Bayesian structure reconstruction from single molecule X-ray scattering data

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

im Promotionsprogramm ProPhys

der Georg-August University School of Science (GAUSS)

vorgelegt von

Michał Walczak

aus Warschau

Göttingen, 2014

Betreuungsausschuss:

Prof. Dr. Helmut Grubmüller, Abteilung für Theoretische und Computergestützte Biophysik, Max-Planck-Institut für biophysikalische Chemie Göttingen

Prof. Dr. Marcus Müller, Institut für Theoretische Physik, Georg-August-Universität Göttingen

Mitglieder der Prüfungskomission:

Referent:

Prof. Dr. Helmut Grubmüller, Abteilung für Theoretische und Computergestützte Biophysik, Max-Planck-Institut für biophysikalische Chemie Göttingen

Koreferent:

Prof. Dr. Marcus Müller, Institut für Theoretische Physik, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskomission:

Dr. Jochen Hub, Abteilung für Molekulare Strukturbiologie, Georg-August-Universität Göttingen

Prof. Dr. Sarah Köster, Institut für Röntgenphysik, Georg-August-Universität Göttingen

Prof. Dr. Hans-Ulrich Krebs, Institut für Materialphysik, Georg-August-Universität Göttingen

Prof. Dr. Simone Techert, Forschungsgruppe Strukturdynamik (bio)chemischer Systeme, Max-Planck-Institut für biophysikalische Chemie Göttingen

Tag der mündlichen Prüfung: 31.10.2014

Vorveröffentlichungen der Dissertation

Teilergebnisse dieser Arbeit wurden im folgenden Beitrag veröffentlicht:

M Walczak and H Grubmüller. Bayesian orientation estimate and structure information from sparse single molecule x-ray diffraction images. *Phys. Rev. E*, 90(2):022714, 2014.

Contents

1 Introduction

Every component of a biological system has its function, knowledge of which is crucial to understand how living organisms work. The function is affected by the dynamics of the component's building blocks (e.g. proteins, lipid membranes etc.), and the dynamics itself depends on the structure of those components [1]. Hence structure determination techniques providing high resolution information are important for understanding biological systems. Structures of biomolecules at atomic resolution are essential to study in detail biological processes with, e.g. molecular dynamics (MD) simulations, which help to explain underlying mechanisms [2, 3].

One such high resolution structure determination technique is X-ray crystallography. However, though a powerful tool, it faces certain limitations. First, this technique is applicable to crystalline samples only. About 40% of all biomolecules cannot be crystallized [4], e.g. certain membrane proteins, and thus are inaccessible to X-ray crystallography. Even for those biomolecules that do form crystals it still might be a tedious process to purify and obtain a well diffracting sample. Furthermore, copies of molecules in individual crystal cells may adopt different conformations. Diffraction images of such samples reflect an average distribution of atomic positions over possible conformations, thus detailed information on molecular subregions is lost. Therefore, due to the structure inhomogeneity it is challenging to obtain high resolution structures of large biomolecules. Spatial resolution aside, temporal resolution achieved for crystals in Laue diffraction experiments at synchrotron sources is limited to $\sim$ 100 ps scales only [5].

The other problem in X-ray crystallography results from registering only the intensities of discrete Bragg reflections. Missing phases have to be determined to reconstruct the electron density of the investigated protein. To circumvent the phase problem, crystallographers use methods such as multiple isomorphous replacement [6], multiwavelength anomalous diffraction [7], or molecular replacement [8].

Instead of using macroscopic crystals, recently developed X-ray sources enabled diffractive imaging of smaller crystals with sizes in the μm to nm range, and thereby solved some of the aforementioned problems. Nanocrystallography has the following advantages over traditional X-ray crystallography. Although a crystalline specimen is still required in nanocrystallography, nanocrystals are easier to grow than macroscopic crystals. Further, due to small crystal sizes, intensities between Bragg reflections are present in diffraction images of nanocrystals, and thus oversampling techniques can be exploited to help retrieving the missing phases [9]. In 2011, Chapman et al. [9] determined the structure of Photosystem I (a membrane protein) at 8.5 Å resolution from nanocrystals exposed to a hard X-ray free electron laser (XFEL) beam. Recently, Barends et al. [10] reported a 2.1 Å structure of a lysozyme determined *de novo* from microcrystals. These experiments confirmed that the use of XFEL sources enables diffractive imaging of samples much smaller than macroscopic crystals.

The key advantage of XFEL beams is their very high intensity; currently available beams deliver $\sim 2.3 \times 10^{12}$ photons per pulse focussed into a 10 μm spot [11]. However, this high intensity causes a tremendously increased radiation damage. Instead of being distributed over many copies of a molecule in a macroscopic crystal, the radiation is absorbed by few molecules in a nanocrystal, or in the most extreme case, by only one molecule. Thus every atom within the irradiated molecule absorbs multiple photons, and as a result loses electrons in the core shell photo-ionization process and subsequent Auger decay. The increasing positive charge of the molecule leads to a Coulomb explosion of the sample [12]. Therefore, it is important that a diffraction image is recorded *before* the radiation damages the illuminated molecule. To avoid imaging the disrupted electron density, femtosecond pulses are necessary. Exposure times in the femtosecond regime will additionally provide high temporal resolution, advantageous for studying conformational dynamics, e.g. during enzymatic reactions [13].

One may ask if it is possible to go even one step further and perform XFEL scattering experiments on *single* molecules, such that no crystals have to be grown at all. Such experiments have the potential to overcome the limitations of crystallography [12–14], save for the purification process. Unlike crystals, single molecules lack translational symmetry, hence they generate continuous diffraction patterns that enable oversampling. Iterative phasing algorithms allow to determine the missing phases from the registered intensities, such that additional constrains are satisfied, and thus retrieve the electron density of the irradiated molecule [15–20]. In the single molecule experiments, molecules are injected into the XFEL beam, e.g. by applying electrospraying techniques, such that during one pulse a diffraction image of only a single molecule can be recorded on the detector. Due to high repetition rates, many images ($10^2 \ldots 10^6$) are obtained. However, each image contains information from only very few elastically scattered photons (of the order of $10 - 10^4$, depending on molecule size and beam intensity) along with substantial noise [18]. Further, and most importantly, molecules entering the beam can rotate freely and assume a random orientation during the exposure. This unknown orientation together with partial structural information is encoded in the recorded diffraction image.

The goal of single molecule XFEL scattering experiments is to determine the structure of the investigated molecule at the highest possible resolution. To achieve this goal, the structural information has to be extracted from sparse diffraction images of a randomly oriented molecule, therein lies the biggest challenge for the prospective structure determination algorithms. Recent calculations showed that a 500 kDa protein scatters only about 4×10^{-2} photons per pixel in the high resolution part of a diffraction image [18]. Such low photon counts reflect a low structural information content of a single diffraction image. The partial structural information results not only from few photon counts, but also from the fact that an image on the detector plane is a 2D projection of a 3D molecular transform obtained from an unknown orientation. Therefore, many diffraction patterns from different orientations are required to fully sample a 3D intensity distribution of the irradiated molecule. Structure determination methods from single molecule XFEL scattering images proposed so far either aim at accurate orientation determination for individual diffraction images and averaging those in 3D reciprocal space, or recovering the structure from intensity correlations and thus omitting the orientation determination. I will now discuss selected methods belonging to the two classes: orientation determination, and correlation based methods.

One of the earliest methods based on orientation determination was the 'common line' method by Huld et al. [21]. This 'common line' refers to a curve in reciprocal space formed by two intersecting Ewald spheres, which correspond to two different diffraction images. Identifying intersection curves of any three diffraction patterns suffices to calculate their relative orientations. However, due to very low photon counts, the images need to be clustered and averaged beforehand, such that enough signal photons for locating the 'common lines' are available. The images are grouped according to cross-correlation function between any two of them, provided that a mean photon count per pixel exceeds 10. Because this threshold value is three orders of magnitude higher than expected in XFEL experiments, it will be very difficult to use the 'common line' method under those conditions [18].

An alternative method for sorting diffraction images into orientational classes and subsequent averaging in 3D reciprocal space was described by Fung et al. [22]. The authors suggest to determine the most likely orientation for every diffraction pattern using generative topographic mapping, which, apart from the images themselves, requires the dimensionality of the orientational space as the only input. However, clustering the diffraction images and averaging them within orientational classes might cause information loss due to insufficient sampling of high resolution regions in 3D reciprocal space, as will be shown in the Results section. Further, the required mean number of elastically scattered photons per picture of about 100 (excluding the central pixels of the detector protected by the beamstop) seems rather high, especially for small molecules. A similar manifold embedding method was proposed recently by Giannakis et al. [23]. A projection from 3D reciprocal space to a 2D diffraction image results in object independent symmetries in those images. This fact is exploited to navigate through the manifold created from the recorded images and determine their relative molecular orientations.

Loh and Elser proposed a method that maximizes the likelihood of an intensity distribution model in reciprocal space to fit a set of diffraction images [24]. This expansion-expectation maximization-compression (EMC) approach uses Bayes' theorem to determine the orientation for individual diffraction image from the intensity model, which is updated by averaging the aligned images in 3D reciprocal space in each iteration. This method was applied to determine the structure of a GroEL (heat shock 60 kDa protein) molecule at 2 nm resolution from up to 10^6 synthetic diffraction images. A similar, though less computationally demanding, algorithm was proposed by Tegze and Bortel [25].

All above methods show how challenging it is to determine the orientation for *individual* sparse diffraction images. As will be shown later, the Bayesian formalism performs this task promisingly despite low photon counts in recorded images.

A second class of structure determination methods circumvents the orientation determination for individual images. Instead, the diffraction intensities in reciprocal space are determined from cross-correlations between diffraction images. The intensities are then expressed in a spherical harmonics basis. Saldin et al. [4] used such an approach to determine a molecular shape. However, the achievable level of detail remains unclear;

also low photon counts in registered patterns might limit the application of this method, similarly to the 'common line' method.

Alternatively, Liu et al. [26] proposed to refine a low resolution electron density model using an angular correlation function of multiple diffraction images. As in approaches typical of small-angle X-ray scattering experiments, the electron density represented on a grid is, in every Monte Carlo step, locally perturbed by a random dilation or an erosion. Resulting intensity correlations are then compared with those from the experimental data. This approach allows for structure determination from diffraction images of many randomly oriented copies of the same molecule.

Starodub et al. [11] devised an alternative correlation based method. The authors calculated an electron density map of two polystyrene spheres with a 91 nm diameter at 20 nm resolution using partial triple correlation of intensity distributions. This approach was applied to a molecule with cylindrical symmetry, thus reducing the complexity of correlation calculations but also the generality. However, a structure determination with full correlation analysis should enable solving high resolution structures of molecules lacking any symmetry [27]. All these correlation based methods have so far been applied to recover low resolution structures. As in the case of the orientation determination approaches, low photon counts in diffraction images might also pose a challenge for the above methods.

In this work, I aimed at developing a structure determination method that extracts high resolution structural information even, in extreme cases, from very sparse and noisy XFEL diffraction images of single molecules. To this end, I proposed two complimentary Bayesian approaches to structure determination at atomic resolution from such images, as depicted in Fig. 1.1. These approaches are referred to as 'Orientational Bayes' and 'Structural Bayes', respectively. In the Orientational Bayes approach, the probability of a molecular orientation $\mathbf{\Theta}$ given a diffraction pattern $\mathbf{X}$, $\pi(\mathbf{\Theta}|\mathbf{X})$, is calculated for every recorded image, and used to align the images in 3D reciprocal space. By contrast, in the Structural Bayes approach, the molecular orientation is not determined for individual images; instead, the probability of a model structure S to give rise to the *entire* recorded set of diffraction patterns $\{\mathbf{X}\}$, $\pi\big(S|\{\mathbf{X}\}\big)$, is computed. Both approaches will be tested for their applicability to solving variously sized molecular structures, with a small tripeptide as the most challenging case, under extreme experimental conditions, such as low photon counts.

The Orientational Bayes approach is similar to the EMC algorithm [24], in the

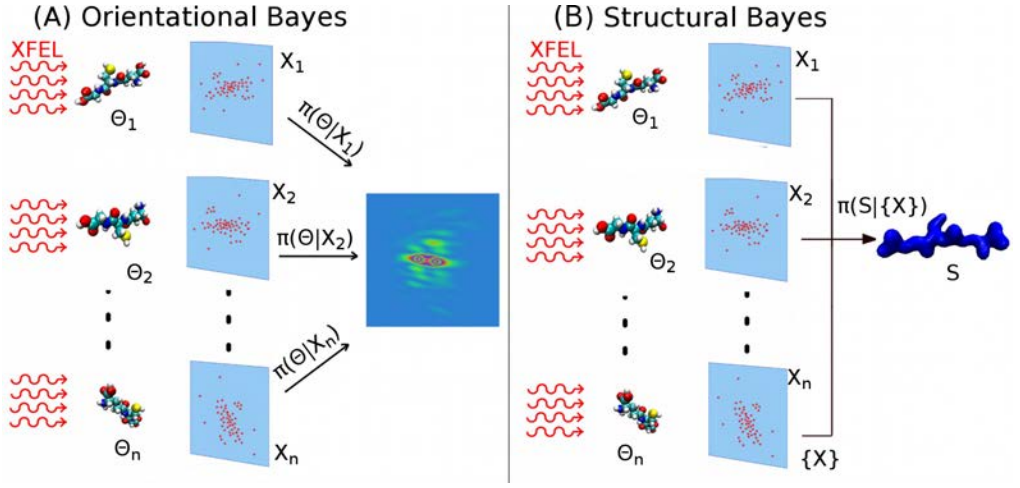

Figure 1.1: Two Bayesian approaches to structure determination from single molecule XFEL diffraction images. (A) The probability of a molecular orientation $\mathbf{\Theta}$, here of a glutathione, $\pi(\mathbf{\Theta}|\mathbf{X})$, is determined for every individual diffraction image $\mathbf{X}$. The underlying molecular transform is obtained by aligning and averaging the images in reciprocal space. (B) The probability of a structure S (defined in real space), $\pi\big(S|\{\mathbf{X}\}\big)$, is calculated to identify a structure that fits best simultaneously to all collected images $\{\mathbf{X}\}$. Figure adapted from Ref. [28].

sense that it uses a rigorous Bayesian framework for determining orientations of individual diffraction images using a 'seed' model, which is similar to the original structure, and can be obtained from, e.g. nanocrystallogrphy [9]. For each recorded image, the probability distribution, $\pi(\mathbf{\Theta}|\mathbf{X})$, determines how probable it is that, given this observed image, the 'seed' structure assumed a particular orientation. The images are aligned according to the calculated probability distributions and subsequently averaged in 3D reciprocal space yielding an improved model of the molecular transform, as illustrated in Fig. 1.1(A).

In the Orientational Bayes approach, the quality of the retrieved electron densities depends on the accuracy of the orientation determination. Therefore, it will be studied how the achievable resolution depends on molecular size, incident beam intensity, and background noise level. In particular, I will investigate whether the Bayesian orientation determination applied to simulated sparse XFEL diffraction images is accurate enough to retrieve an electron density at atomistic resolution despite background noise levels up to 50% of the average signal photon counts per image. This question has not yet been addressed by the EMC approach [24]. In contrast to the EMC algorithm that considers diffraction images in terms of photon counts per pixel and uses a Poisson approximation to calculate the likelihood of observing a diffraction pattern, here, probabilities of all individual photons in recorded diffraction patterns are used, thereby providing a more general likelihood formulation. In particular, the shot noise is directly and accurately accounted for by the likelihood defined in terms of a multinomial distribution. Also, additional background noise is straightforwardly included in the probabilities of registering individual photons.

As discussed above, the Orientational Bayes approach requires a 'seed' model. To avoid this requirement, I propose the Structural Bayes approach. In this approach, all orientations accessible to the model structure are sampled and the probability distribution is integrated over those orientations, leaving the model structure as the variable that is optimized to yield the highest probability. By maximizing the probability $\pi\big(S|\{\mathbf{X}\}\big)$, a structure that simultaneously fits best to the entire set of recorded images is found among several candidate structures. This way, I aim at distinguishing between different structure models by calculating ratios of evidence between them, which is a common practice in Bayesian model comparison [29].

Using a tripeptide as a test molecule, I investigated to what extent the Structural Bayes approach enables a *de novo* structure determination. To this aim, a structure model was refined in a Monte Carlo (MC) simulation by sampling different amino acid conformations and comparing resulting structure models with synthetic diffraction images generated from a reference residue configuration.

