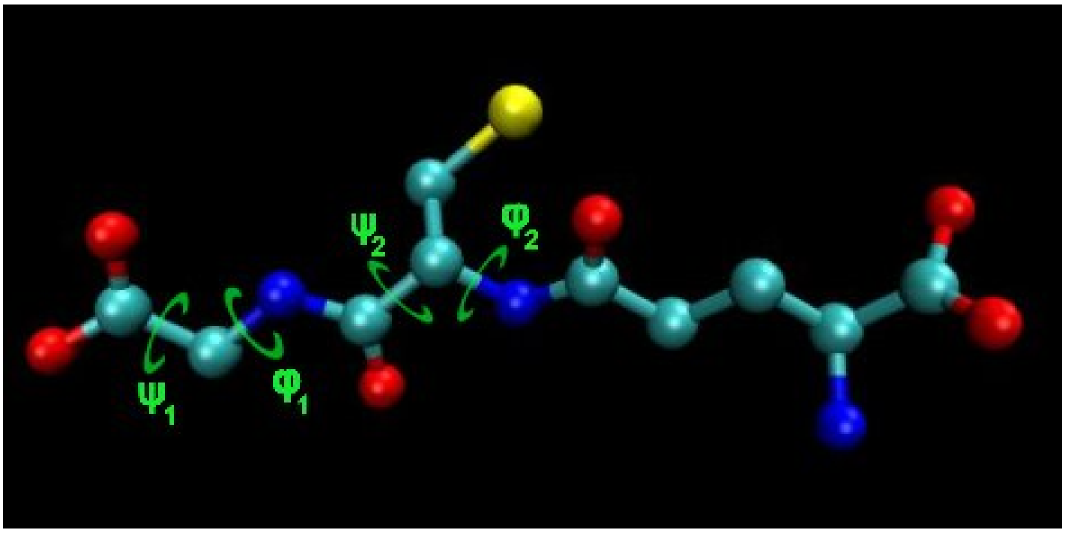

Figure 3.3: Four dihedral angles that are changed to generate different conformations of the tripeptide

A new set of dihedral angles is obtained from a Gaussian distribution with the mean equal to the values of last accepted angles and an arbitrarily chosen standard deviation. To prevent the method from being trapped in a local minimum of the energy landscape, simulated annealing [5] is applied. By introducing a virtual temperature ratio $T_r = T_a/T_b$, Metropolis criterion is given by

$$\xi < e^{\frac{(\ln \pi_j - \ln \pi_{j-1}) k_B T_a}{k_B T_b}} = \left(\frac{\pi_j}{\pi_{j-1}} \right)^{T_r}. \tag{3.8}$$

Hence starting with a low value of T_r initially improves the sampling. Each MC step the temperature ratio grows exponentially until it asymptotically reaches 1, by that time the system should have reached the global minimum of the energy landscape. The annealing scheme used in this project is given by $T_r(j) = 1 - e^{\ln(1-T_0) - j\tau}$, where j denotes the MC step,

$T_0 = 0.001$ is the starting ratio, $\tau = 0.002$ is the time constant. Values of those parameters were adjusted heuristically.
To additionally improve the sampling, the value of the standard deviation (starting value of $\pi/10$) of the Gaussian distribution of the dihedral angles is halved when the acceptance ratio drops below a chosen threshold (0.01).

Chapter 4

Results

In this section results of the numerical study of single molecule x-ray scattering are presented. Influence of molecule symmetry and of noise on the posterior probability distribution is shown. Here I will also provide a comparison of performance of proposed reconstruction methods, estimate the reconstructed electron density resolution dependence on incident beam intensity, and demonstrate the outcome of the MC approach to structure prediction.

4.1 Intensity distributions

Intensity distributions play a crucial role in the proposed molecule orientation determination method as they appear in the Bayes' formula. Therefore, it is essential that different oriented molecules should produce distinguishable intensity distributions. Hence, in the first stage of the project I investigated the intensity distributions obtained for two different molecules, benzene and glutathione.

As can be seen in figure 4.1, intensity distributions for planar and symmetric molecules, like benzene, are hardly distinct for different orientations. For each of the presented orientations, the intensity distributions are two dimensional Gaussians differing in the tilt of their axes with respect to the frame of reference of the screen, and in their widths. Changes in the θ angle, being the angle between the aromatic ring plane and the incident beam, are traceable within the intensity distributions, as they clearly correspond to the tilt of the Gaussians. However, the changes in remaining two angles, which are reflected in different widths of the Gaussians, do not seem to be distinctive, as for different ψ and φ combinations one obtains similar distributions (e.g. the ones for $\theta = 60°$, $\psi = \varphi = 0°$, and $\theta = 60°$, $\psi = 30°$, $\varphi = 90°$).