However, such amino acid based structural refinement may not be feasible for larger biomolecules due to the vast search space. Therefore, to address also those large biomolecules with the Structural Bayes approach, I reformulated the previous question: is it still possible to distinguish among different conformations of large structure models obtained otherwise than by exhaustive conformational sampling? To answer this question, different conformations of three immunoglobulin (Ig) domains of a titin from a constrained MD simulation were used as a limited test set to be compared with diffraction images generated from a reference conformation.

Because in large biomolecular complexes structural changes happen at various length scales, I also investigated if those changes can be detected using the Structural Bayes approach; in particular, if local structural changes are traceable against a large structural background. To this aim, cryo-electron microscopy (cryoEM) derived ribosomal structures of seven different translocation states were used as a test set to check if the reference state can be identified by calculating the probability of a structure given a set of images. In the ribosome, tRNA chains constitute only a small fraction of the entire complex. Hence their movement during the translocation process results in local structural changes against the large structural background of ribosomal subunits.

Finally, to reduce computational time, a multiscale structural model is introduced. I investigated if it is possible to distinguish among structures with important regions modelled at atomic resolution and the remaining parts described at a lower resolution as coarse grain (CG) beads. CG modelling is commonly used in MD simulations to study molecular processes of large systems at relevant time scales, e.g. vesicle fusion [30], because it reduces the computational cost.

When addressing the above questions, I found that the two proposed Bayesian approaches can indeed serve to determine molecular structures from single molecule XFEL diffraction data. The results presented in this work suggest that these approaches are able to extract structural information at atomic resolution from sparse and noisy diffraction images of single biomolecules of different sizes (e.g. with molecular masses ranging from several hundred Da to MDa). Further, the Structural Bayes approach should also be able to trace structural changes happening at multiple length scales in complex molecules, including localized structural changes against a large structural background. This feature might be useful in studying biological processes such as enzymatic reactions or ligand binding.

2 Theory

This chapter introduces the theoretical concepts of my project. First, I will focus on the basics of X-ray scattering theory and Bayesian analysis that were used for developing the two structure determination approaches sketched in the introduction. Then, I will describe the Bayesian framework applied for orientation determination and model comparison based on structure probability.

2.1 X-ray scattering

In X-ray scattering experiments, matter, mostly built of light atoms, interacts in three major ways with radiation. The most important type of interaction for diffraction imaging is the elastic scattering. In that process, the photon energy is conserved and only the momentum direction alters. Elastically scattered photons by the target molecule interfere coherently and form diffraction patterns and thereby convey information about the atomic structure. The most dominant interaction for high energy XFEL beams, however, is the photoelectric effect, during which photon absorption leads to a core shell ionization in most cases, followed by subsequent Auger decay. For 12 keV (1Å wavelength) photons, as planned in single molecule scattering experiments to achieve atomic resolution, the photoelectric cross-section of a carbon atom is approximately 10 times larger than the elastic scattering cross-section [12]. In the third possible event, a photon is inelastically scattered (Compton scattering). In fact, the photoelectric cross-section is about 33 times larger than the inelastic scattering cross-section for a carbon atom [12]. Photons that transfer part of their momentum to bound electrons during inelastic scattering contribute to background noise in diffraction images.

All these processes can be described using perturbation theory and Hamiltonian nonrelativistic quantum electrodynamics. In particular, the derivation of an elastic scattering cross section not only reveals the link between diffraction pattern and electron density function, but also explains how the phase problem in diffractive imaging arises. Contents of this section are based on the full derivation by R. Santra in his paper [31].

Here, I will only focus on certain steps and assume atomic units: reduced Planck constant $\hbar = 1$, electron mass $\mathrm{m_e} = 1$, and the speed of light $\mathrm{c} = 1/\alpha$, where α is the fine-structure constant.

The total Hamiltonian describing a molecule interacting with an electromagnetic field has three major components

$$\hat{H} = \hat{H}_{\mathrm{mol}} + \hat{H}_{\mathrm{EM}} + \hat{H}_{\mathrm{int}}, \tag{2.1}$$

where $\hat{H}_{\mathrm{mol}}$ is the molecular Hamiltonian, $\hat{H}_{\mathrm{EM}}$ is the Hamiltonian for the free electromagnetic field, $\hat{H}_{\mathrm{int}}$ describes the interaction between photon and electron fields. The molecular Hamiltonian comprises of the nuclear kinetic energy term, the nucleus-nucleus repulsion term, and the electronic Hamiltonian. However, for further considerations, nuclei movement is neglected, which is justified by anticipated pulse lengths in the femtosecond regime. The vector potential $\mathbf{A}$ describing the electromagnetic field, expanded in plane waves in a box of volume V, is given by an operator

$$\hat{\mathbf{A}}(\mathbf{r}) = \sum_{\mathbf{k},\lambda} \sqrt{\frac{2\pi}{V\omega_k \alpha^2}} \left(\hat{a}_{\mathbf{k},\lambda}\epsilon_{\mathbf{k},\lambda}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \hat{\mathrm{a}}^{\dagger}_{\mathbf{k},\lambda}\epsilon_{\mathbf{k},\lambda}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\right), \tag{2.2}$$

where $\mathbf{k}$ is a wave vector, $\omega_{\mathrm{k}} = \frac{|\mathbf{k}|}{\alpha}$ is corresponding angular frequency, $\epsilon_{\mathbf{k},\lambda}$ is a polarization vector with $\lambda = 1, 2$, α is the fine-structure constant, $\hat{a}^{\dagger}_{\mathbf{k},\lambda}$ and $\hat{a}_{\mathbf{k},\lambda}$ are creation, annihilation operators, respectively, acting on a photon in mode $(\mathbf{k}, \lambda)$. The Hamiltonian for the free electromagnetic field, in the Coulomb gauge, is then

$$\hat{H}_{\mathrm{EM}} = \sum_{\mathbf{k},\lambda} \omega_k \hat{a}^{\dagger}_{\mathbf{k},\lambda}\hat{a}_{\mathbf{k},\lambda} + \sum_{\mathbf{k},\lambda} \omega_k/2. \tag{2.3}$$

The interaction Hamiltonian reads

$$\hat{H}_{\mathrm{int}} = \alpha \int \hat{\psi}^{\dagger}(\mathbf{r}) \left[\hat{\mathbf{A}}(\mathbf{r}) \cdot \frac{\nabla}{\mathrm{i}}\right] \hat{\psi}(\mathbf{r}) \mathrm{d}^3\mathbf{r} + \frac{1}{2}\alpha^2 \int \hat{\psi}^{\dagger}(\mathbf{r})\hat{A}^2(\mathbf{r})\hat{\psi}(\mathbf{r})\mathrm{d}^3\mathbf{r}. \tag{2.4}$$

The field operator,

$$\hat{\psi}(\mathbf{r}) = \begin{pmatrix} \hat{\psi}_{+1/2}(\mathbf{r}) \\ \hat{\psi}_{-1/2}(\mathbf{r}) \end{pmatrix},$$

has two components that either create $[\hat{\psi}^{\dagger}_{\sigma}(\mathbf{r})]$ or annihilate $[\hat{\psi}_{\sigma}(\mathbf{r})]$ an electron with spin projection quantum number σ at position $\mathbf{r}$.

The interaction Hamiltonian $\hat{H}_{\mathrm{int}}$ in Eq. (2.1) is in the following treated as a perturbation of the system described by $\hat{H}_{\rm mol}+\hat{H}_{\rm EM}$. Assuming that initially the molecule with the number of electrons $\rm N_{el}$ is in the electronic ground state $|\Psi_0^{\rm N_{el}}\rangle$, and the photon field is in the Fock state $|N_{\rm EM}\rangle$ containing $N_{\rm EM}$ photons in the mode $(\mathbf{k}_{\rm I},\lambda_{\rm I})$, then the initial state of the system is $|{\rm I}\rangle = |\Psi_0^{\rm N_{el}}\rangle|{\rm N_{EM}}\rangle$. After elastic scattering of an X-ray photon, the final state of the system is $|{\rm F}\rangle = \hat{\rm a}^\dagger_{\mathbf{k}_{\rm F},\lambda_{\rm F}}|\Psi_0^{\rm N_{el}}\rangle|{\rm N_{EM}}-1\rangle$. In the first order, only the $\hat{A}^2$ term in the interaction Hamiltonian (Eq. (2.3)) contributes to elastic scattering of a single photon, according to Fermi's golden rule, the transition rate Γ from the initial $|{\rm I}\rangle$ to the final $|{\rm F}\rangle$ state is

$$
\begin{aligned}
\Gamma_{\rm FI} = \quad & 2\pi\delta(E_{\rm F}-E_{\rm I})\big|\langle F|\hat{H}_{\rm int}|I\rangle\big|^2 \\
= \quad & 2\pi\delta(\omega_{\rm F}-\omega_{\rm I})\Bigg|\langle N_{\rm EM}-1|\langle\Psi^{\rm N_{el}}|\hat{a}_{\mathbf{k}_{\rm F},\lambda_{\rm F}} \\
\times \quad & \frac{\alpha^2}{2}\int {\rm d}^3\mathbf{r}\hat{\psi}^\dagger(\mathbf{r})\hat{A}^2(\mathbf{r})\hat{\psi}(\mathbf{r})|\Psi^{\rm N_{el}}\rangle|N_{\rm EM}\rangle\Bigg|^2 \\
= \quad & \frac{(2\pi)^3}{V^2\omega_{\rm F}\omega_{\rm I}}\delta(\omega_{\rm F}-\omega_{\rm I})|\epsilon^*_{\mathbf{k}_{\rm F},\lambda_{\rm F}}\cdot\epsilon_{\mathbf{k}_{\rm I},\lambda_{\rm I}}|^2 \\
\times \quad & \big|\langle N_{\rm EM}-1|\hat{a}_{\mathbf{k}_{\rm F},\lambda_{\rm F}}(\hat{a}_{\mathbf{k}_{\rm I},\lambda_{\rm I}}\hat{a}^\dagger_{\mathbf{k}_{\rm F},\lambda_{\rm F}}+\hat{a}^\dagger_{\mathbf{k}_{\rm F},\lambda_{\rm F}}\hat{a}_{\mathbf{k}_{\rm I},\lambda_{\rm I}})|N_{\rm EM}\rangle\big|^2 \\
\times \quad & \left|\int {\rm d}^3\mathbf{r}\langle\Psi^{\rm N_{el}}|\psi^\dagger(\mathbf{r}){\rm e}^{{\rm i}(\mathbf{k}_{\rm I}-\mathbf{k}_{\rm F})\cdot\mathbf{r}}\hat{\psi}(\mathbf{r})|\Psi^{\rm N_{el}}\rangle\right|^2 \\
= \quad & \frac{(2\pi)^3N_{\rm EM}}{V^2\omega_{\rm F}\omega_{\rm I}}\delta(\omega_{\rm F}-\omega_{\rm I})|\epsilon^*_{\mathbf{k}_{\rm F},\lambda_{\rm F}}\cdot\epsilon_{\mathbf{k}_{\rm I},\lambda_{\rm I}}|^2\,|f^0(\Delta\mathbf{k})|^2, \qquad (2.5)
\end{aligned}
$$

where $E_{\rm I}$, $E_{\rm F}$ are the energies of the initial and final state, respectively, $\Delta\mathbf{k}=\mathbf{k}_{\rm I}-\mathbf{k}_{\rm F}$ is the scattering vector. The form factor $f^0(\Delta\mathbf{k})$ in Eq. (2.5) is a Fourier transform of the ground state electron density

$$
f^0(\Delta\mathbf{k}) = \int\langle\Psi^{\rm N_{el}}|\hat{\psi}^\dagger(\mathbf{r}){\rm e}^{{\rm i}\Delta\mathbf{k}\cdot\mathbf{r}}\hat{\psi}^\dagger(\mathbf{r})|\Psi^{\rm N_{el}}\rangle{\rm d}^3\mathbf{r} = \int\rho(\mathbf{r}){\rm e}^{{\rm i}\Delta\mathbf{k}\cdot\mathbf{r}}{\rm d}^3\mathbf{r}. \qquad (2.6)
$$

The differential scattering cross section for elastic scattering into an solid angle ${\rm d}\Omega$ is calculated as a sum of the transition rates divided by X-ray photon flux $J_{\rm EM}=\frac{N_{\rm EM}}{\alpha V}$ over the scattered photon states

$$
\begin{aligned}
{\rm d}\sigma = \quad & \sum_{\lambda_{\rm F}}\frac{V}{(2\pi)^3}{\rm d}\Omega\int_0^\infty {\rm d}k_{\rm F}k_{\rm F}^2\Gamma_{\rm FI}/J_{\rm EM} \\
= \quad & \frac{V}{(2\pi)^3}{\rm d}\Omega\sum_{\lambda_{\rm F}}\int_0^\infty {\rm d}\omega_{\rm F}\omega_{\rm F}^2\alpha^3\frac{(2\pi)^3N_{\rm EM}}{V^2}\frac{1}{\omega_{\rm F}\omega_{\rm I}}\delta(\omega_{\rm F}-\omega_{\rm I})|\epsilon^*_{\mathbf{k}_{\rm F},\lambda_{\rm F}}\cdot\epsilon_{\mathbf{k}_{\rm I},\lambda_{\rm I}}|^2 \\
\times \quad & |f^0(\Delta\mathbf{k})|^2\frac{\alpha V}{N_{\rm EM}}. \qquad (2.7)
\end{aligned}
$$

Thus the elastic scattering differential cross section,

$$\frac{\mathrm{d}\sigma(\Delta\mathbf{k})}{\mathrm{d}\Omega} = \alpha^4 |f^0(\Delta\mathbf{k})|^2 \sum_{\lambda_\mathrm{F}} |\epsilon^*_{\mathbf{k}_\mathrm{F},\lambda_\mathrm{F}} \cdot \epsilon_{\mathbf{k}_\mathrm{I},\lambda_\mathrm{I}}|^2, \tag{2.8}$$

relates the electron density of the irradiated molecule to the observed diffraction pattern, and shows that the phase information is inaccessible in scattering experiments. According to Eq. (2.8), only the amplitude of the Fourier transformed electron density is measured in the experiment. For an unpolarized X-ray beam, the polarization-dependent factor $\sum_{\lambda_\mathrm{F}} |\epsilon^*_{\mathbf{k}_\mathrm{F},\lambda_\mathrm{F}} \cdot \epsilon_{\mathbf{k}_\mathrm{I},\lambda_\mathrm{I}}|^2$ integrated over all accessible polarizations (orthogonal to the incident wave vector) is $(1 + \cos^2 2\theta)/2$, thus Eq. (2.8) reads

$$\frac{\mathrm{d}\sigma(\Delta\mathbf{k})}{\mathrm{d}\Omega} = r_\mathrm{e}^2 \frac{(1 + \cos^2 2\theta)}{2} |f^0(\Delta\mathbf{k})|^2, \tag{2.9}$$

where r_e is the classical electron radius, and θ is a scattering angle.

In single molecule XFEL experiments, both the incident beam intensity and the electron density vary during the exposure. Changes in the electron density result from the radiation damage of the sample. Due to this time evolution for unpolarized X-ray radiation, the intensity distribution recorded by a detector is given by

$$I(\Delta\mathbf{k}) = r_\mathrm{e}^2 \frac{1 + \cos^2 2\theta}{2} \Delta\Omega \int\limits_{-\infty}^{\infty} \mathrm{d}t\, I_0(t) \left| \int \mathrm{d}^3\mathbf{r}\, \rho(\mathbf{r}, t) \mathrm{e}^{\mathrm{i}\Delta\mathbf{k}\cdot\mathbf{r}} \right|^2, \tag{2.10}$$

where I_0 is the incident beam intensity, $\Delta\Omega$ is a solid angle subtended by a detector pixel [12]. Here, sufficiently short pulses (few fs in length) with low temporal coherence were assumed, thus the scattering amplitudes are summed incoherently over time slices. When needed, Eq. (2.10) can be generalized to account for potential coherence between the time slices and the pulse polarization. However, these issues are peripheral to the presented structure reconstruction methods and thus will not be discussed in more detail.

2.2 Bayesian analysis

Bayesian analysis allows to extract hidden information indirectly from sparse and noisy data measured in experiments. In single molecule XFEL experiments, the orientation of the irradiated molecule is encoded in the diffraction images, but it is not directly measured. Therefore, to determine the structure from diffraction patterns, I apply Bayes' theorem.