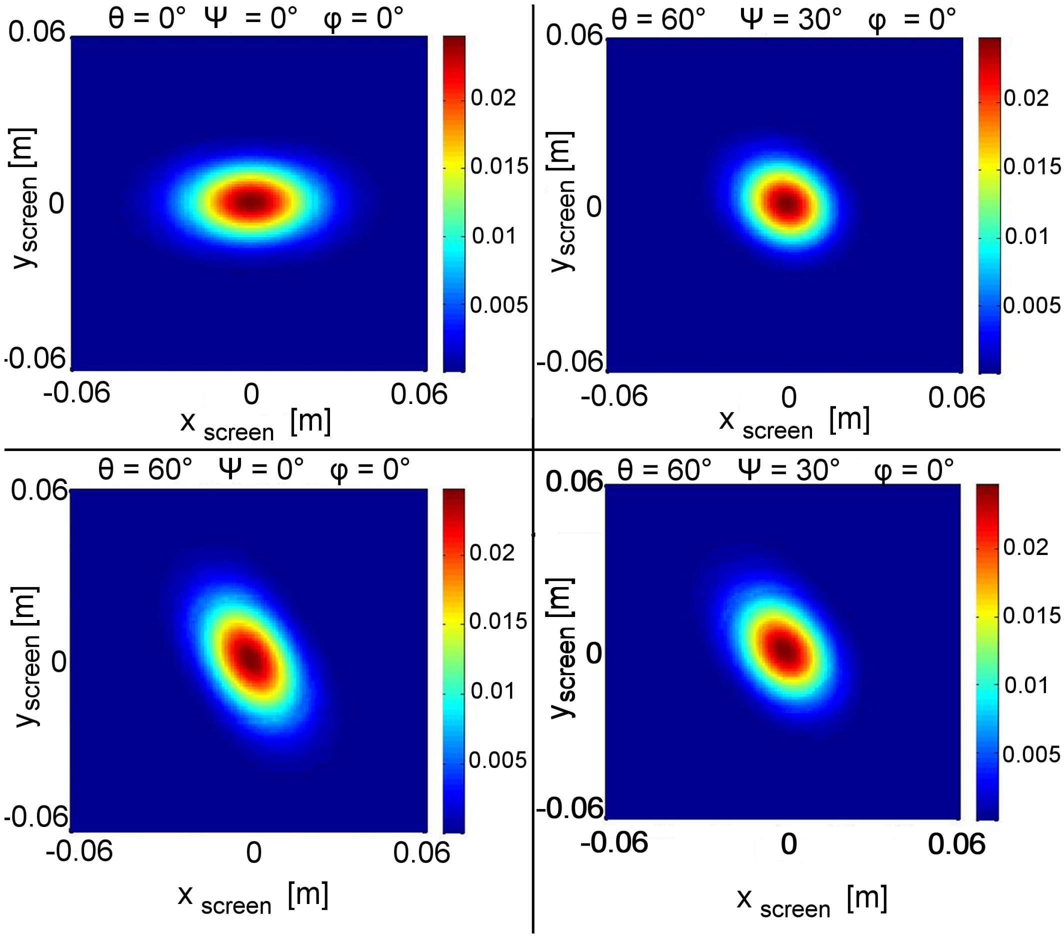

Figure 4.1: Intensity distributions recorded on the detector plane for different orientation of benzene molecule. In the upper left corner is the distribution obtained for the reference orientation, i.e the plane of the benzene ring is parallel to the incident X-ray beam.

Since the glutathione molecule reveals no symmetry, neither is it planar nor linear, one might expect the intensity distributions to be more distinguishable than in the case of benzene. Indeed, as shown in figure 4.2, nonzero values of ψ and φ angles are associated with speckles accompanying the global, Gaussian shaped, intensity maximum, and thus contributing to the uniqueness of the intensity distribution. On the other hand, the changes in the θ angle do not always lead to the same tilt of the Gaussian for different ψ and φ angles. This, however, should not influence the distinction between the intensity distributions for different orientations, as the afore mentioned speckles seem to give rise to the differences in the first place.

The distinction between the intensity distributions for different orientations has a direct influence on the posterior probability landscape. Judging by shown intensity distributions, one might expect that the probability distributions have a more pronounced maximum for glutathione compared to benzene. In the latter case one would additionally expect to observe several maxima corresponding to equivalent structures due to the rotational symmetry of the aromatic ring.

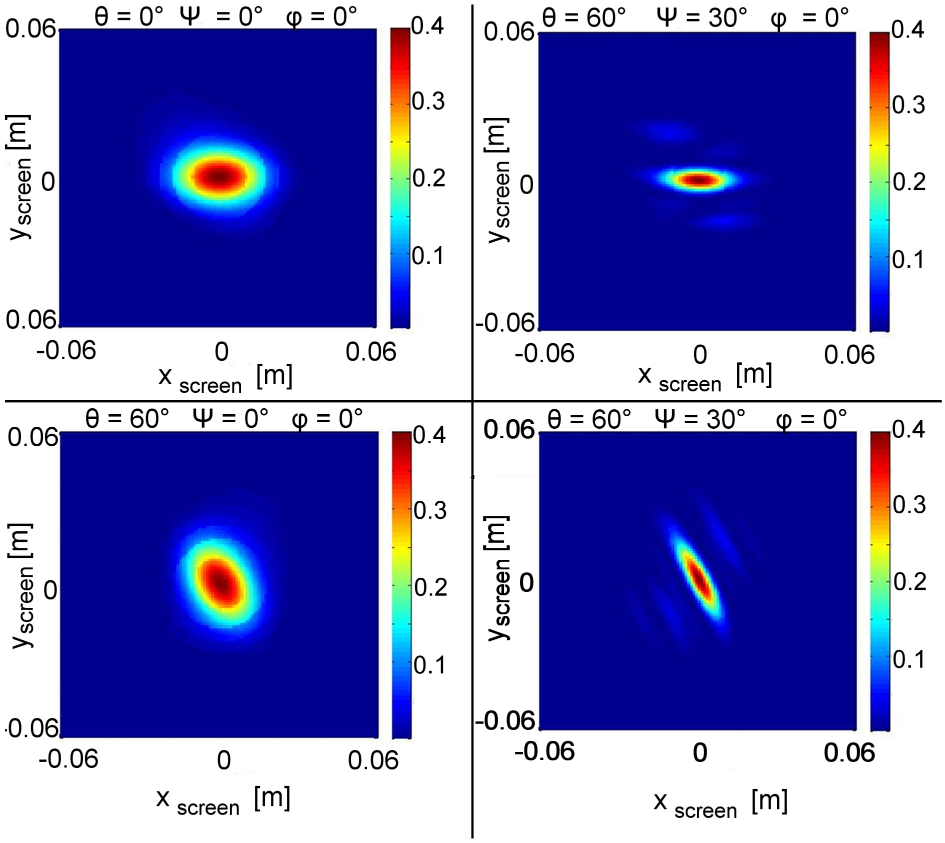

Figure 4.2: Intensity distributions recorded on the detector plane for different orientation of glutathione molecule. In the upper left corner is the distribution obtained for the reference orientation, i.e the molecule is parallel to the incident X-ray beam.

4.2 Posterior probability distributions

The accuracy in estimating the orientation of a molecule, based on the obtained diffraction pattern, depends on the sharpness of the posterior probability distribution maximum. This, according to the formula 2.16, is influenced both by the uniqueness of the intensity distributions for different orientations and the number of scattered photons. The shape of the posterior probability distribution is also affected by the presence of Poisson shot noise and the background noise.

The influence of the shape of the molecule on the intensity distributions, and thereby on the posterior probability distribution, is seen in the example of the benzene molecule.

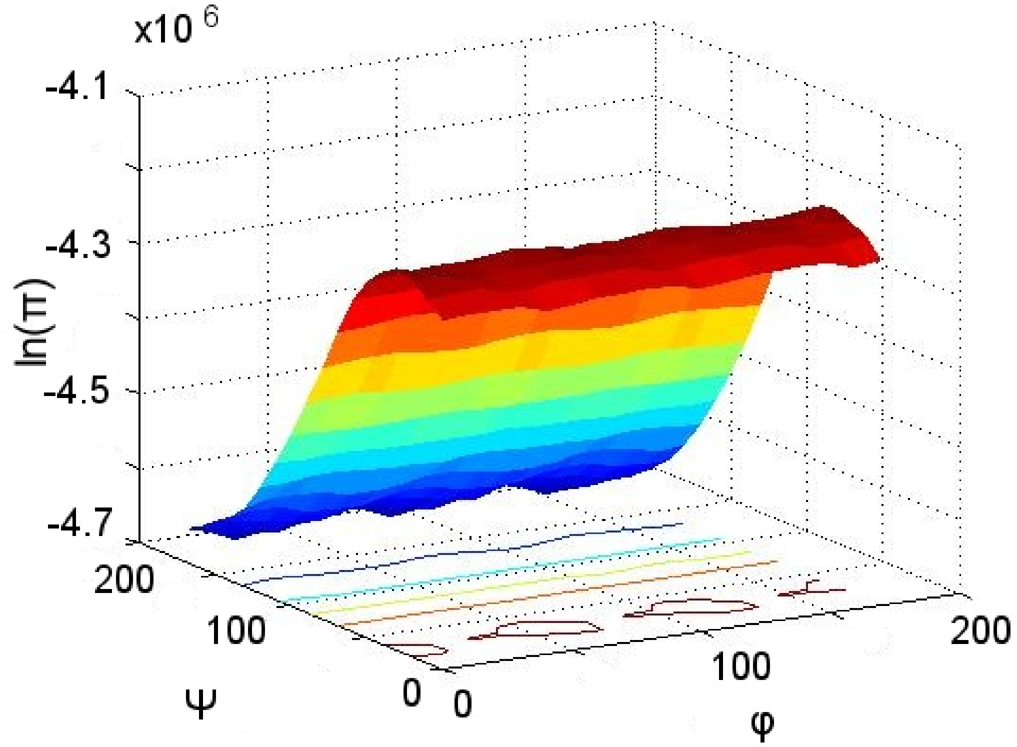

Figure 4.3: Slice at $\theta = 30°$ through the posterior probability distribution for benzene molecule oriented as follows: $\theta = 30°, \psi = 30°, \varphi = 0°$. Calculated for 500,000 scattered photons registered on a 121 x 121 pixel detector.

The symmetry of the benzene ring clearly affects the shape of the posterior probability distribution. That influence is manifested by the four maxima spaced at 60° along the φ axis. Furthermore, the landscape along that axis is relatively flat, making accurate estimation of the orientation questionable. Hence, to avoid loss of generality in testing the performance of proposed method the rest of research was carried out using the unsymmetrical tripeptide.

Using the glutathione molecule I intended to investigate the influence of noise on the posterior probability distribution. The lack of symmetry clearly influences the shape of the posterior probability distribution. Unlike for the benzene molecule, in case of the tripeptide there is a well pronounced maximum in obtained landscape, the single peak is sharp within the selected $\psi\varphi$-plane, which is easily seen in the linear scale plot. Despite only about 280 scattered photons, ψ and φ coordinates of the maximum in presented slice (θ was set to the true value) agree with the true orientation. However, the height of the maximum being less than 1 (posterior probability was normalized using Chebyshev norm) indicates that the global maximum is located beyond the selected plane. This deviation of global maximum location with respect to the true orientation is caused by Poisson shot noise in the recorded diffraction pattern. The presence of the shot noise also effects the widening of the peak around the posterior probability maximum.

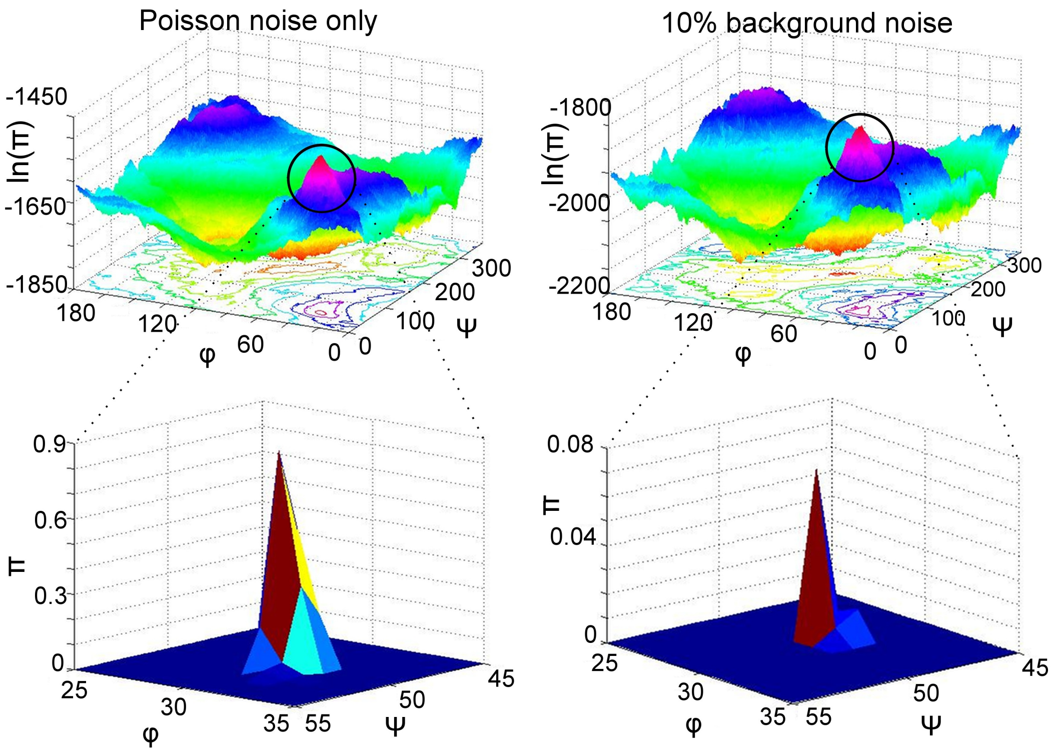

Figure 4.4: Slice at $\theta = 73°$ through the posterior probability distribution for glutathione molecule oriented as follows: $\theta = 73°, \psi = 52°, \varphi = 34°$. Calculated for 280 scattered photons registered on a 121 x 121 pixel detector.