Assuming that n disjoint events $B_1, \ldots, B_n$ are not directly observed in an experiment, but beliefs about their occurrence are expressed in therms of *a priori* probabilities $P(B_i)$. Further, for an observable A directly measured in the experiment, the conditional probabilities $P(A|B_i)$, also called likelihood, are known. According to the Bayes' theorem,

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

the posterior probability $P(B_i|A)$ combines the beliefs about B_i prior to the experiment with knowledge gained from observing the event A, thus completing the information about B_i. For a continuous observable x and a parameter θ belonging to a parameter space Θ, Eq. (2.11) reads

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

Applied to single molecule XFEL experiments, this Bayesian formalism has the following interpretation. Direct observables are photon arrival positions $\mathbf{X}$ recorded on the detector plane. Parameters indirectly observed are the molecular orientation $\boldsymbol{\Theta}$ for a particular diffraction image and the underlying molecular structure S. The likelihood of recording a diffraction pattern given an orientation $f(\mathbf{X}|\boldsymbol{\Theta})$, calculated using an assumed structure model and combined with the *a priori* orientation distribution $p(\boldsymbol{\Theta})$, allows, according to the Bayes' theorem Eq. (2.11), to obtain the posterior probability of the molecular orientation given the diffraction pattern $\pi(\boldsymbol{\Theta}|\mathbf{X})$.

In the Orientational Bayes approach, I apply the Bayesian parameter estimation [29] to determine the molecular orientation for each diffraction image individually. A model of the physical system relates the parameter (orientation) to an ideal signal (intensity distribution). By using this model the likelihood function for the measurement outcome (diffraction pattern) is calculated. Finally, an improved molecular transform is obtained from diffraction images aligned according to the orientation estimated from the posterior probability distributions. In the Structural Bayes approach, however, instead of parameter determination, ratios of posterior probabilities for different structure models are determined. This Bayesian model comparison [29] approach is used to identify a model that fits best to the entire set of diffraction images.

2.3 Posterior probability distribution and orientation determination

To estimate the molecular orientation $\mathbf{\Theta}$ for a single diffraction pattern $\mathbf{X}$, the posterior probability distribution $\pi(\mathbf{\Theta}|\mathbf{X})$ is calculated from the *a priori* orientation distribution $p(\mathbf{\Theta})$ and the likelihood $f(\mathbf{X}|\mathbf{\Theta})$ that this diffraction pattern results from a particular orientation of the model structure. This posterior probability density is used to align the diffraction images on corresponding Ewald spheres in 3D reciprocal space, and thus recover the molecular transform of the irradiated molecule.

The orientation of a molecule exposed to XFEL radiation is denoted by $\mathbf{\Theta}_i = (\theta_i, \psi_i, \varphi_i)$. A diffraction pattern recorded from the molecule oriented according to $\mathbf{\Theta}_i$ is defined by positions of all n_i recorded photons on the detector plane $\mathbf{X}_i = \{(x_i^{(l)}, y_i^{(l)})\}_{l=1\ldots n_i}$. Assuming an incident beam intensity I_0 focused into a focal spot area F_A, from the resulting constant number of total incident photons $N_\mathrm{total} = I_0 F_\mathrm{A}$, only n_i are registered in the i-th image, the rest $N_\mathrm{total} - n_i$ are not. The likelihood of observing a particular arrangement of photons $\mathbf{X}_i$ scattered by the target molecule oriented according to $\mathbf{\Theta}_i$ is given by a product of independent probabilities $I_{\mathbf{\Theta}_i}/N_\mathrm{total}$ of recording a photon at a position $(x_i^{(l)}, y_i^{(l)})$ and the probability of the remaining $N_\mathrm{total} - n_i$ photons not being recorded

$$\begin{aligned} f(\mathbf{X}_i|\mathbf{\Theta}_i) \propto \quad & \left(1 - \frac{A_{\mathbf{\Theta}_i}}{N_\mathrm{total}}\right)^{N_\mathrm{total}-n_i} \prod_{l=1}^{n_i} \frac{I_{\mathbf{\Theta}_i}[\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)})]}{N_\mathrm{total}} \\ \propto & \left(1 - \frac{A_{\mathbf{\Theta}_i}}{N_\mathrm{total}}\right)^{N_\mathrm{total}-n_i} \prod_{l=1}^{n_i} I_{\mathbf{\Theta}_i}[\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)})]. \end{aligned} \tag{2.13}$$

$I_{\mathbf{\Theta}_i}[\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)})]$ is the intensity value in a detector pixel at the l-th recorded photon position $(x_i^{(l)}, y_i^{(l)})$ for an orientation $\mathbf{\Theta}_i$ and $A_{\mathbf{\Theta}_i} = \sum_{l=1}^{N_\mathrm{pixels}} I_{\mathbf{\Theta}_i}[\Delta\mathbf{k}(x^{(l)}, y^{(l)})]$ is the expected amount of elastic scattering registered by the detector in all of its N_pixels pixels. The intensity values were calculated from the model 'seed' structure using Eq. (2.10).

By expressing the likelihood function as a multinomial distribution, Eq. (2.13) automatically accounts for the shot noise. However, additional background noise requires a modification of $I_{\mathbf{\Theta}_i}(\Delta\mathbf{k})$ with an appropriate noise model, which is described in the Methods section.

According to the Bayes' theorem [Eq. (2.12)], the *a priori* distribution of the molecular orientation is also required to calculate the posterior probability. In single molecule experiments, the orientations are assumed to be uniformly distributed; therefore, the probability $\pi(\mathbf{\Theta}_i|\mathbf{X}_i)$ is proportional to the likelihood expressed in Eq. (2.13). The posterior probability distribution carries the complete information that can be gained from the experiments about the underlying molecular orientation for an individual diffraction pattern. I will explore two ways to estimate the orientation and name them 'Maximum Likelihood' and 'Bayesian'.

In the Maximum Likelihood approach, the position of the maximum in the calculated posterior probability distribution is used as a point estimate of the orientation. Photon positions from the diffraction image are then projected onto an Ewald sphere corresponding to that orientation. The entire process is then repeated for all collected diffraction images. Thus the recorded photons are averaged in 3D reciprocal space and yield a molecular transform of the irradiated molecule.

The Maximum Likelihood approach does not use the complete information contained in the posterior probability distribution. Hence in the Bayesian approach I will investigate how much can be gained from the entire orientational information. To this end, photon positions from a single diffraction images are projected onto multiple Ewald spheres with weights given by an appropriate posterior probability value for a particular orientation. Again, this process is repeated for all recorded images, though, in this case yielding a molecular transform that is a weighted average of the registered photons.

2.4 Posterior probability of a structure

In the Orientational Bayes approach, the molecular orientation is estimated for each diffraction pattern assuming a model 'seed' structure. Certain *a priori* knowledge about the molecule is therefore necessary. To limit the extent of prerequisite information, I developed the Structural Bayes approach that compares structure models in a search for one that simultaneously fits best to the entire set of recorded diffraction images. The model comparison is done by calculating posterior probability ratios.

A structure model is described by N atomic positions $S = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$. The likelihood of observing a photon configuration $\mathbf{X}_i = \{(x_i^{(l)}, y_i^{(l)})\}_{l=1,\ldots,n_i}$ scattered by structure

S_j oriented according to $\mathbf{\Theta}_i^{(j)} = (\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$ is

$$f(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}) \propto \Big[1 - \frac{A(\mathbf{\Theta}_i^{(j)}, S_j)}{N_{\text{total}}}\Big]^{N_{\text{total}}-n_i} \prod_{l=1}^{n_i} I\big[R(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})\Delta\mathbf{k}(x_i^{(l)}, y_i^{(l)}), S_j\big], \tag{2.14}$$

where $I(\Delta\mathbf{k}, S_j)$ is the intensity in a detector pixel corresponding to a scattering vector $\Delta\mathbf{k}$ rotated by a rotation matrix $R(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$ corresponding to the orientation $\mathbf{\Theta}_i^{(j)} = (\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})$, $A(\mathbf{\Theta}_i^{(j)}, S_j) = \sum_{l=1}^{N_{\text{pixels}}} I\big[R(\theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)})\Delta\mathbf{k}(x^{(l)}, y^{(l)}), S_j\big]$ is the expected amount of elastic scattering for the orientation $\mathbf{\Theta}_i^{(j)}$ of structure S_j registered by a detector in all its N_{pixels} pixels, and N_{total} is the total number of incident photons.

The likelihoods of registering individual pictures $f\big(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}\big)$ are independent, thus the likelihood of observing an entire set of diffraction patterns $\{\mathbf{X}_i\}$ is given by the product

$$f\big(\{\mathbf{X}_i\}|S_j, \{\mathbf{\Theta}_i^{(j)}\}\big) = \prod_i f\big(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}\big). \tag{2.15}$$

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

$$\pi\big(S_j, \{\mathbf{\Theta}_i^{(j)}\}|\{\mathbf{X}_i\}\big) \propto \prod_i f\big(\mathbf{X}_i|S_j, \mathbf{\Theta}_i^{(j)}\big). \tag{2.16}$$

Finally, the posterior probability of structure S_j giving rise to the set of registered diffraction images $\{\mathbf{X}_i\}$ is calculated by integrating Eq. (2.16) with respect to the molecular orientation $\mathbf{\Theta}_i^{(j)}$

$$\begin{aligned}\pi\big(S_j|\{\mathbf{X}_i\}\big) \propto \quad & \prod_i \iiint f\big(\mathbf{X}_i|S_j, \theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)}\big) \\ & \times \sin\theta_i^{(j)} \mathrm{d}\theta_i^{(j)} \mathrm{d}\psi_i^{(j)} \mathrm{d}\varphi_i^{(j)}.\end{aligned} \tag{2.17}$$

This expression is used to find amongst a set of proposed structure models the one that fits best to the entire set of images.

3 Methods

This chapter describes the performed simulations of single molecule XFEL scattering experiments and the computational details of the two presented Bayesian approaches, the Orientational Bayes approach and Structural Bayes approach. So far, no atomic resolution XFEL diffraction images of single molecules are available, hence the first section focuses on modelling the experiments.

3.1 Modelling single molecule X-ray experiments

For both the simulation and analysis of the diffraction images, a model of the intensity distribution is required. As mentioned in the Theory chapter, the registered intensity $I(\Delta\mathbf{k})$ [defined in Eq. (2.10)] is in general a function of the time dependent electron density $\rho(\mathbf{r}, t)$. However, assuming pulses shorter than 10 fs, the electron density was considered constant during the exposure time [12].

3.1.1 Describing electron density and intensity distribution

In an all atom (AA) representation, the electron density was defined as follows

$$\rho(\mathbf{r}) = \sum_{i=1}^{N_{\text{atoms}}} N_i^{\text{el}} \mathrm{e}^{-(\mathbf{r}-\mathbf{r}_i)^2/(2\sigma_i^2)}, \tag{3.1}$$

where N_i^{el} is the number of electrons in the i-th atom, $\mathbf{r}_i$ is its position, and σ_i is its radius. The sum is performed over all N_{atoms} non-hydrogen atoms in the molecule. Similarly, in the coarse grain (CG) representation, the electron density is calculated as a sum of

Gaussian functions over separate CG beads

$$\rho(\mathbf{r})_{\mathrm{CG}} = \sum_{i=1}^{N_{\mathrm{beads}}} N_i^{\mathrm{CG}} \mathrm{e}^{-(\mathbf{r}-\mathbf{r}_i^{\mathrm{CG}})^2/(2\sigma_i^{\mathrm{CG}^2})}, \tag{3.2}$$

where $\mathbf{r}_i^{\mathrm{CG}} = \sum_{i=1}^{N_{\mathrm{atoms}}}(\mathbf{r}-\mathbf{r}_i)N_i^{\mathrm{el}} / \sum_{i=1}^{N_{\mathrm{atoms}}} N_i^{\mathrm{el}}$ is the position of the i-th bead, $N_i^{\mathrm{CG}} = \sum_{i=1}^{N_{\mathrm{atoms}}} N_i$ is the total number of electrons in all non-hydrogen atoms constituting that bead, and σ_i^{CG} is the bead radius calculated as a standard deviation of the AA electron density representation of the bead.

A 1Å wavelength was assumed for modelling all intensity distributions. For the glutathione, those were computed on a $200 \times 200 \times 200$ grid with a $6.3 \times 10^{-2}\,\text{Å}^{-1}$ spacing, a $300 \times 300 \times 300$ grid with a $6.3 \times 10^{-3}\,\text{Å}^{-1}$ spacing for the titin, and a $300 \times 300 \times 300$ grid with a $1.3 \times 10^{-3}\,\text{Å}^{-1}$ spacing for the ribosome. An incident beam intensity $I_0 = 2 \times 10^8\,\text{photons/Å}^2$ (obtained by focusing approx. 1.57×10^{12} photons to a 10 nm diameter spot) was assumed for the glutathione, and $I_0 = 4 \times 10^6\,\text{photons/Å}^2$ (approx. 3.14×10^{12} photons in a 100 nm diameter spot) for the titin and the ribosome.

3.1.2 Generating diffraction patterns

To mimic single molecule XFEL scattering experiments, the calculated intensity distributions were used to generate diffraction images, which in turn were used to test the proposed structure determination methods. In the experiments, elastically scattered photons are registered at random positions following the intensity distribution. For efficiency reasons, the distribution of photon counts n in a detector pixel of the simulated diffraction images was approximated by a Poisson distribution

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

where $\Delta\mathbf{k}$ is the scattering vector corresponding to a particular pixel. Photon counts at $\Delta\mathbf{k} = 0$ were used to estimate orientations in the Orientational Bayes approach, but not for calculating the structure probabilities in the Structural Bayes approach.

In the diffraction images, only the elastically scattered photons carry structural information; all other registered photons are considered background noise. To simulate the experiments, the background noise was included in the diffraction images by adding Gaussian distributed photons. The standard deviation of the distribution was chosen to be 1/10 of the detector size to model experimental conditions, in which background noise is mostly present in the centre of the image and decays towards the high resolution regions [32]. Consequently, a corresponding Gaussian function was added to the intensity distribution to calculate posterior probability distributions using Eq. (2.13). Assuming that part of the inelastic scattering is not recorded in a diffraction image due to energy filtering of the detector, the amount of additional background photons in the generated images was considered at 10% and 50% ratios of noise to the mean signal photon counts per picture.

The detector size was assumed to be 121 × 121 pixel in a 6 cm × 6 cm area for the glutathione, 241 × 241 pixel (1.2 cm × 1.2 cm) for the titin, and 241 × 241 pixel (2.4 mm×2.4 mm) for the ribosome. In all simulations the distance between the irradiated molecule and the detector plane was assumed to be 10 cm.

3.1.3 Generating random orientations

The orientation distribution for single molecules entering the XFEL beam was assumed to be uniform. To generate orientations following such a distribution, Euler angles used for the orientation description were drawn from a probability density $g(\theta, \psi, \varphi) = (8\pi)^{-1} \sin\theta$ [33], i.e., $\psi \in I[0, 2\pi)$, $\varphi \in I[0, \pi)$, and $\theta = \arccos z$, where $z \in I[-1, 1]$. I used the Gnu Scientific Library [34] implementation of the 'Mersenne twister' algorithm [35] to generate the diffraction images of randomly oriented single molecules.

3.2 Computing posterior distributions

The posterior probability distributions $\pi(\mathbf{\Theta}_i|\mathbf{X}_i)$ for an individual diffraction image were computed from Eq. (2.13) for accessible orientations sampled on a grid. The intensity distribution $I_{\mathbf{\Theta}}[\Delta\mathbf{k}(x, y)]$ registered on the detector plane for a molecular orientation $\mathbf{\Theta}$ is a projection of the intensity distribution on a corresponding Ewald sphere obtained via trilinear interpolation from the molecular transform computed earlier on a 3D cubic grid. To avoid numerical underflows, logarithms of the posterior probabilities were calculated and exponentiated when required.

To improve the orientational resolution without unnecessary computational cost increase, high probability regions were sampled with better accuracy. This orientational resolution enhancement was achieved by first finding probability maxima on a coarse grid and subsequently sampling surrounding relevant regions with a finer step. The coarse grid covered the entire Euler angles range $\theta = (0, \pi)$, $\psi = [0, 2\pi)$, and $\phi = [0, \pi)$ with a 10° step. The fine sampling with a 2° step was done in regions defined as those, where the fine sampled probability exceeded the maximum of coarse sampled probability $\pi^{\text{fine}}(\mathbf{\Theta}_i|\mathbf{X}_i)/\pi^{\text{coarse}}_{\text{max}} \geq 10^{-3}$ times for the glutathione and $\pi^{\text{fine}}(\mathbf{\Theta}_i|\mathbf{X}_i)/\pi^{\text{coarse}}_{\text{max}} \geq 5 \times 10^{-4}$ times for the titin and the ribosome.

In the Maximum Likelihood approach, the orientation of a diffraction pattern was estimated as the position of the fine sampled posterior probability maximum. In contrast, the Bayesian approach considered all fine sampled orientations with assigned weights $W_i^{\text{fine}}(\mathbf{\Theta}_i) = \pi^{\text{fine}}(\mathbf{\Theta}_i|\mathbf{X}_i)/\pi^{\text{fine}}_{\text{max}}$.

The angular resolution dependency $\Delta\Theta(N_{\text{phot}})$ was obtained from posterior probability distributions sampled with a 1° step. Diffraction images of the glutathione molecule rotated from the reference by $\theta = 58°, \psi = 74°$, and $\varphi = 136°$ were used for those calculations.

In the Structural Bayes approach, the posterior portability of a structure given a set of diffraction images $\pi\big(S_j|\{\mathbf{X}_i\}\big)$ was obtained from the product of likelihood functions for individual images $f\big(\mathbf{X}_i|S_j, \theta_i^{(j)}, \psi_i^{(j)}, \varphi_i^{(j)}\big)$ integrated over all orientations using the rectangle rule.

3.3 Retrieving electron densities

The reconstructed molecular transform carries only partial information on the underlying electron density. To retrieve the latter, the relaxed averaged alternating reflections algorithm (RAAR) [20] was used to calculate the missing phases. The amplitudes $|F(\Delta\mathbf{k})| = \sqrt{I(\Delta\mathbf{k})}$ of the reconstructed molecular transform were combined with random phases and provided to the algorithm. The amplitude at $\mathbf{\Delta k} = 0$ was also included in the calculations.

The positivity constraint for the electron density was enforced in a finite support defined as a cube centered at the origin and with a 9Å long edge (twice the radius of gyration of the glutathione). 300 iterations were performed to retrieve the missing phases. The β parameter relaxed from its initial value $\beta_0 = 0.75$ to its final value $\beta_{\text{max}} = 0.99$ in seven iterations following a smooth approximation of a step function [20]. The projections between real and reciprocal space were carried out with the fast Fourier transform implementation from the FFTW library [36].

To quantify the quality of the reconstruction method, R-factors were calculated for both obtained molecular transforms and electron densities defined as

$$R = \frac{\sum ||F_{\mathrm{ref}}(\Delta\mathbf{k})| - |F_{\mathrm{det}}(\Delta\mathbf{k})||}{\sum |F_{\mathrm{ref}}(\Delta\mathbf{k})|}, \tag{3.4}$$

where $|F_{\mathrm{ref}}(\Delta\mathbf{k})|$ is the amplitude of the reference molecular transform and $|F_{\mathrm{det}}(\Delta\mathbf{k})|$ of the recovered molecular transform. The R-factors were computed up to a 0.22 Å resolution ($|\Delta\mathbf{k}| \leq 4.4\,\text{Å}^{-1}$).

3.4 *De novo* structure refinement of the glutathione

Random conformations of the glutathione constituting the search space for the MC refinement were generated by simultaneously changing the dihedral angles in the glycine and cysteine residues of the model molecule. The four dihedral angles in the starting structures were drawn from a uniform distribution. Then, in every MC step, a new set of angles was obtained from the previously accepted ones by varying them according to a normal distribution with a given standard deviation. Initially, the standard deviation was set to 10° for all MC runs. To keep a constant acceptance ratio of 0.2, the standard deviation was doubled when the ratio exceeded this threshold and halved otherwise.

To avoid the system getting trapped in a local minimum of the sampled energy landscape at an early stage of the simulation, simulated annealing was used [37]. To this end, a dimensionless temperature ratio $T_{\mathrm{r}} = T/T_{\mathrm{a}}$ was implemented in the Metropolis criterion,

$$\xi < \exp\left[\frac{(\ln \pi_j - \ln \pi_{j-1})T}{T_{\mathrm{a}}}\right] = \left(\frac{\pi_j}{\pi_{j-1}}\right)^{T_{\mathrm{r}}}, \tag{3.5}$$

where T_{a} is the annealing temperature and T is a pseudo-temperature that reduces the dimension of the temperature ratio. The annealing was achieved through an exponential temperature ratio increase with every accepted MC step from the initial value $T_{\mathrm{r}}^{0} = 0.002$ to the final value $T_{\mathrm{r}}^{\mathrm{f}} = 1.2$, $T_{\mathrm{r}}(n) = T_{\mathrm{r}}^{\mathrm{f}} + (T_{\mathrm{r}}^{0} - T_{\mathrm{r}}^{\mathrm{f}})\mathrm{e}^{-n\tau}$, where n is the number of accepted MC steps and $\tau = 0.005$ is a time constant. These values were adjusted empirically.

3.5 Generating a set of titin conformations

To generate a set of proposed structures, an MD simulation of the titin molecule in vacuum was performed using the GROMACS 4.5 simulation package [38] with the OPLS-AA forcefield [39]. Long range electrostatic interactions (exceeding a 1.0 nm cutoff) were computed with the particle mesh Ewald method [40]. Lennard-Jones interactions were calculated up to a cutoff of 1.4 nm. The protein was coupled to a 300 K thermal bath using the velocity rescale algorithm [41] with a time constant of 0.2 ps. All bonds were constrained with the LINCS algorithm [42]. To avoid intradomain structural changes, additional distance restrains were applied to atoms within the same Ig domains. An integration time step of 2 fs was used. The total length of the simulation was 2.81 ns. The proposed structures were obtained from snapshots 100 ps apart. During the last 10 ps of the simulation, snapshots were taken every 1 ps to obtain conformations with small structural changes compared with the reference structure, and thereby to sample the small RMSD values regime in Fig. 4.6.

4 Results and Discussion

In this chapter, results of the two proposed Bayesian structure determination approaches, the Orientational Bayes approach and the Structural Bayes approach, are presented and discussed. First, I will discuss the accuracy of the orientation determination for individual diffraction images. The quality of reconstructed electron densities will be assessed under consideration of challenging experimental conditions, such as low photon counts and background noise, demonstrating the robustness of the Orientational Bayes approach. Finally, the achievable spatial resolution will be estimated as a function of the incident beam intensity and molecular mass, revealing a scaling of the resolution with the molecular mass as $M^{-1/6}$.

Secondly, a potential application of the Structural Bayes approach to *de novo* structure determination of a small biomolecule will be studied. By limiting the search space, this approach is applied to distinguish among different structures of large biomolecules; here, demonstrated on conformations of three Ig domains and ribosomal translocation states. Additionally, for the ribosome, the sensitivity of the Structural Bayes approach to localized structural changes against structural background and its robustness to model inaccuracy will be discussed.

4.1 Orientation determination

The aim of the Orientational Bayes approach is to estimate the molecular orientation for each diffraction image individually. The accuracy in orientation determination influences the quality of the recovered electron density. A challenge presents itself in achieving sufficient accuracy despite very few signal photons and the presence of background noise. In this section, I will investigate if it is possible to determine electron density maps at atomic resolution from sparse diffraction images containing substantial background noise. Further, an estimate of achievable resolution for molecules of different sizes exposed to various beam intensities will be provided.

To test the Orientational Bayes approach, I simulated XFEL diffraction images of a glutathione molecule. For those, I calculated the posterior probability distributions $\pi(\mathbf{\Theta}_i|\mathbf{X}_i)$. The images contained shot noise modelled by Poisson distribution of photon counts per pixel. Background noise was considered by adding normal distributed photons at levels corresponding to 10% and 50% of the mean photon count per picture.

4.1.1 Posterior probability landscape

To accurately determine the molecular orientation, the posterior probability distribution $\pi(\mathbf{\Theta}_i|\mathbf{X}_i)$ should possess a well pronounced maximum around the actual orientation. By applying the Bayesian formalism, a high accuracy is expected already at very low numbers of signal photons. This expectation is corroborated by an example cut in the ψ, φ-plane through a 3D posterior probability landscape calculated from a simulated diffraction image of a glutathione molecule containing only 65 elastically scattered photons shown in Fig. 4.1. A dominant maximum is already visible in the logarithmic plot (top left row), but the peak becomes pronounced in the linear scale (zoom below). Shot noise causes a deviation of the maximum position ($\theta = 71°, \psi = 52°, \varphi = 33°$) from the actual orientation ($\theta = 73°, \psi = 52°, \varphi = 34°$) by about 2.2°. This shift remains still within the half width of the peak (about 3.2°).

Adding 50% background noise changes marginally the posterior probability surface (right two plots in Fig. 4.1). The resulting shift of the maximum by 6.3° to $\theta = 67°, \psi = 51°, \varphi = 36°$ is more pronounced than in the previous case. The width of the peak increased to about 4.3°.

These results suggest that even at low photon counts and despite additional background noise the orientation information can be extracted from diffraction images. Whether the proposed orientation estimate is accurate enough to resolve a structure at an atomic level will be studied in the following sections.

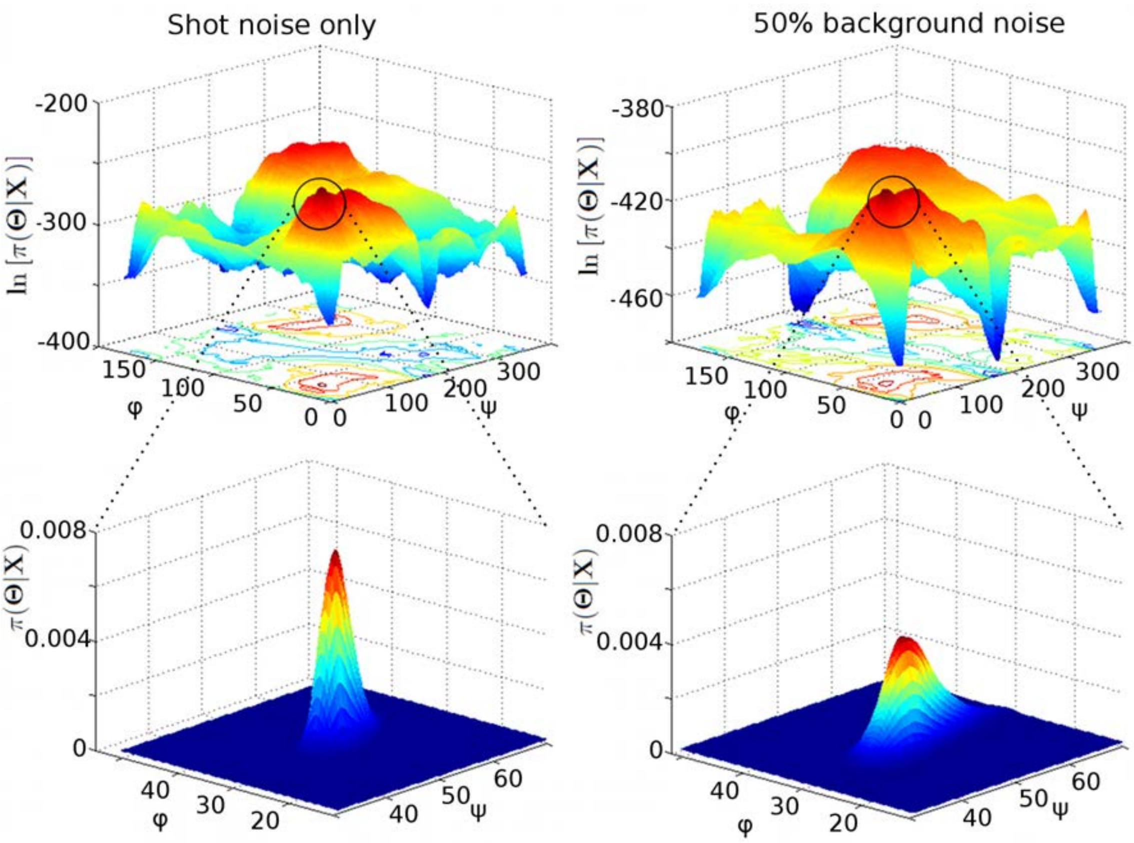

Figure 4.1: Example cuts through 3D posterior probability landscapes $\pi(\mathbf{\Theta}|\mathbf{X})$. To illustrate the accuracy in orientation determination, posterior probability distributions were calculated for diffraction images of a glutathione molecule oriented as follows $\theta = 73°, \psi = 52°, \varphi = 34°$, containing shot noise only (left) and additional 50% background noise (right). The top row shows ψ, φ-cuts at a logarithmic scale taken at the θ coordinate of the posterior probability maximum ($\theta_{\max} = 71°$ for shot noise only, $\theta_{\max} = 67°$ for background noise). The bottom row depicts in linear scale how pronounced the maximum peaks are. Figure adapted from Ref. [28].

4.1.2 Orientation determination and electron density retrieval

Once the molecular orientation is accurately estimated for a particular diffraction image, photons forming that pattern are mapped on an Ewald sphere corresponding to the determined orientation. Photons from many diffraction images are accumulated in voxels of a Cartesian grid, yielding an averaged 3D molecular transform. I investigated two possible ways of performing those photon averages, as described in the Theory and Methods sections. I name these methods and will refer to them in the following as 'Maximum Likelihood' and 'Bayesian', respectively.

The Maximum Likelihood method locates the position of the maximum in the posterior probability landscape. This point estimate of the most likely molecular orientation to generate a given diffraction image is used to transfer the photon positions to a corresponding Ewald sphere. The Maximum Likelihood method does not exploit the entire information contained in the posterior probability distribution. In contrast, the Bayesian method assigns a weight defined by the posterior probability value to every possible orientation. Therefore, the Bayesian method should be less prone to lose information due to incomplete sampling and the discretization of reciprocal space.

To compare the Maximum Likelihood and Bayesian methods, I retrieved the molecular transform of a glutathione tripeptide from 20,000 synthetic diffraction images, each with 82 elastically scattered photons on average. The influence of the background noise on the quality of the reconstruction outcome was studied by including additional 10% and 50% photons relative to the mean signal photons count per picture. Figure 4.2 shows profiles of the retrieved molecular transforms along the k_x axis (red lines) compared to the reference (blue). The plots in the top row collate the results of the Maximum Likelihood and Bayesian methods. The molecular transform profiles in the bottom row illustrate the impact of background noise on the reconstruction accuracy of the Bayesian method. To assess the quality of the reconstruction, the difference between the reference and calculated profiles is plotted underneath each graph.

As shown in the top part of Fig. 4.2, the Maximum Likelihood method reconstructs the molecular transform reasonably well only in the low k-vector regime, which suggests that only low resolution electron density can be determined using this method. The Bayesian method outperforms the Maximum Likelihood method, as it also captures the high resolution details in the reconstructed molecular transform. This increased accuracy seems to result from the use of the entire information contained in the posterior probability distribution, thereby ensuring a better coverage of reciprocal space with Ewald spheres. The improved quality of the reconstructed molecular transforms is also visible in respective R-factors in the three upper rows of Table 4.1.