Adding 10% background noise, with respect to the number of elastically scattered photons, to the registered diffraction pattern changes the position of the global maximum in the posterior probability distribution. The height of the local maximum on the selected slice is one order of magnitude lower than in the case with Poisson noise only. This indicates that not only is the global maximum shifted with respect to the true orientation, but also the true orientation is either not taken into account during the reconstruction process at all, or a very low weight is assigned to it. Further, due to the background noise, the local maximum in the selected plane is slightly displaced relative to the true orientation as well.

The nonzero width of the peak in the 3D posterior probability landscape, caused by the presence of noise, means that in the weighted average reconstruction method each diffraction pattern is mapped to several Ewald spheres corresponding to different orientations. Thereby a question regarding the influence of the scattered photons count on the peak width and on the resolution of reconstructed electron density arises. To answer that question I have performed bootstrap sampling from posterior distributions for a certain orientation, and computed the mean peak width for 50 diffraction patterns corresponding to the same orientation.

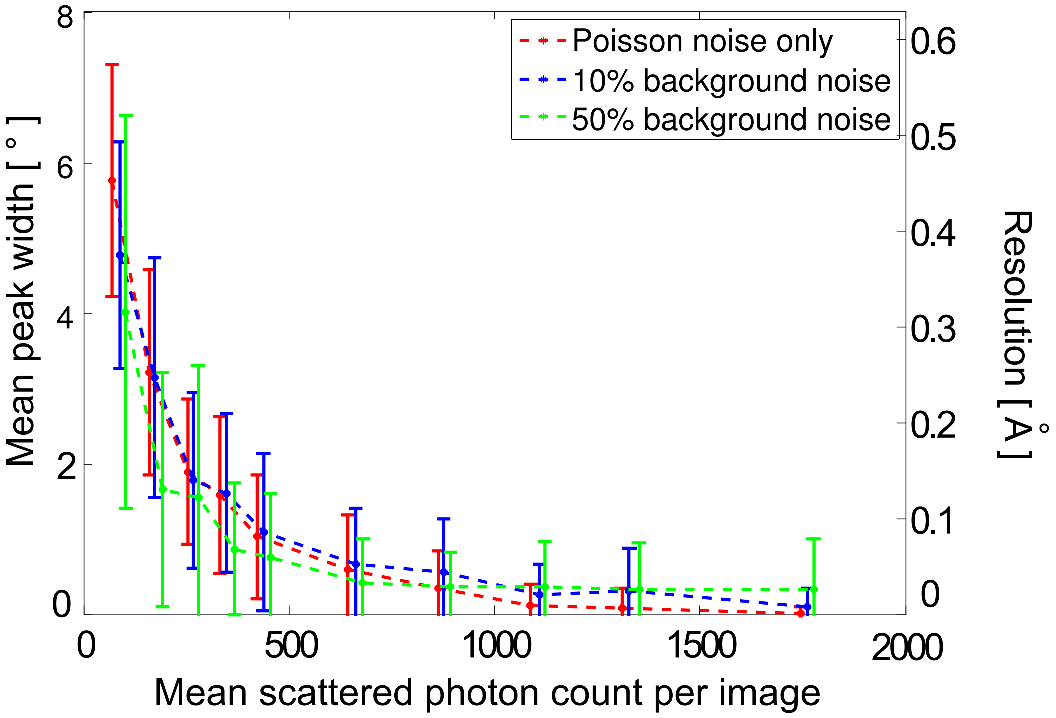

Figure 4.5: Estimate of the mean peak width and resolution for a single diffraction pattern at different levels of background noise. Points are artificially shifted for a clearer view.

As expected, increasing number of incident photons causes narrowing of the peak in the posterior probability landscape. At mean scattered photon count starting from about 500 per image, a mean peak width, comparable to the error estimation of orientation determination reported by Fung et al. [11], is achieved. The background noise does not seem to influence the shape of the dependence of mean peak width on the photon count.

4.3 Performance of the reconstruction methods

As seen in the previously presented posterior probability landscapes, the presence of the noise causes dislocation of the peak with respect to the true orientation, additionally the peak has a nonzero width. Thus one might expect the 'maximum likelihood' method perform worse than the 'Bayes' method.

Conducted simulations prove that assumption right (see figure 4.7). While both proposed methods reconstruct the molecular transform well in the low wave vectors regime, the 'maximum likelihood' method fails in the high k-vectors regime. Since only the high k-values regions carry high resolution information about the electron density, electron density function reconstructed with the 'Bayes' method is better resolved. In the high k-values regions of the reconstruction profile of the 'maximum likelihood' method intervals with zero value are present, whereas corresponding intervals of the reference profile have nonzero values, which leads to structural information loss. In the 'Bayes' method no such regions are present, because due to assigning a weight to different orientations for a single diffraction image a better coverage of the reciprocal space with Ewald spheres is achieved.

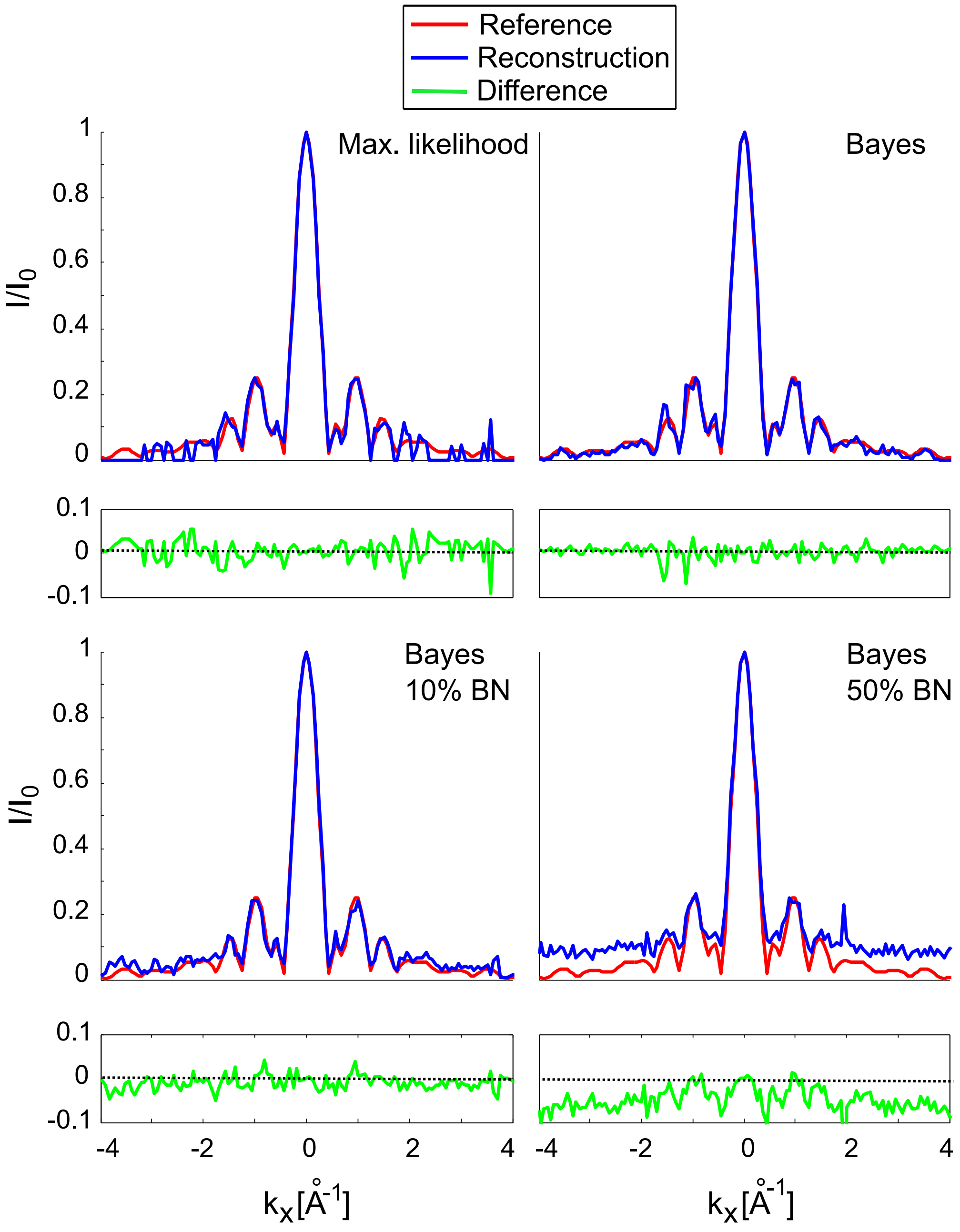

Figure 4.6: Profiles of reconstructed Fourier transformed electron density along k_x axis. Calculated for 20,000 diffraction images containing on average about 280 scattered photons per picture. The upper two profiles were reconstructed from diffraction images containing Poisson shot noise only.

Inclusion of the background noise is manifested in the high k-values regions by a vertical shift of the baseline of reconstruction profile with respect to the reference. Despite that shift, at background noise level of 10% (relative to the mean scattered photon count), the shape of the reference profile is still recognizable in the reconstructed profile. With 50 % level, however, the background noise causes too much distortion for the high resolution structural information to be retrieved from the reconstructed molecular transform.

The profiles of reconstructed molecular transform are helpful for understanding the results of electron density retrieval. The well reconstructed low k-values regions (both methods with Poisson noise only) contribute to the good agreement of the overall shape of the electron density. The loss of high resolution structural information in the 'maximum likelihood' method becomes evident in worse resolved details in the retrieved electron density, as compared to the 'Bayes' method.

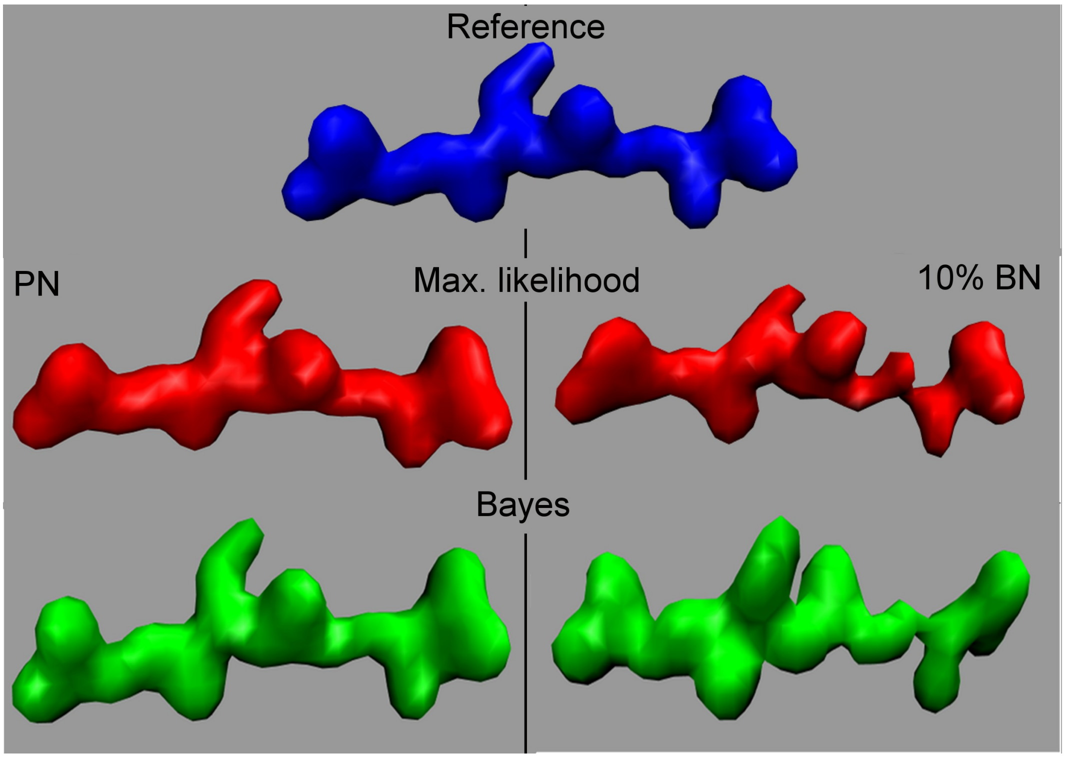

Figure 4.7: Comparison of retrieved electron density. Left side: Poisson noise only; right side: 10% background noise.