To assess the influence of the background noise on the quality of reconstruction (bottom row in Fig. 4.2), Gaussian distributed random photon positions were added to the diffraction images containing shot noise only, as well as an appropriate Gaussian

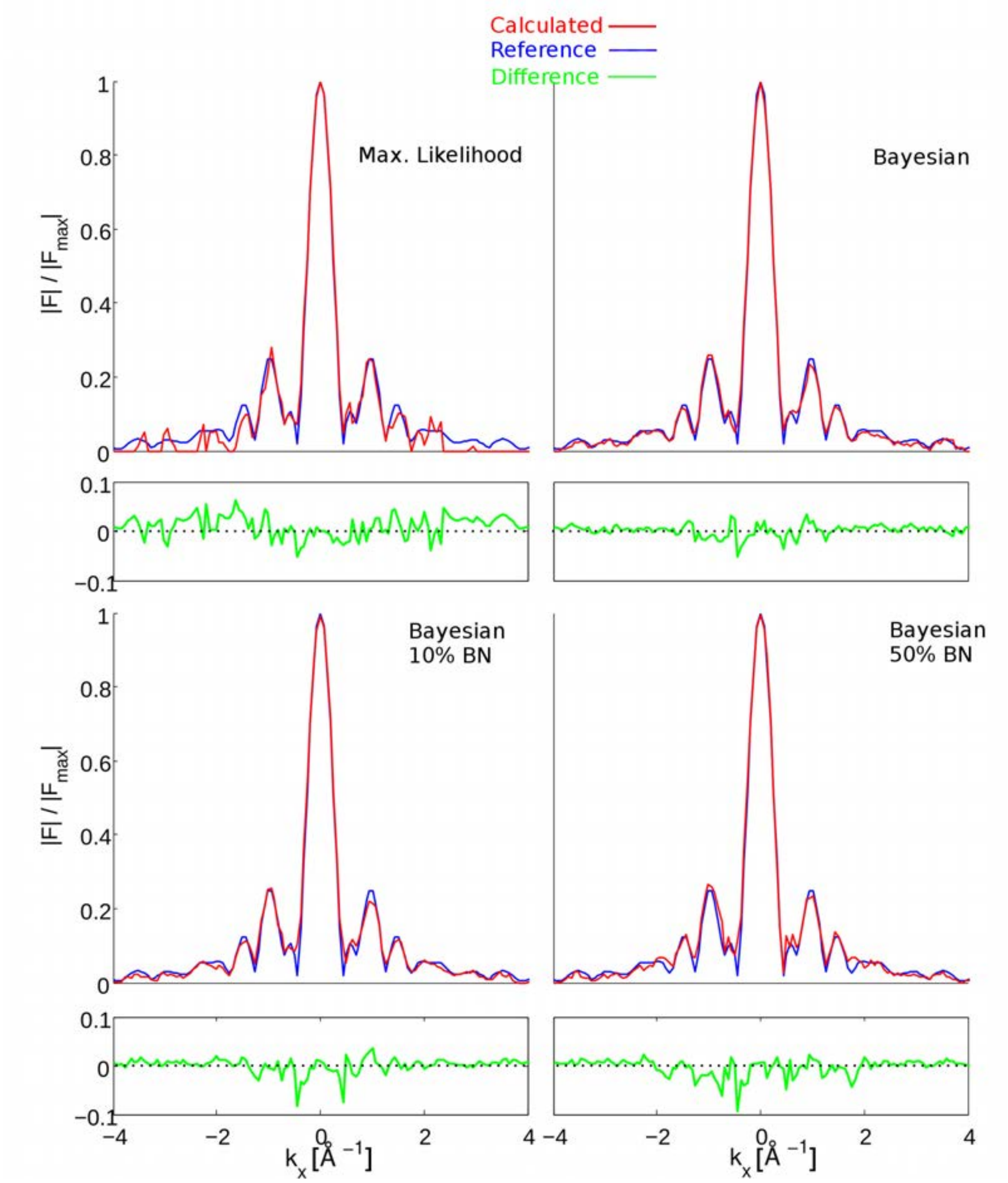

Figure 4.2: Quality of retrieved molecular transforms. Cuts through calculated molecular transforms along k_x axis (red lines) are compared to the reference (blue), and their difference is plotted below in green. The top row contrasts the performance of the Maximum Likelihood (left) and the Bayesian (right) methods in the high resolution regime. The bottom row depicts the influence of background noise at two different levels on the molecular transform both obtained with the Bayesian method. Figure adapted from Ref. [28].

Table 4.1: R-factors as a quality measure for structures determined with the Maximum Likelihood and Bayesian methods from diffraction images containing shot noise only (SN) and additional 50% background noise (BN). R-factors in the three upper rows quantify the accuracy of the retrieved molecular transforms. The three lower rows contain R-factors that measure the similarity in reciprocal space between the retrieved and reference electron densities. All R-factors were calculated up to a 0.22 Å resolution ($|\Delta k| \leq 4.4\,\text{Å}^{-1}$).

	Method	Noise level	R-factor
Molecular transform determination	Max. Lik.	SN	0.48
	Bayesian	SN	0.21
		50% BN	0.23
Electron density determination	Max. Lik.	SN	0.54
	Bayesian	SN	0.27
		50% BN	0.28

model was included in Eq. (2.13) for calculating the posterior distributions. After histogramming the photons from all recorded images in 3D reciprocal space, the background noise was subtracted from the obtained molecular transform. Despite the assumed background noise levels of 10% and 50%, respectively, the Bayesian method yielded still accurate molecular transforms. Whereas the calculated molecular transforms deviate slightly from the reference, the corresponding R-factor (third row of Table 4.1) remains similar to the one obtained for the shot noise only scenario (second row of Table 4.1), thus suggesting no significant deterioration in the quality of the reconstruction despite the additional background noise.

To check whether the anticipated quality and the level of detail of the reconstructed electron densities obtained from the reconstructed molecular transforms reflects the robustness of the Orientational Bayes approach, Fig. 4.3 shows these electron density maps retrieved with a relaxed averaged alternating reflections algorithm (RAAR) [20]. As expected, the Maximum Likelihood method (left side of the middle row) yields a low resolution map, lacking the high resolution details visible in the electron density map retrieved from diffraction images containing shot noise only with the Bayesian method (right side of the middle row). A better performance of the Bayesian method is also reflected in the R-factors listed in the fourth and fifth row of Table 4.1. This loss of detail in case of the Maximum Likelihood method was anticipated from the missing high resolution information in the reconstructed molecular transforms. The bottom row of Fig. 4.3 depicts the robustness of the Bayesian method in the presence of up to 50%

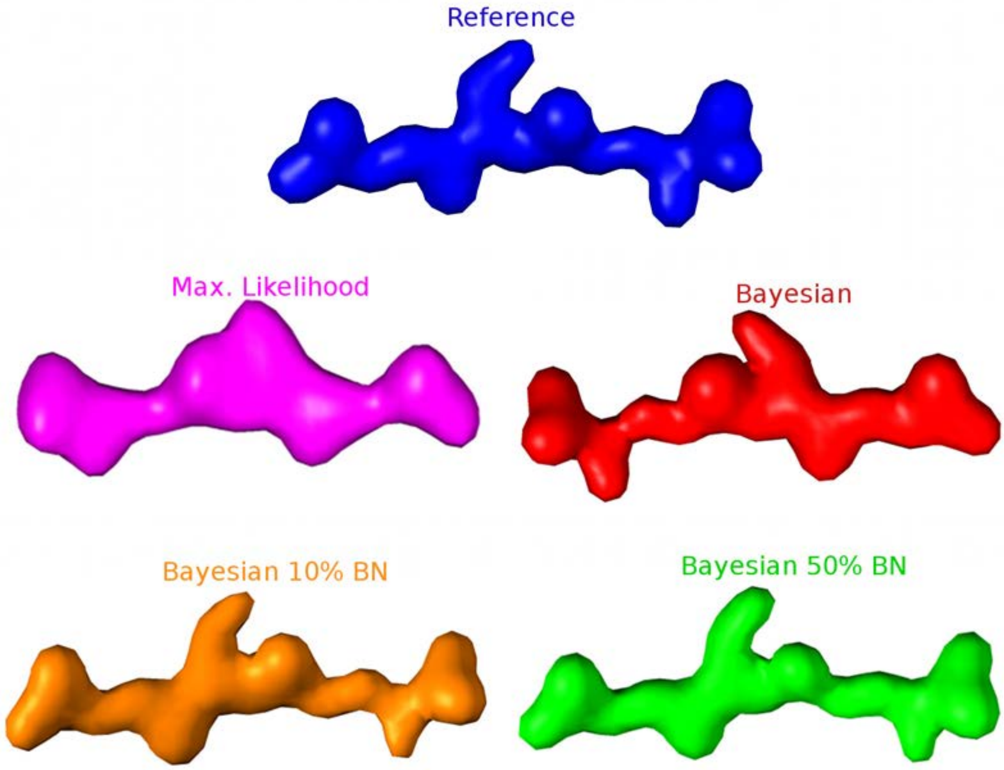

Figure 4.3: Quality of retrieved electron densities compared with the reference (blue). Middle row shows electron densities calculated with the Maximum Likelihood (pink) and Bayesian (red) methods from images with shot noise only. The bottom row illustrates how the Bayesian method copes with additional background noise (BN) at levels of 10% (orange) and 50% (green). Figure adapted from Ref. [28].

background noise. In fact, no significant difference caused by neither 10% nor 50% background noise level is visible in the retrieved electron density maps, as compared to the one calculated from images with shot noise only. The similarity between the latter map and the reconstructed maps from images with 50% background noise is further reflected in the R-factor values presented in the two bottom rows of Table 4.1. Above results suggest that the Bayesian method is robust against shot noise and low photon counts but, more importantly, also against substantial background noise. In contrast to the Maximum Likelihood method, the Bayesian method is not affected by high resolution detail loss because it utilizes the entire information contained in the posterior probability distribution.

4.1.3 Achievable resolution dependence on molecular mass

I have shown that it is possible to solve the structure of a small biomolecule despite low photon counts registered in single molecule diffraction images and in the presence of background noise. In the following, I studied how the achievable resolution depends on molecular masses spanning a wider range, given different beam intensities and background noise levels.

To answer this question, I estimated the achievable spatial resolution Δx by a product of angular resolution $\Delta\Theta$, being a measure of the orientation determination accuracy, and the radius of gyration R_g of the molecule used in the experiment. The angular resolution $\Delta\Theta$ was estimated as a mean distance to the actual orientation, calculated from the posterior probability distribution. The distance between orientations was expressed in Riemannian metrics [43].

The spatial resolution defined as $\Delta x = R_\mathrm{g}\Delta\Theta$ is influenced by two opposing effects. Similar to pointillistic methods in fluorescence microscopy [44], where the resolution scales with the number of photons as $N_\mathrm{phot}^{-1/2}$, I expected the accuracy of the orientation determination $\Delta\Theta$ to increase with the number of photons registered in a diffraction pattern, $\Delta\Theta \propto N_\mathrm{phot}^{-1/2}$. Here, this scaling was anticipated for the following reason. First, I consider a diffraction pattern with N_phot recorded photons resulting in a likelihood function $f(\mathbf{X}|\mathbf{\Theta})$, which yields a posterior probability landscape $\pi(\mathbf{\Theta}|\mathbf{X})$ that I assume to have a well pronounced maximum at $\mathbf{\Theta}_\mathrm{max}$ (e.g. as shown in Fig. 4.1). Next, I assume a diffraction image with m times more photons, mN_phot. This image can be described as a superposition of m images of the first sort with N_phot photons each because the scattering of the individual photons are independent events. Specifically, each of those m subsets of photons are drawn, by construction, from the same likelihood distribution $f(\mathbf{X}|\mathbf{\Theta})$. The likelihood of the superimposed image $f_m(m\mathbf{X}|\mathbf{\Theta})$ is thus proportional to $f(\mathbf{X}|\mathbf{\Theta})^m$. Taylor expansion up to the second order term of $\log f(\mathbf{X}|\mathbf{\Theta})$ and of $log f_m(m\mathbf{X}|\mathbf{\Theta})$ around $\mathbf{\Theta}_\mathrm{max}$ shows the expected scaling of the posterior probability standard deviation with $m^{-1/2}$. N_phot is proportional to the incident beam intensity I_0, and presumably to the molecular mass, thus yielding $\Delta\Theta \propto (I_0 M)^{-1/2}$. *Vice versa*, for a specific orientation accuracy $\Delta\Theta$, the achievable spatial resolution should decrease with the molecular mass, due to increasing radius of gyration $R_\mathrm{g} \propto M^{1/3}$. Put together, these two opposing effects result in the spatial resolution increasing with the molecular mass as $\Delta x \propto I_0^{-1/2} M^{-1/6}$. In this light, using the glutathione as a test case presents the biggest challenge as opposed to larger biomolecules such as the titin or the ribosome.

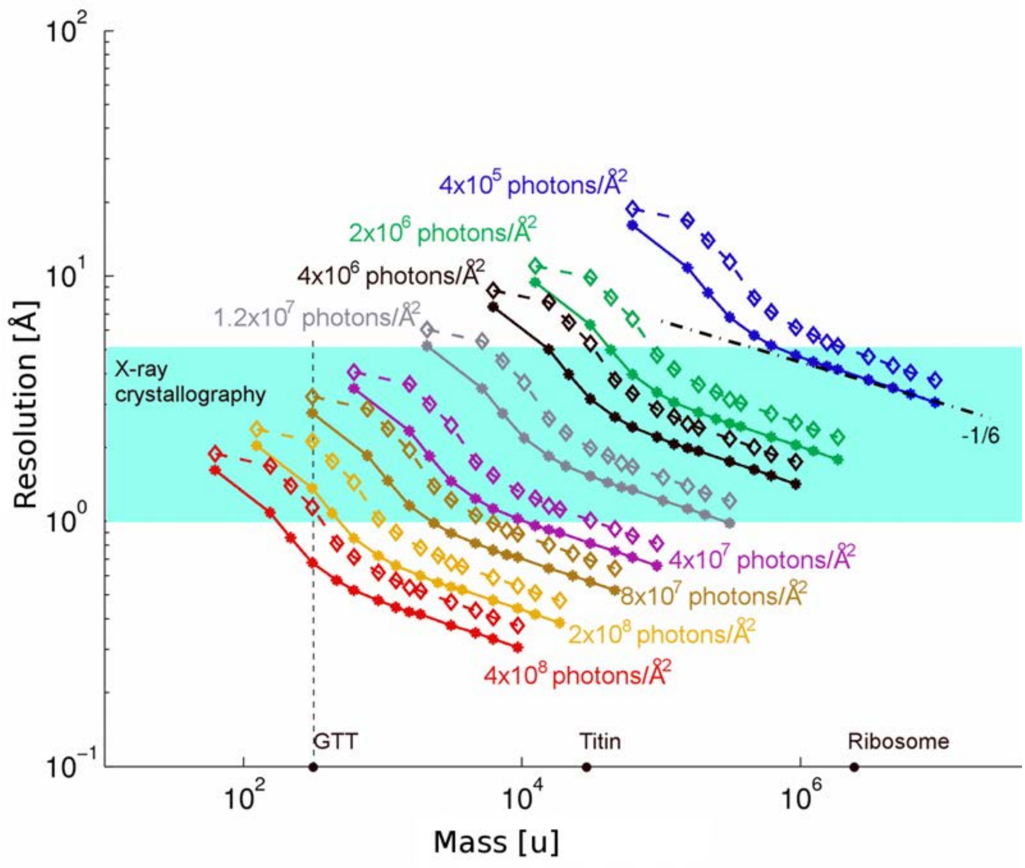

Figure 4.4: Achievable spatial resolution for differently sized molecules and incident beam intensities. Solid lines (dots) correspond to diffraction images with shot noise only and dashed lines (diamonds) to images with additional 50% background noise. Line colours refer to different incident beam intensities (photons/Å^2). For comparison, the resolution range typical for X-ray crystallography (ca 1-5 Å) is marked with a cyan background. The masses of the test molecules used in this work are labeled on the x-axis; glutathione (GTT), titin and ribosome. The black dot-dash line illustrates the expected scaling as $M^{-1/6}$. Figure adapted from Ref. [28].

To verify this scaling, I simulated scattering experiments with varying beam intensities using the glutathione as a target molecule. Thereby, posterior probability distributions were calculated for images containing on average from $N_\text{phot} = 24$ to 3724 scattered photons. For each of those average photon counts, 500 diffraction images with a corresponding N_phot were generated to calculate the average orientational accuracy

$\Delta\Theta$. The resulting achievable spatial resolution for the glutathione can be extracted from the coloured lines, corresponding to different beam intensities, intersecting with the vertical dashed line in Fig. 4.4.

Ultimately, I intended to generalize the calculated $\Delta\Theta(N_{\text{phot}})$ dependence obtained for the glutathione to predict the expected resolution for biomolecules of different sizes. Those molecules were modelled by scaling up the glutathione α-times in size and $\alpha^3 M$ in molecular mass, accordingly (horizontal axis in Fig. 4.4). As mentioned previously, the number of registered photons was assumed to scale with the molecular mass as $N_{\text{phot}} \propto I_0 M$. As expected, the colour-coded curves for different beam intensities I_0 show that the achievable spatial resolution increases $\propto M^{-1/6}$ (exemplified by a dot-dash line with a slope of $-1/6$) and $\propto I_0^{-1/2}$ for large molecular masses, corresponding to $N_{\text{phot}} > 200$. For lower photon counts, smaller masses, respectively, the resolution changes more rapidly as a result of comprised orientational accuracy $\Delta\Theta$. For very sparse images, it is difficult to distinguish the correct orientation from those rotated by about 180$^{\text{o}}$, hence the misaligned orientations become almost equally probable as the ones around the correct orientation. Therefore, the achievable orientational accuracy $\Delta\Theta$ approaches 90° at very low photon counts.