With added background noise reconstruction of the electron density becomes problematic. Subtracting a heuristically chosen number from the reconstructed molecular transform is required prior to application of phase retrieval algorithm, so as to reduce the vertical shift shown in the profiles in the k-space. Even at the level of 10%, the background noise is not averaged out in the reconstructed molecular transform, so that it also distorts the retrieved electron density. It is still possible to recognize the overall shape of the molecule, but part of the detailed information is lost. The 'maximum likelihood' method is more vulnerable to loss of the high resolution structural information.

4.4 De novo structure prediction

Several factors have an impact on the outcome of the MC based structure prediction method. Such parameters as intensity of the incident beam and number of provided diffraction images influence the posterior probability distribution of a structure given a set of diffraction images. With increasing number of included images the most probable structure gets closer to the reference. I have performed several simulations, and the presented results were obtained for 400 diffraction images with mean photon count of abut 280 photons per picture, being the optimal values yielding satisfactory outcome.

With fixed dihedral angles within the glutamic acid residue of the glutathione molecule, the posterior probability, or the energy, landscape is a 4D one. It is rugged and steep. Figure 4.8 shows a 2D slice of energy landscape close to the global minimum. With two dihedrals fixed to optimal values, an energy barrier is present in the shown landscape. It might be possible for a MC simulation to get trapped in the low energy region for values of φ_2, ψ_2 around $(78°, -20°)$. The steepness of the landscape increases with the number of provided diffraction images. Depending on the starting structure the simulation might get trapped in a local energy minimum. I have observed such a behaviour for a starting structure 'far' from the reference, i.e. with a root mean square deviation (RMSD) of 2.15 Å.

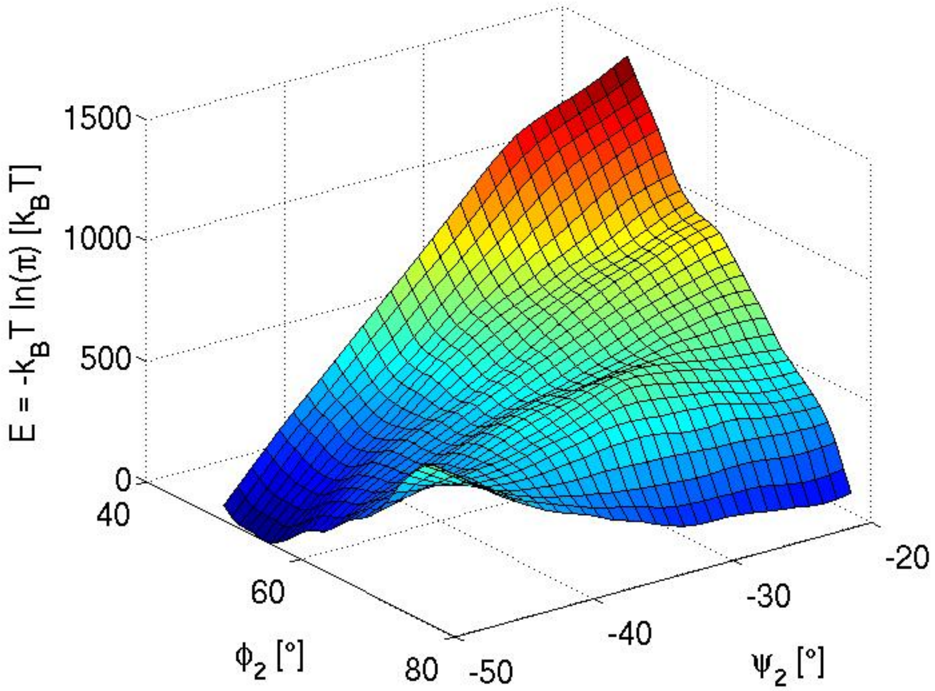

Figure 4.8: Energy landscape as a function of the dihedral angles in the cysteine residue (see Figure 3.3). The global minimum is not shown.

However, starting from structures 'close' to the reference (with RMSD values of about 1.45 Å) results in a prompt convergence. Two MC runs with different starting structures have yielded almost identical end structures after about 1,600 MC steps (40 accepted MC steps).

The resulting structure matched closely to the reference one, with a RMSD value of 0.8 Å.

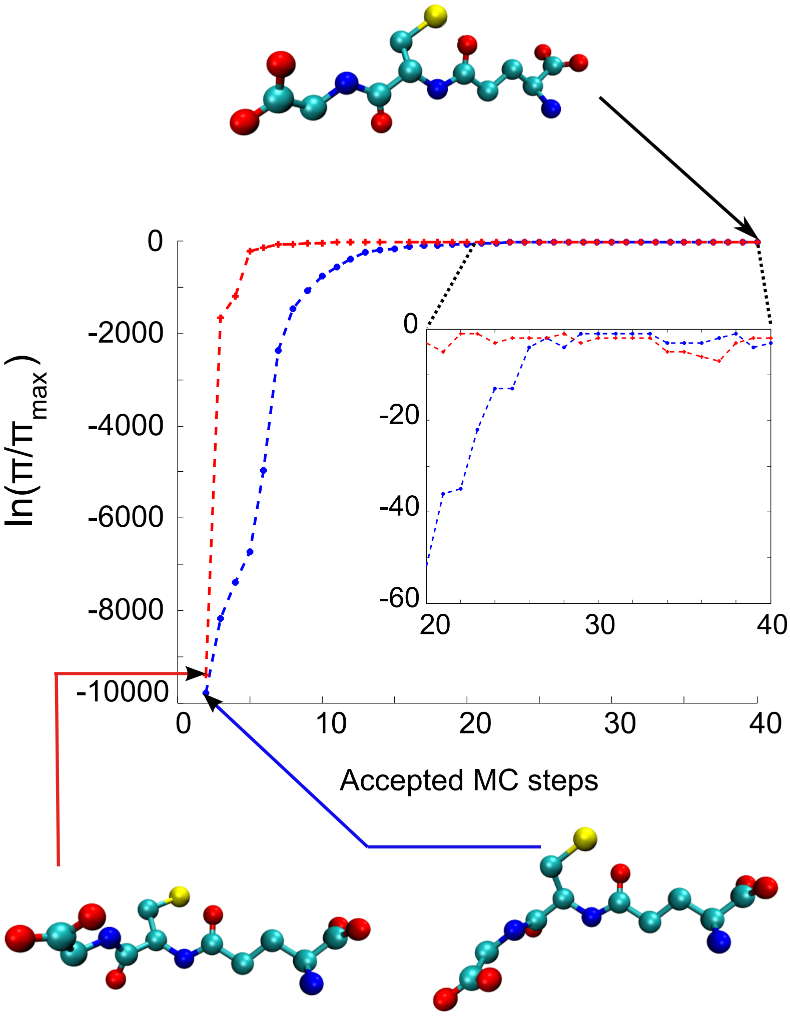

Figure 4.9: Two MC runs with a total length of 1,600 MC steps for random starting structures. Both of them converge quickly to almost the same end structure.

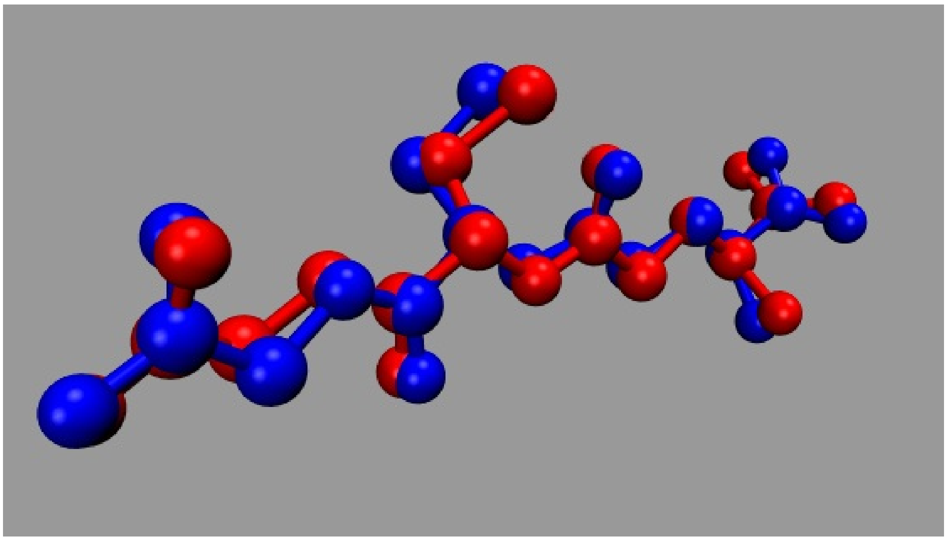

Figure 4.10: Comparison of RMSD aligned structures: blue - reference, red - final structure from MC simulation (RMSD value of 0.8 Å).

Chapter 5

Conclusions

In this project a Bayesian based approach to structure reconstruction from single molecule scattering data has been studied. The 'maximum likelihood' and 'Bayes' methods, both requiring a model input structure, are fundamental to the MC-based method, which finds the most probable structure. A reconstruction method is required to handle sparse and noisy diffraction patterns. This study shows that it is possible to recover the structure of a biomolecule from diffraction images with very low photon count and affected by Poisson and background noise.

The 'maximum likelihood' and 'Bayes' methods average the provided set of the diffraction pattern in the 3D reciprocal space by determining the orientation of the molecule for each of the diffraction patterns. The 'maximum likelihood' method uses the position of the maximum of the posterior distribution function as the orientation estimate, thus it is vulnerable to information loss. The 'Bayes' method, in contrast, treats the posterior probability distribution as a weighting function for the orientations, as a result the high resolution regions in the reciprocal space are better sampled, compared to the 'maximum likelihood' method.

I have observed that the shape of the posterior probability landscape is influenced by several factors. Rotational symmetry of molecules is manifested in the landscape by the presence of multiple maxima corresponding to the equivalent orientations. Additionally, the landscape is shallower compared to the one obtained for non symmetric molecules. The shape of the posterior probability landscape is also affected by the number of registered photons and the level of noise. While with increasing numbers of photons the peak of the distribution becomes narrower, for nerly infinite numbers of incident photons the posterior probability distribution resembles a delta function centered at the true orientation, the presence of the background noise causes dislocation and broadening of the peak. For numbers of scattered photons used in this numerical study, which were still larger than the ones one could expect for such small molecules in the real XFEL experiments, the Poisson shot noise alone caused broadening and dislocation of the global maximum, and including background noise enhanced that effect. Unlike the 'maximum likelihood' method, the 'Bayes' method reconstructs the molecular transform without any structural information loss, though it is sensitive to the background noise. However, using a larger number of diffraction patterns might reduce the background noise due to better averaging. On the other hand, structure reconstruction of molecules with larger scattering cross sections will not be affected by the background noise, thus applying the 'Bayes' method should yield satisfactory results.

Both 'Bayes' and 'maximum likelihood' methods require a model structure to generate the ensemble of intensity distributions necessary to compute the posterior probability distribution. In this project the same structure was used for generating the diffraction patterns and as the model structure for the reconstruction methods. An interesting question, which has not been answered here, is how much the model structure can vary from the true structure so that the output of the reconstruction method is still acceptable. An answer to that question will explain weather the 'Bayes' method can be used as a structure refinement tool.

The MC approach to structure determination was derived from the two other methods. Its goal is not determining explicitly the orientation for each diffraction pattern, but instead searching for the most probable structure given a set of diffraction images. Thus it does not require a model structure. However, it has to be provided with fragments with known internal structure, so as to determine their relative orientation. For simple polypeptides these fragments can be single amino acids, or subunits in case of larger proteins.

The energy landscape, whose dimensionality depends on the number of angles needed for description of the entire structure, is steep and rugged, thus sampling problems may arise. Whilst starting from conformations close to the reference structure enables recovering the structure, in other cases simulation might become trapped in a local energy minimum. Therefore, one might consider applying replica exchange [14, 31] to improve the convergence of this method. An important parameter influencing the energy landscape is the number of included diffraction patterns. It has to be chosen carefully, too small causes a large difference between the most probable and the reference structure, whereas too large unnecessarily prolongs the computation time of a single MC step.

The outcome of the 'Bayes' reconstruction method is the molecular transform, thus the atomic positions have to be computed from the retrieved electron density. Using the phase retrieval algorithm gives rise to an additional bias. Unlike the 'Bayes' method, the MC based structure retrieval yields explicit atomic structure, hence it doesn't suffer from phase retrieval errors.