The achievable spatial resolution for a beam intensity comparable with the one currently available at Stanford Linear Accelerator Center (SLAC) is plotted with black lines in Fig. 4.4. For these lines, I assumed a 12 keV beam with an intensity of $I_0 = 4.0 \times 10^6$ photons/Å^2 focused to a 100 nm spot [12], whereas an intensity of approximately 10^5 photons/Å^2 photons in a 1 μm focal spot was achieved recently; however, for up to 2 keV XFEL beams [45]. According to the estimated resolution dependence on molecular mass, an intensity of $I_0 = 4.0 \times 10^6$ photons/Å^2 should already suffice to solve large structures, e.g three Ig domains of a titin molecule or the ribosome, within a resolution range typically achieved in X-ray crystallography, indicated by a shaded area. To achieve atomic resolution for smaller molecules, higher beam intensities are necessary. For instance, imaging the glutathione would require increasing the beam intensity to $I_0 = 2.0 \times 10^8$ photons/Å^2 by reducing the focal spot size to 10 nm, which should be possible, at least for 6 keV XFEL radiation [46].

4.2 Structure optimization

Up to this point, I used a 'seed' model structure to determine the molecular orientation for each diffraction image separately. Using the Bayesian formalism allowed to extract the orientational information even from sparse and noisy scattering data, thereby enabling reliable structure determination at atomic resolution. Further, I intended to investigate how to circumvent the need of the 'seed' model for determining the structure from single molecule scattering images. This alternative approach aims at finding a structure that simultaneously fits best to the entire set of diffraction images. Here, the posterior probability of a structure giving rise to a set of observed images serves as a comparison criterion to distinguish between proposed structures and is implemented in a refinement procedure. The structure model defined in real space can thus be iteratively optimized according to the probability measure, as will be shown in the following section.

4.2.1 *De novo* structure determination

I will assess the ability of the developed Bayesian approach to solve molecular structures *de novo*. To this end, the posterior probability of a structure given a set of diffraction images, Eq. (2.17), was implemented in a Monte Carlo (MC) structure optimization of the glutathione. In contrast to the Orientational Bayes approach, here, the molecular structure S_j is treated as an additional parameter that is optimized in a MC simulation to find the structure that fits best to the entire set of recorded diffraction images; for details refer to Methods section.

The search space of the proposed structures consisted of glutathione conformations differing in four dihedral angles between cysteine and glycine residues. The internal structure of the three amino acids constituting the peptide was assumed to be known. The search was performed from starting structures with randomly chosen dihedral angles. In each MC step, the posterior probability of a newly proposed structure, $\pi_{j+1} = \pi\big(S_{j+1}|\{\mathbf{X}_i\}\big)$, generated by changing all four dihedral angles according to a normal distribution, was calculated using Eq. (2.17). The posterior probability ratio of the newly proposed and previously accepted structure π_{j+1}/π_j was used as the Metropolis criterion [47] with associated energies $E_j = -k_BT \ln \pi_j$. Consequently, the proposed structure was accepted if $\xi < \exp(-\Delta E/k_BT) = \pi_{j+1}/\pi_j$, where ξ is a random number between [0,1).

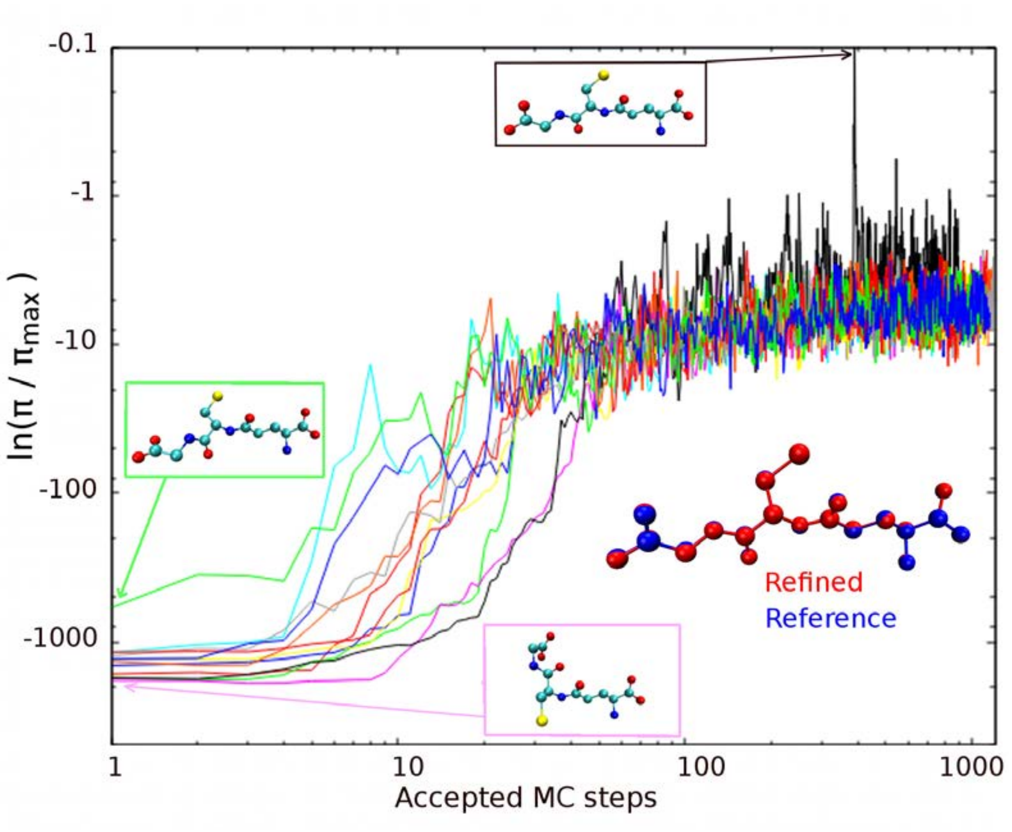

Figure 4.5: *De novo* structure determination in a Monte Carlo simulation. To optimize the structure of the glutathione, the posterior probability was used as a criterion to find a structure that most likely gave rise to 200 simulated diffraction images. The logarithm of the normalized probability was plotted for accepted structures in twelve independent MC runs (colour lines), each of them started from different random configurations. Two example initial structures are depicted in green and pink boxes. The most probable structure was observed after about 500 steps (black box); its overlay (blue) with the reference structure (red) illustrates their similarity. Figure adapted from Ref. [28].

The glutathione structure was refined against only 200 simulated diffraction images with ca. 76 elastically scattered photons per picture on average, but without background noise. Twelve MC runs from random structures were performed, for each of them Fig. 4.5 shows in colour lines the increasing posterior probability $\pi\big(S|\{\mathbf{X}\}\big)$, normalized to the probability of the most probable structure $\pi_{\max}$, as a function of accepted MC steps. The most and least probable starting conformations are shown in blue and green boxes, respectively. After about 700 accepted MC steps, all runs seem to converge, and the most probable structure, shown in a black box, was found after about 380 accepted MC steps. The remarkable structural similarity between the refined (blue) and the reference (red) structure is depicted in the overlay of these structures and is also reflected in a root mean square deviation (RMSD) of 0.02 Å.

These results show that the Structural Bayes approach is able to accurately solve *de novo* the structure of single molecules; however, only if the search space is limited, as in the presented case of the glutathione. For larger biomolecules such MC optimization might not be feasible because of a sampling problem. It was not the scope of this work to propose a solution to overcome the sampling problem; instead, I applied the Structural Bayes approach to distinguish among different conformations of larger structures.

4.2.2 Structure discrimination for large biomolecules

An exhaustive amino acid based search space for *de novo* structure optimization of biomolecules larger than a peptide might be computationally too demanding. Therefore, I limited the structural search space for two other example molecules: titin and the 70S ribosome of *E. coli*. Here, the goal is to distinguish the correct conformation of a molecule from the incorrect ones.

To test the developed approach on a relatively large protein, I used a 283 residues long titin molecule with three Ig domains (Ig67-Ig69). The internal structure of the domains remained rigid, whereas the domains were flexibly connected via proline-, glutamate-, valine-, and lysine-rich (PEVK) linkers (PDB entry 2RIK [48]). 290 conformations, differing in the mutual arrangement of the domains, were obtained in a 2.81 ns MD simulation with distance restrains put on the domain atoms, yet allowing the flexibility of the linkers. The snapshot at 2800 ns was chosen as reference structure to generate 200 diffraction images, containing on average 376 photons per picture; the images contained shot noise only. For each of the generated structures, the posterior probability of that structure giving rise to the observed images $\pi\big(S|\{\mathbf{X}\}\big)$ and the RMSD to the reference structure were calculated and plotted in Fig. 4.6 (blue asterisks). Structural differences between conformations are shown for the reference structure (blue) and three other sample structures along the RMSD range (magenta, orange and red).

As expected, the reference structure is the most probable one and any structural differences in the other sampled conformations lower their posterior probability. Even

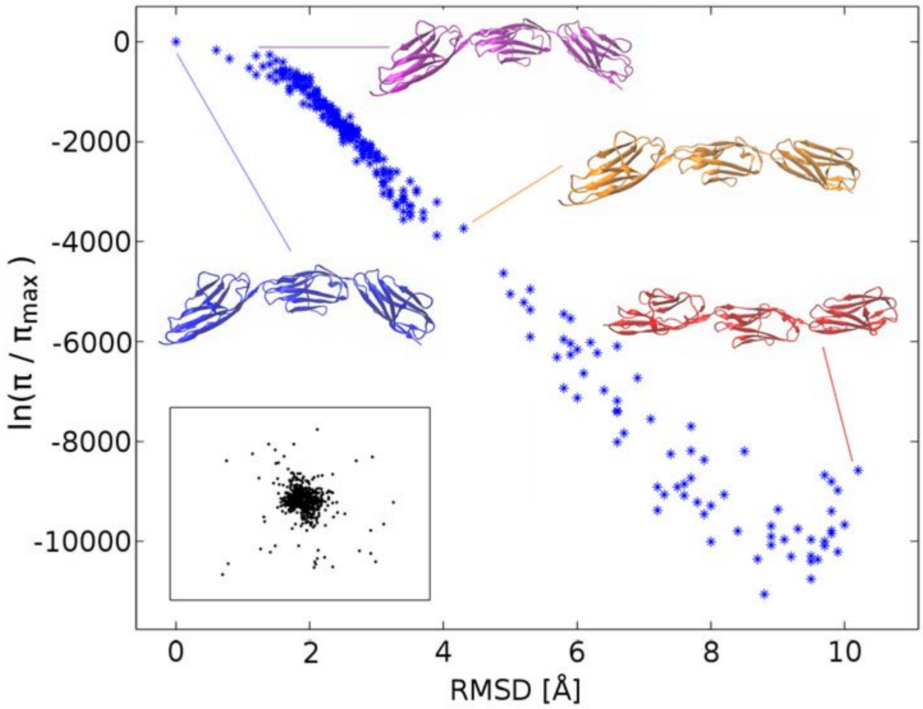

Figure 4.6: Finding the correct titin conformation within a limited set of proposed structures. 290 different conformations were compared against 200 synthetic diffraction images (example image shown in the bottom left corner) generated from the reference (blue, cartoon representation). For each of these structures its posterior probability was plotted versus the RMSD with respect to the reference (blue asterisks). Three intermediate structures in a cartoon representation are shown in magenta, orange and red colour. Figure adapted from Ref. [28].

the smallest structural change of 0.6 Å decreases the probability about 1.24×10^{72} times, suggesting that the reference structure could be correctly identified with much certainty amongst the sampled conformations with an accuracy better than 0.6 Å RMSD.

The largest molecule used as a test case is the bacterial 70S ribosome with a molecular mass of about 2.5 MDa. However, the ribosome size was not the only criterion for the choice as a test molecule. During the translocation process, the ribosome undergoes structural changes at different length scales as described recently [3, 49], and thereby might challenge the developed approach. The translocation states of the ribosome are classified according to the tRNA chain positions with respect to the binding sites of the 30S and 50S subunits. Apart from the tRNA displacement, the structural changes between the states also stem from different subunit configurations.

To test whether the Structural Bayes approach can be used to identify the reference structure amongst a set of proposed structures, I chose seven translocation state structures that were obtained as atomic fits to cryo-EM maps and kindly provided by my colleagues [3]. The reference structure was chosen from the pre-translocation (pre1) state, as previously defined in Ref. [49], and used to generate 200 diffraction images containing on average 1.075×10^5 photons per picture. The images contained shot noise only, and were used to test the Structural Bayes approach for its capacity to discriminate between different ribosomal structures at three difficulty levels.

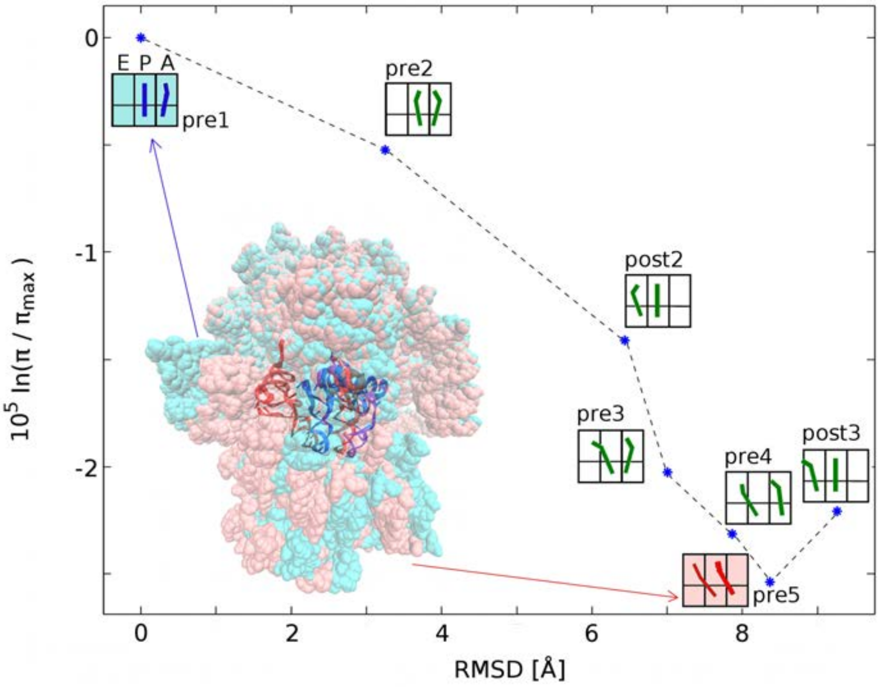

Figure 4.7: Identifying the correct ribosomal translocation state among seven proposed structures. The different translocation states were compared against 200 synthetic diffraction images generated from the reference (pre1 state). Structural differences were expressed in terms of normalized posterior probabilities and the RMSD with respect to the reference. These differences are exemplified by an overlay of the pre1 (blue) and the pre5 (red) state that contrasts the subunit (surface) arrangement and the tRNA chain (cartoon) configurations. Boxes next to each of the points show the tRNA positions with respect to the binding sites (E, P, A). Figure adapted from Ref. [28].

First, differences between entire ribosomal structures were regarded. Posterior probabilities were calculated for all seven structures and plotted in Fig. 4.7 versus the RMSD to the reference (blue asterisks). The box representations at each point illustrate the location of the tRNA chains with respect to the three binding sites of the subunits: aminoacyl (A), peptidyl (P), and exit (E) (as defined in previous studies [49]). An overlay of pre1 (red) and pre5 (blue) structures depicts the overall structural change between these two states resulting from different subunit arrangements (surface representation) and the translocation of the tRNA chains (cartoon representation). As anticipated, the reference structure was correctly identified as the most probable one to give rise to the recorded diffraction images. Due to high photon counts per image, the reference structure was determined with almost certainty in contrast to the remaining six structures; note the large posterior probability ratios compared with the remaining structures. Apparently, the posterior probability of a structure decreases with increasing structural difference to the reference, here, expressed in terms of RMSD. Though, the post-translocation (post3) structure deviates from this trend, possibly because the relatively small change in the subunit arrangement compared to the reference structure masks to some extent the tRNA chains displacement.

The next challenge for the developed method was to detect local structural changes against a large structural background. In the ribosome, the tRNA chains constitute only a small part of the entire complex, yet tracing their movement along the mRNA chain is important to understand the translocation process. Hence following question emerges: is it possible to detect the structural changes of the tRNA chains alone against the structural background of ribosomal units? To answer it, I constructed seven test structures that consisted of the tRNA chains from the seven translocation states embedded in the pre1 subunit configuration. This way, these test structures differed only in the tRNA chain positioning and their internal conformation.

Figure 4.8 depicts how the posterior probabilities of the seven test structures giving rise to 200 diffraction images generated from the pre1 state decreased along the tRNA chain displacement, characterized in terms of RMSD compared with the reference structure. The reference tRNA chain positions were successfully identified as most probable. The lower x-axis shows the RMSD values of the entire complex compared with the reference, whereas the upper x-axis depicts RMSDs of the tRNA chains only. Due to size differences, the structural background partly masks the local structural changes, which is visible in different lower and upper RMSD ranges. The inset in the upper right corner illustrates the size comparison between the tRNA chains (cartoon representation) and the ribosomal units of the pre1 state (surface representation). Overlays of tRNA chains

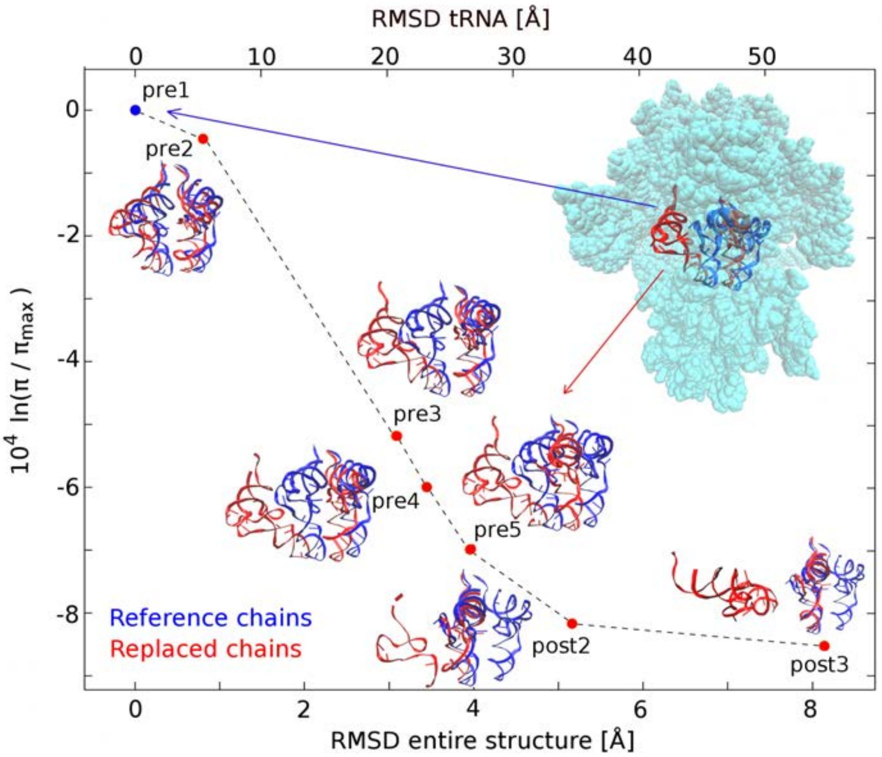

Figure 4.8: Detecting local structural changes appearing during the ribosomal tRNA translocation process. Normalized posterior probabilities that each of the test structures, consisting of native tRNA chains embedded in pre1 state subunits, gave rise to 200 synthetic diffraction images of the reference structure (pre1) decrease along the chains displacement. The RMSDs on the lower x-axis quantify the overall structural change, whereas the upper x-axis shows the RMSDs of the local structural changes resulting from the tRNA movement. The conformation change from the reference chains (blue) to the native for other translocation states (red) is depicted at each of the red dots. The inset in the top right corner illustrates the local structural change in tRNA chain configurations from the pre1 to the pre5 state (cartoon) against the structural background of the pre1 state (surface). Figure adapted from Ref. [28].

next to red points in the plot show the structural differences between the reference chains (blue) and the replaced ones (red). As shown in the plot, the translocation process can be tracked in terms of decreasing posterior probability. Even small localized structural changes of tRNA chains are detected against a large structural background causing a well pronounced drop in posterior probability values along the increasing RMSD. This result suggests that single molecule X-ray scattering experiments might be suitable for studying, e.g. ligand binding processes.

The goal of the third difficulty level was to challenge the Structural Bayes approach by introducing an inaccuracy in the structure model. In particular, I tested if detailed structural information can also be retrieved against distorted structural background. To answer this question, I created test structures by embedding the native tRNA configurations of the seven different translocation states into the pre2 state subunit arrangement and calculated their probabilities to give rise to diffraction images generated from the pre1 state. This way, I introduced an inaccuracy in the structural background model.

As in the previous case, Fig. 4.9 illustrates decreasing posterior probability along the translocation process. Despite the inaccuracy in modelling the subunits arrangement (note the offset in the lower RMSD axis), it is still possible to correctly identify the reference position of tRNA chains and track the local structural changes in the translocation process. The decreasing posterior probability trend is similar to the one obtained previously. Here, the posterior probability ratio between the reference and the second most probable structure is slightly less pronounced yet still very large $\ln(\pi_{\text{reference}}/\pi_{\text{2nd}}) \approx 4.04{\times}10^3$, indicating an almost certain structure discrimination.

These results show that it is indeed possible to distinguish between among conformations of large biomolecules. The developed method is also sensitive to local structural changes that can be tracked even against large and inaccurately modelled structural background.

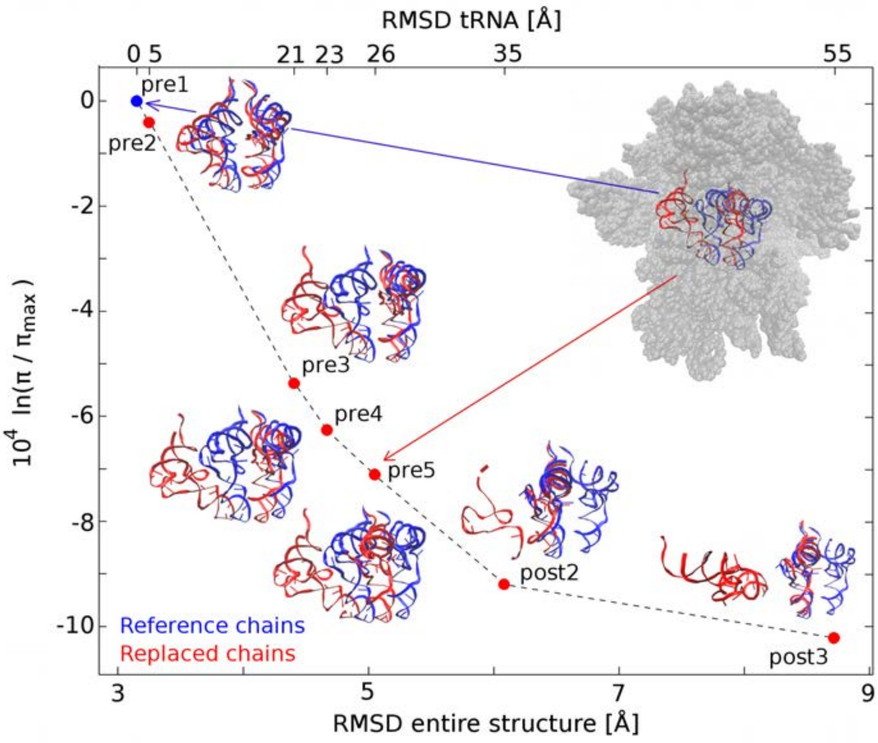

Figure 4.9: Detecting local structural changes in tRNA configurations against inaccurate structural background model. Normalized posterior probabilities that each of the test structures consisting of native tRNA chains embedded in pre2 state subunits give rise to 200 synthetic diffraction images of the reference structure (pre1) decrease along the chains displacement. The RMSDs on the lower x-axis describe the overall structural change (offset in the scale is caused by the model inaccuracy), whereas the upper x-axis shows the RMSDs of the translocated tRNA chains alone. The conformation change from the reference chains (blue) to the native for other translocation states (red) is depicted at each of the red dots. The inset in the top right corner illustrates the local structural change in tRNA chain configurations from the pre1 to the pre5 state against the inaccurately modelled structural background, i.e. the pre2 subunit arrangement (surface). Figure adapted from Ref. [28].

4.2.3 Structure discrimination using a multiscale structure model

So far, all molecular structures were modelled at atomic resolution. The above results suggest that the developed method can be used to extract local structural information despite inaccurate structural background. Hence next, I investigated whether certain less important regions, such as the structural background, can be modelled at a lower resolution while maintaining atomic resolution in the regions of interest. This way, computation cost would be reduced without sacrificing accuracy in retrieval of relevant structural details.

To achieve this multiscale structure model, the electron density of tRNA chains was, as previously, calculated atom-wise whereas that of the ribosomal units residue-wise, i.e., single amino acids and nucleic acids were represented as coarse grain (CG) beads (for a detailed description refer to Methods section). These models were compared against 200 diffraction images with 1.075×10^5 photons per picture generated from an all atom (AA) representation of the pre1 state.

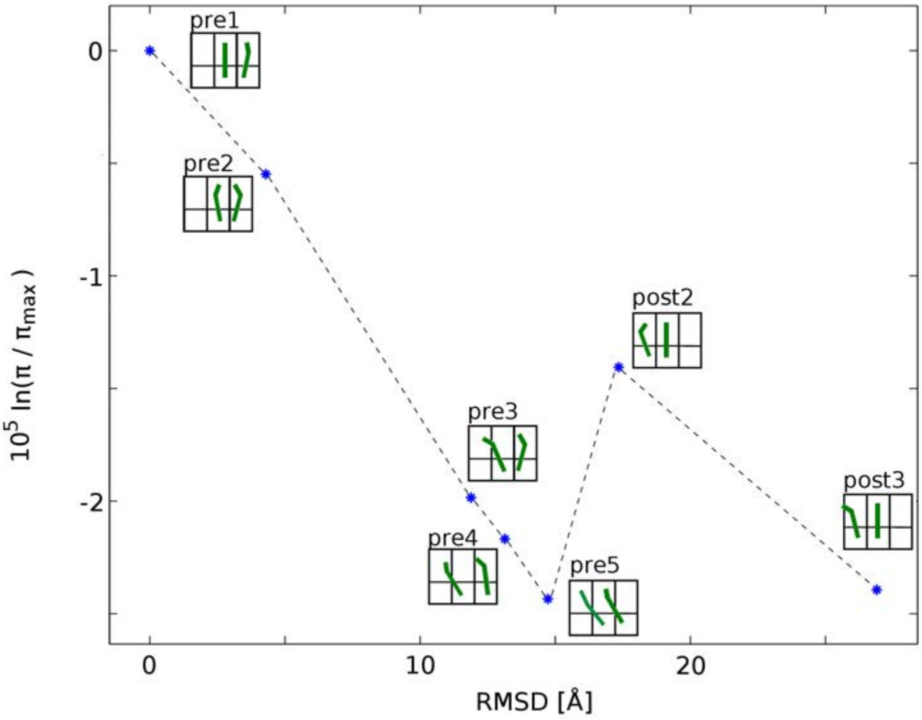

Figure 4.10: Identifying the correct ribosomal translocation state using mixed CG/AA structure representations. Test structures were compared against 200 synthetic diffraction images that were generated from the AA representation of the pre1 state. For seven different translocation states, the logarithm of the normalized posterior probability and the RMSD with respect to the CG/AA representation of the pre1 state were calculated. Boxes next to each of the points depict the location of the tRNA chains in the ribosome at corresponding translocation states.

First, to test whether the multiscale description would be suitable for structure discrimination with the developed method, the posterior probabilities with respect to the generated diffraction images were calculated for CG/AA representation structures of seven different translocation states. The probability values of these states were plotted in Fig. 4.10 versus the RMSD to the CG/AA structure of the pre1 state. As in the AA representation, the correct state was identified with almost certainty. The post2 and post3 states appear as outliers from the decreasing trend of the posterior probabilities of the other states. Differences in subunit arrangements might partly compensate the structural changes caused by tRNA translocation thus leading to the observed deviation.

Next, I asked if structural details at atomic resolution can be extracted from a multiscale model with an additional inaccuracy introduced in the structural background. To answer this question, test structures were created by modelling the tRNA chain arrangements native to the seven translocation states in AA representation and placing them in CG subunit environment of the pre2 state. Figure 4.11 shows a decrease in posterior probability that the test structures fit to the diffraction images of the pre1 state in AA representation plotted versus structural difference to the CG/AA representation of the pre1 state expressed in terms of RMSD. Local structural changes are still traceable at a high resolution even though both the accuracy and the resolution in modelling of the subunits were comprised. Compared to the results obtained for an AA representation (Fig. 4.9), the probability ratio between the most probable and the second most probable structure is slightly less pronounced $\ln(\pi_{\text{max}}/\pi_{\text{2nd}}) \approx 1.60 \times 10^3$, but still assuring an almost certain identification of the correct tRNA chain arrangement amongst all proposed structures.

Taken together, these results suggest that it should be possible to extract high resolution structural information in regions of interest even using an inaccurate and low resolution model of the structural background. Hence computational effort can possibly be spared without comprising the high accuracy in small, yet important regions. This CG description might be useful in devising a structure refinement scheme with increasing levels of detail, i.e., decreasing size of CG beads; however, discussion of such a procedure exceeds the scope of this work.

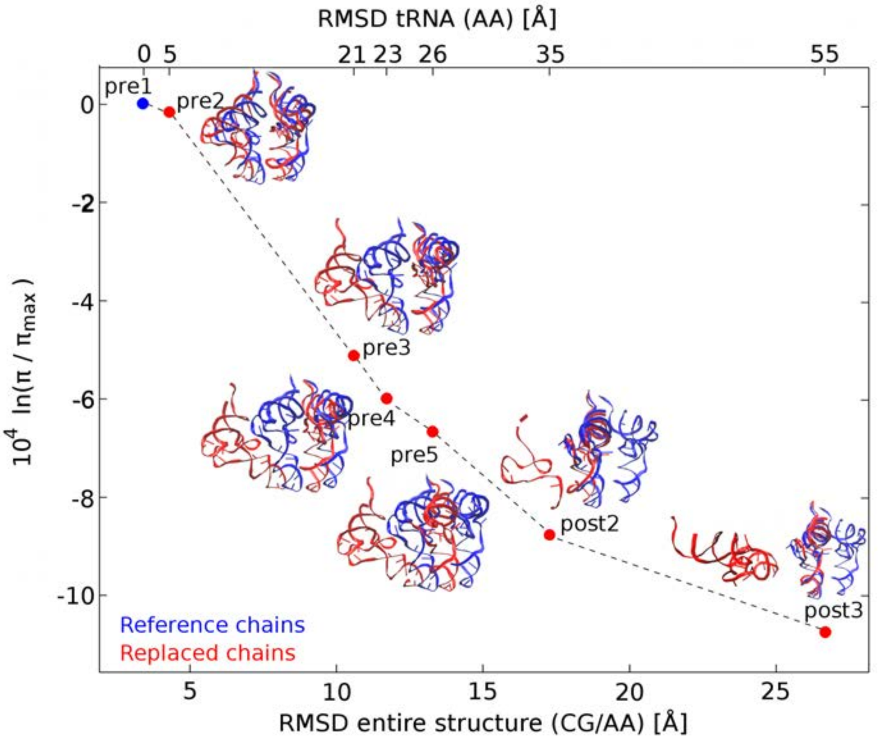

Figure 4.11: Tracing local structural changes during tRNA translocation using the inaccurate multiscale structure model. For each translocation state model, obtained by embedding the native tRNA chain conformations within the subunits of the pre2 state (inaccuracy of the model), the normalized posterior probability was calculated (y-axis) using 200 synthetic diffraction images generated from the AA representation of the pre1 state. The bottom x-axis corresponds to the RMSDs of entire structures (mixed CG/AA representation), whereas the upper x-axis shows the RMSDs of the tRNA chains alone (AA representation). The difference between the translocated (red) and the reference chains (blue) is shown for each of the states.