The computational effort of the 'Bayes' method scales with the number of diffraction patterns, their size and the sampling step size of the posterior probability distribution. The number of atoms, however, does not influence much the computational time, because the Fourier transform of the electron density of the model molecule is computed only once at the beginning. Since in each MC step the electron density of the proposed structure is Fourier-transformed, the computational time of the MC based method scales with the number of atoms.

Use of proposed approaches depends on the formulation of the structure determination problem. The 'Bayes' approach could be used for structure refinement, whereas the MC structure determination method might be used for determining the relative orientation of fragments of biomolecules. B conducting numerical experiments, I have shown that those methods are capable of working with sparse and noisy data. However, the performance of the methods needs to be improved.

Chapter 6

Acknowledgments

I would like to thank my supervisor Prof. Dr. Helmut Grubmüller for introducing me to this very interesting subject, discussions and valuable advice. I am also grateful to Prof. Dr. Marcus Müller for his suggestions regarding my thesis. Further, I would like to thank my colleagues from the Department of Theoretical and Computational Biophysics, in particular Dr. Carsten Kutzner for helping me to debug my source code, Christian Blau for discussions and proofreading, Carl Burmeister and Stephanus Fengler for valuable hints, and helping me getting started.

Bibliography

[1] J.O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

[2] D. Bilderback. Review of third and next generation synchrotron light sources. *Journal of physics. B, Atomic, molecular, and optical physics*, 38(9):S773–S797, 2005.

[3] W.M. Bolstad. *Introduction to Bayesian statistics*. Wiley, 2004.

[4] I.N. Bronstein, K.A. Semendjajew, G. Musiol, and H. Mühlig. *Taschenbuch der Mathematik*. Deutsch Harri GmbH, 2008.

[5] S.P. Brooks and B.J.T. Morgan. Optimization using simulated annealing. *The Statistician*, 44(2):241–257, 1995.

[6] C.R. Cantor and P.R. Schimmel. *Biophyiscal chemistry Part II: Techniques for the study of biological structure and function*. WH Freeman & Co, 1980.

[7] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of statistics*, 7(1):1–26, 1979.

[8] V. Elser. Solution of the crystallographic phase problem by iterated projections. *Acta Crystallographica Section A: Foundations of Crystallography*, 59(3):201–209, 2003.

[9] J. Feldhaus, J. Arthur, and J.B. Hastings. X-ray free-electron lasers. *Journal of physics. B, Atomic, molecular, and optical physics*, 38(9):S799–S819, 2005.

[10] J.R. Fienup. Reconstruction of an object from the modulus of its fourier transform. *Optics letters*, 3(1):27–29, 1978.

[11] R. Fung, V. Shneerson, D.K. Saldin, and A. Ourmazd. Structure from fleeting illumination of faint spinning objects in flight. *Nature Physics*, 5(1):64–67, 2009.

[12] KJ Gaffney and HN Chapman. Imaging atomic structure and dynamics with ultrafast x-ray scattering. *Science*, 316(5830):1444, 2007.

[13] B. Gough. *GNU Scientific Library Reference Manual*, 2009.

[14] U.H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, 1997.

[15] S.P. Hau-Riege. X-ray atomic scattering factors of low-z ions with a core hole. *Physical Review A*, 76(4):42511, 2007.

[16] G. Huldt, A. Szoke, and J. Hajdu. Diffraction imaging of single particles and biomolecules. *Journal of structural biology*, 144(1-2):219–227, 2003.

[17] S.G. Johnson and M. Frigo. The design and implementation of fftw3. *Proc. IEEE*, 93:216–231, 2005.

[18] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM transactions on modeling and computer simulation*, 8(1):3–30, 1998.

[19] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087, 1953.

[20] J. Miao, K.O. Hodgson, and D. Sayre. An approach to three-dimensional structures of biomolecules by using single-molecule diffraction images. *Proceedings of the National Academy of Sciences of the United States of America*, 98(12):6641, 2001.

[21] J. Miao, D. Sayre, and H.N. Chapman. Phase retrieval from the magnitude of the fourier transforms of nonperiodic objects. *Journal of the Optical Society of America A*, 15(6):1662–1669, 1998.

[22] R. Miles. On random rotations in r3. *Biometrika*, 52:636, 1965.

[23] M. Moakher. Means and averaging in the group of rotations. *SIAM journal on matrix analysis and applications*, 24(1):1–16, 2002.

[24] R. Neutze, G. Huldt, J. Hajdu, and D. van der Spoel. Potential impact of an x-ray free electron laser on structural biology. *Radiation physics and chemistry*, 71(3-4):905–916, 2004.

[25] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu. Potential for biomolecular imaging with femtosecond x-ray pulses. *Nature*, 406(6797):752–757, 2000.

[26] G. Oszlanyi and A. Suto. Ab initio structure solution by charge flipping. *Acta Crystallographica Section A: Foundations of Crystallography*, 60(2):134–141, 2004.

[27] D.K. Saldin, V.L. Shneerson, R. Fung, and A. Ourmazd. Structure of isolated biomolecules obtained from ultrashort x-ray pulses: exploiting the symmetry of random orientations. *Journal of physics. Condensed matter*, 21(13):134014, 2009.

[28] F. Schotte, J. Soman, J.S. Olson, M. Wulff, and P.A. Anfinrud. Picosecond time-resolved x-ray crystallography: probing protein function in real time. *Journal of structural biology*, 147(3):235–246, 2004.

[29] G.F. Schröder and H. Grubmüller. Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer experiments. *The Journal of chemical physics*, 119(18):9920–9924, 2003.

[30] V.L. Shneerson, A. Ourmazd, and D.K. Saldin. Crystallography without crystals. i. the common-line method for assembling a three-dimensional diffraction volume from single-particle scattering. *Acta crystallographica. Section A, Foundations of crystallography*, 64(2):303–315, 2008.

[31] R.H. Swendsen and J.S. Wang. Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 1986.

[32] W.F. van Gunsteren and H.J.C. Berendsen. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition in English*, 29(9):992–1023, 1990.

[33] L. Young, E. Kanter, B. Krassig, Y. Li, A . March, and S. Pratt. Femtosecond electronic response of atoms to ultra-intense x-rays. *Nature*, 466(7302):56–61, 2010.