5 Summary and Future Perspectives

In this work, I developed two Bayesian approaches to determine the structure of biomolecules from single molecule X-ray scattering images. The Orientational Bayes approach, which is similar to the EMC algorithm by Loh and Elser [24], determines the molecular orientation for individual diffraction images. The Structural Bayes approach is used to identify a structure that simultaneously fits best to all recorded diffraction images. These approaches were tested using simulated scattering images containing very few photons and substantial shot noise to mimic the challenging conditions of future experiments. The anticipated low photon counts are indeed one of the challenges for structure reconstruction algorithms; as recent calculations showed, a relatively small 500 kDa protein would scatter about 100 photons per picture, yielding photon counts of 4×10^{-2} photons per pixel in the high resolution regions of a diffraction image [18].

The Orientational Bayes approach uses a 'seed model' (a structure similar to the reference, obtained from, e.g. nanocrystallography or from homology modelling [50, 51]) to determine the molecular orientation for individual images. The underlying molecular transform is thereby determined by aligning and averaging the collected images in 3D reciprocal space. Here, I compared two orientation determination schemes differing in the extent of information contained in the calculated posterior probability distribution they actually use. The Maximum Likelihood scheme estimates the molecular orientation as the position of the maximum in the posterior probability landscape. In contrast, the Bayesian scheme assigns weights to all sampled orientations according to the obtained posterior probabilities. The results obtained in this work showed that, by using the entire available information encoded in the posterior probability distribution, the Bayesian scheme is superior to the Maximum Likelihood one. The Bayesian scheme ensures a more accurate coverage of the 3D reciprocal space, and consequently, it recovers high resolution structural information.

The ability to distinguish between different orientations in the Orientational Bayes approach depends on the shape of the posterior probability landscape, which in turn is determined by the number of recorded elastically scattered photons and the level of additional background noise. With increasing number of photons, the maximum of the posterior probability distribution becomes narrower, leading to a better accuracy in the orientation determination. An opposite effect is observed in the presence of background noise; the additional noise results in a less pronounced posterior probability maximum. The anticipated and calculated scaling of the angular resolution with the number of recorded photons as N_{phot} translates into a spatial resolution increase with the molecular mass as $M^{-1/6}$. This scaling implies that, for a given beam intensity, a better resolution is expected for larger molecules; also when background noise is present. This result appears counter-intuitive from the X-ray crystallography perspective. There, obtaining high resolution structures for large molecules is challenging due to structural inhomogeneity resulting in imaging an average distribution of atomic position over many possible conformations. For instance, the crystal structure of three Ig domains (∼30.8 kDa) was solved at a 1.6 Å [48] resolution, whereas the bacterial 70S ribosome (∼2.5 MDa) structures were obtained at 2.8 Å [52] and 3.7 Å [53] resolutions, respectively. Here, assuming an incident beam intensity anticipated in single molecule XFEL experiments of about 4.0×10^6 photons/Å^2 [12], the estimated resolution for the Ig domains construct is about 3.2 Å and the resolution increases for the ribosome to 1.2 Å. The results further suggest that by reducing the focal spot size and thereby increasing the beam intensity, even higher resolutions should be achieved. Small focal sizes, about 10 nm in diameter, would be required to image smaller molecules, such as the glutathione (308 Da), at about 1.4 Å resolution.

The Orientational Bayes approach requires a 'seed' model structure to determine the molecular orientation for each of the recorded images. To avoid this requirement, I developed and studied an alternative: the Structural Bayes approach. This approach does not treat the collected diffraction images individually; instead it considers the entire set of images to find a structure in real space that fits best to all images simultaneously. By defining the structures in real space and distinguishing between them according to their probabilities, the phase retrieval stage is circumvented. Also, a better spatial resolution of recovered structures is expected because the need to accurately align the images, which limits the resolution in the Orientational Bayes approach, is avoided.

To assess the possibility of *de novo* structure determination, I implemented the posterior structure probabilities in a Metropolis MC refinement procedure for a small test molecule, glutathione. From 200 simulated diffraction images containing on average

~ 70 elastically scattered photons each, a structure almost identical to the reference was obtained. The refined structure differed from the reference by only 0.02 Å in terms of RMSD.

Whereas the presented amino acid based *de novo* structure refinement was successfully applied to the glutathione, the sampling problem might hinder this procedure in refining larger molecules. To avoid this problem, I reformulated the goal and aimed at distinguishing among different conformations of large molecules forming a limited set of test structures. Indeed, obtained results showed that the reference conformation of three Ig domains was correctly identified among 290 proposed structures as the most probable one to give rise to the generated diffraction images. The posterior probabilities obtained using only 200 simulated diffraction images with on average ~ 380 photons did suffice to detect structural changes as small as 0.6 Å measured in terms of RMSD to the reference structure.

Structural changes of large biomolecular systems might span several length scales for their individual components. Hence I investigated whether the Structural Bayes approach is able to distinguish among different states of a complex molecule. First, I asked if the overall structural change can be detected in terms of calculated posterior probabilities. Indeed, it was possible to correctly identify a reference structure among seven different translocation states of the 70S ribosome. The test structures were compared against 200 simulated diffraction images of the reference structure, and any structural deviation from the reference resulted in a lowered probability of a particular structure. This result suggests that the Structural Bayes approach can be used to differentiate between states of a complex biomolecule by detecting overall structural changes; also when those changes result from local structural changes, happening at multiple length scales, of individual components.

In certain biological processes, important structural changes happen in confined regions of a complex system. Thus the next challenge for the Structural Bayes approach was to trace local structural changes against large structural background. As a test case, tRNA configurations native to seven ribosomal translocation states were embedded in a subunit arrangement of a selected state. These seven chimera structures were compared against 200 diffraction images generated from the reference state. The reference tRNA configuration was correctly identified as the most probable one. Moreover, the tRNA displacement in the translocation process was traceable in terms of decreasing structure probabilities.

Next, to test the robustness of the developed method to inaccuracies in the structural background, I introduced an inaccuracy in the structural background model. With an incorrectly modelled subunit arrangement, it was still possible to identify the reference tRNA chain configurations as the most probable among the seven test structures. Despite the introduced inaccuracy, the tRNA displacement was still reflected in decreasing posterior probabilities along the translocation process. The results obtained for the ribosome suggest that the planned XFEL single molecule experiments might be applied to study localized structural changes even of small regions of interest against a large structural background. Extracting such structural information might help to understand mechanisms of, e.g. ligand binding or enzymatic reactions. The conformational changes of a small ligand within a binding pocket or a substrate in a catalytic reaction centre, traced in single molecule XFEL experiments, might give insights into the mechanisms governing these processes.

Further, to reduce computational cost, I studied a multiscale structure model applied to distinguish among complex structures. Less important regions were modelled at a lower resolution (coarse grain representation) whereas the atomic details were maintained in the important parts of the investigated molecules. Similar to the previously described all atom structure models, the multiscale structure models of ribosomal translocation states allowed to trace tRNA displacement at atomic resolution against the coarse grained subunit arrangements. Such multiscale structure models might also facilitate a *de novo* structure refinement of larger biomolecules.

In summary, the results presented in this work suggest that it should be possible to extract structural information at atomic resolution from single molecule XFEL experiments for molecules of various sizes, ranging from small peptides to large complexes. The challenging experimental conditions, e.g. anticipated low photon counts and substantial background noise in the diffraction images, can be addressed by the two Bayesian approaches described here to solve structures of various molecular sizes. In contrast to X-ray crystallography, solving structures of single molecules at a high resolution should be less challenging with their increasing sizes. These results combined with no requirement of crystalline specimen suggest that single molecule XFEL scattering experiments might indeed become a powerful tool for structure determination and help to understand mechanisms governing biological processes.

6 Acknowledgments

I would like to thank my supervisor Prof. Dr. Helmut Grubmüller for introducing me to this very interesting subject, fruitful discussions and valuable advice. It has been a privilege to work in a friendly and stimulating atmosphere in the Department of Theoretical and Computational Biophysics. I am also grateful to Prof. Dr. Marcus Müller for his helpful suggestions regarding my thesis.

Further, I would like to thank my colleagues from the Department of Theoretical and Computational Biophysics, in particular Carsten Kutzner for helping me to debug my source code, Christian Blau, Ludger Inhester, Benjamin von Ardenne and Plamen Dobrev for discussions, and Jelger Risselada for sharing his insights into various matters. Petra Kellers, Ludger Inhester, Andrea Vaiana and Benjamin von Ardenne kindly agreed to proofread my thesis, their remarks greatly improved the readability of this thesis. I would like to thank Lars Bock, Christian Blau and Andrea Vaiana for providing me with the ribosomal structures. I am also indebted to our secretary Eveline Heinemann, who is always of great help when dealing with formalities, as well as our system administrators Ansgar Esztermann and Martin Fechner for promptly solving any IT problems.

This work was supported by the Deutsche Forschungsgemeinschaft, grant SFB 755/B4.

Bibliography

[1] U Hensen, T Meyer, J Haas, R Rex, G Vriend, and H Grubmüller. Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PLoS ONE*, 7(5):e33931, 2012.

[2] J Czub and H Grubmüller. Torsional elasticity and energetics of F1-ATPase. *Proc. Natl. Acad. Sci. USA*, 108(18):7408–7413, 2011.

[3] L V Bock, C Blau, G F Schröder, I I Davydov, N Fischer, H Stark, M V Rodnina, A C Vaiana, and H Grubmüller. Energy barriers and driving forces of tRNA translocation through the ribosome. *Nat. Struct. Mol. Biol.*, 20(12):1390–6, 2013.

[4] 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. *J. Phys.: Condens. Matter*, 21(13):134014, 2009.

[5] Richard Neutze and Keith Moffat. Time-resolved structural studies at synchrotrons and X-ray free electron lasers: opportunities and challenges. *Curr. Opin. Struct. Biol.*, 22(5):651–659, 2012.

[6] F H C Crick and Beatrice S Magdoff. The theory of the method of isomorphous replacement for protein crystals. I. *Acta Crystallogr.*, 9(10):901–908, 1956.

[7] W A Hendrickson. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, 254(5028):51–58, 1991.

[8] M G Rossmann. The molecular replacement method. *Acta Crystallogr. Sect. A*, 46:73–82, 1990.

[9] H N Chapman, P Fromme, A Barty, T A White, R A Kirian, A Aquila, M S Hunter, J Schulz, D P DePonte, U Weierstall, et al. Femtosecond X-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.

[10] T R M Barends, L Foucar, S Botha, R B Doak, R L Shoeman, K Nass, J E Koglin, G J Williams, S Boutet, M Messerschmidt, et al. De novo protein crystal structure determination from X-ray free-electron laser data. *Nature*, 505(7482):244–247, 2014.

[11] 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.*, 3:1276, 2012.

[12] 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.

[13] 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.

[14] K J Gaffney and H N Chapman. Imaging atomic structure and dynamics with ultrafast X-ray scattering. *Science*, 316(5830):1444–1448, 2007.

[15] J Miao, D Sayre, and H N Chapman. Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects. *J. Opt. Soc. Am. A*, 15(6):1662–1669, 1998.

[16] J Miao, K O Hodgson, and D Sayre. An approach to three-dimensional structures of biomolecules by using single-molecule diffraction images. *Proc. Natl. Acad. Sci. USA*, 98(12):6641–6645, 2001.

[17] G Oszlanyi and A Suto. Ab initio structure solution by charge flipping. *Acta Crystallogr. Sect. A*, 60(2):134–141, 2004.

[18] 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 Crystallogr. Sect. A*, 64(2):303–315, 2008.

[19] V Elser. Solution of the crystallographic phase problem by iterated projections. *Acta Crystallogr. Sect. A*, 59(3):201–209, 2003.

[20] D R Luke. Relaxed averaged alternating reflections for diffraction imaging. *Inverse Problems*, 21(1):37–50, 2005.

[21] G Huldt, A Szoke, and J Hajdu. Diffraction imaging of single particles and biomolecules. *J. Struct. Biol.*, 144(1–2):219–227, 2003.

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

[23] D Giannakis, P Schwander, and A Ourmazd. The symmetries of image formation by scattering. I. Theoretical framework. *Opt. Express*, 20(12):12799–12826, 2012.

[24] Ne-Te Duane Loh and V Elser. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E*, 80(2):026705, 2009.

[25] M Tegze and G Bortel. Atomic structure of a single large biomolecule from diffraction patterns of random orientations. *J. Struct. Biol.*, 179(1):41–5, 2012.

[26] H Liu, B K Poon, D K Saldin, J C H Spence, and P H Zwart. Three-dimensional single-particle imaging using angular correlations from X-ray laser data. *Acta Crystallogr. Sect. A*, 69(4):365–373, 2013.

[27] B von Ardenne. Reconstruction of electron densities from few photon single molecule x-ray scattering experiments. Master's thesis, Georg-August-Universität Göttingen, 2012.

[28] M Walczak and H Grubmüller. Bayesian orientation estimate and structure information from sparse single molecule x-ray diffraction images. *Phys. Rev. E*, 90(2):022714, 2014.

[29] U von Toussaint. Bayesian inference in physics. *Rev. Mod. Phys.*, 83(3):943, 2011.

[30] H J Risselada, C Kutzner, and H Grubmüller. Caught in the Act: Visualization of SNARE-Mediated Fusion Events in Molecular Detail. *ChemBioChem*, 12(7):1049–1055, 2011.

[31] R Santra. Concepts in x-ray physics. *J. Phys. B*, 42(2):023001, 2009.

[32] 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):225501, 2010.

[33] R Miles. On random rotations in R3. *Biometrika*, 52:636–639, 1965.

[34] M Galassi, J Davies, J Theiler, B Gough, G Jungman, P Alken, M Booth, and F Rossi. *GNU Scientific Library Reference Manual.* Network Theory Ltd., 2009.

[35] M Matsumoto and T Nishimura. Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM T. Model. Comput. S.*, 8(1):3–30, 1998.

[36] M Frigo and S G Johnson. The design and implementation of FFTW3. *Proc. IEEE*, 93(2):216–231, 2005.

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

[38] B Hess, C Kutzner, D van der Spoel, and E Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[39] G A Kaminski, R A Friesner, J Tirado-Rives, and W L Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105(28):6474–6487, 2001.

[40] T Darden, D York, and L Pedersen. Particle mesh Ewald: An N.log (N) method for Ewald sums in large systems. *J. Chem. Phys*, 98(12):10089, 1993.

[41] G Bussi, D Donadio, and M Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys*, 126(1):014101, 2007.

[42] B Hess, H Bekker, H J C Berendsen, and Johannes G E M Fraaije. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[43] M Moakher. Means and averaging in the group of rotations. *SIAM J. Matrix Anal. Appl.*, 24(1):1–16, 2002.

[44] L Schermelleh, R Heintzmann, and H Leonhardt. A guide to super-resolution fluorescence microscopy. *J. Cell Biol.*, 190(2):165–175, 2010.

[45] 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.

[46] D Nilsson, F Uhlén, J Reinspach, H M Hertz, A Holmberg, H Sinn, and U Vogt. Thermal stability of tungsten zone plates for focusing hard x-ray free-electron laser radiation. *New J. Phys.*, 14(4):043010, 2012.

[47] N Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller, E Teller, et al. Equation of state calculations by fast computing machines. *J. Chem. Phys*, 21(6):1087, 1953.

[48] E von Castelmur, M Marino, D I Svergun, L Kreplak, Z Ucurum-Fotiadis, P V Konarev, A Urzhumtsev, D Labeit, S Labeit, and O Mayans. A regular pattern of Ig super-motifs defines segmental flexibility as the elastic mechanism of the titin chain. *Proc. Natl. Acad. Sci. USA*, 105(4):1186–1191, 2008.

[49] N Fischer, A L Konevega, W Wintermeyer, M V Rodnina, and H Stark. Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature*, 466(7304):329–333, 2010.

[50] D Baker and A Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.

[51] K Ginalski. Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.*, 16(2):172–177, 2006.

[52] M Selmer, C M Dunham, F V Murphy, A Weixlbaumer, S Petry, A C Kelley, J R Weir, and V Ramakrishnan. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, 313(5795):1935–1942, 2006.

[53] A Korostelev, S Trakhanov, M Laurberg, and H F Noller. Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell*, 126(6):1065–1077, 2006.