Bayesian Structural Ensemble Determination from Single-Molecule X-ray Scattering

Dissertation
for the award of the degree
Doctor rerum naturalium
of the Georg-August-Universität Göttingen
within the doctoral program
International Max Planck Research School
Physics of Biological and Complex Systems
of the Georg-August University School of Science

submitted by
Steffen Schultze
from Jever

Göttingen, 2023

Thesis Committee

Prof. Dr. Helmut Grubmüller (first referee)
Max-Planck-Institute for Multidisciplinary Sciences
Department of Theoretical and Computational Biophysics

Prof. Dr. Simone Techert (second referee)
Georg-August-Universität Göttingen, Institut für Röntgenphysik
and Deutsches Elektronen-Synchrotron

Prof. Dr. Thorsten Hohage
Georg-August-Universität Göttingen
Institut für Numerische und Angewandte Mathematik

Examination Board

Prof. Dr. Helmut Grubmüller

Prof. Dr. Simone Techert

Prof. Dr. Thorsten Hohage

Prof. Dr. Holger Stark
Max-Planck-Institute for Multidisciplinary Sciences
Department of Structural Dynamics

Prof. Dr. Jörg Enderlein
Georg-August-Universität Göttingen
III. Physikalisches Institut

Prof. Dr. Stefan Klumpp
Georg-August-Universität Göttingen
Institut für Dynamik komplexer Systeme

Date of the Disputation: November 24, 2023

Abstract

Single-molecule X-ray scattering experiments using ultrashort X-ray free electron laser (XFEL) pulses have opened a new route for the structure determination of biomolecules. They also hold the potential to extract the structural ensemble of a molecule without the need for synchronization. In these experiments, a stream of single copies of the molecule to be studied enters the pulsed XFEL beam, and for each pulse, the scattered photons are recorded as a scattering image. However, structure refinement from single-molecule X-ray scattering images is quite challenging due to unknown molecular orientations, typically very low numbers of recorded photons per scattering image, and low signal-to-noise ratios in this extreme Poisson regime.

In the first and main part of this thesis I therefore develop and assess a novel Bayesian approach and demonstrate that it should be possible to determine not only a single structure, but an entire structural ensemble from these experiments. This approach allows for the systematic treatment of noise and other complicating experimental effects and, simultaneously, eliminates the need for classification, hit selection, and orientation determination. In fact, I explicitly include many complicating experimental effects, such as Ewald curvature, intensity fluctuations, hits vs. misses, beam polarization, irregular detector shapes, incoherent scattering and background scattering.

On the single structure level, I demonstrate that my approach can achieve near-atomistic resolutions for the protein crambin from noise-free synthetic scattering images, and that it achieves the same resolution of 9 nm from experimental data for the coliphage PR772 virus as previous approaches, using only a very small fraction of the available data. On the structural ensemble level, I demonstrate that my approach can determine the conformational ensemble of alanine dipeptide and even the unfolded ensemble of the mini-protein chignolin. I further demonstrate using synthetic images that my approach can reliably determine electron densities even in the extreme low hit rate and high noise regime.

Further, I systematically analyze the scaling behavior of my approach, finding, for instance, that the number of images required to determine a structural ensemble is proportional to the square of the number of conformers, that the amount of structural information per image is proportional the square of the number of photons, and that already a small amount of noise strongly decreases the achievable resolutions.

In a second part of this thesis, I present an analysis of time-lagged independent component analysis (tICA), a widely used dimension reduction method for the analysis of molecular dynamics trajectories. I seek to understand how much information on the actual protein dynamics is contained in the tICA-projections of MD-trajectories, as opposed to noise due to the inherently stochastic nature of each trajectory. To that end, I analyze the tICA-projections of high dimensional random walks using a combination of analytical and numerical methods, finding that they resemble cosine functions and strongly depend on the lag time, exhibiting strikingly complex behavior. Further, I demonstrate that the tICA-projections of protein trajectories can indeed be strikingly similar to those of random walks, suggesting that not only the ensemble properties of the non-converged protein trajectories resemble those of random walks, as has been shown earlier via PCA, but also the time correlations of the underlying protein dynamics.

Contents

Chapter 1

Introduction

This thesis consists of two independent parts. The first and main part focuses on a novel Bayesian approach for structure determination from single-molecule X-ray scattering, and the second part on an analysis of the dimension reduction technique time-lagged independent component analysis (tICA), specifically its application to high-dimensional random walks. This chapter serves as an introduction for both of these parts.

1.1 Single-molecule X-ray scattering

Proteins are the primary building blocks of all known living organisms. They perform a vast variety of functions, from the catalysis of biochemical reactions as enzymes to the defense against foreign invaders as antibodies, to signal transport and to DNA-replication. Almost every protein adopts a specific three-dimensional configuration, its *structure*, determined by its sequence of amino acids. However, proteins are far from being static objects, featuring exceedingly complex and high-dimensional dynamics, structural heterogeneity, and conformational transitions between different distinct conformations or metastable structures [1, 2].

Knowledge of both protein structures and protein dynamics is required to understand their functions and, in particular, the underlying functional mechanisms. Structural models are also required for drug design or biotechnology. However, determining the structure of a protein, or even its dynamics, is a far from trivial task, due to the fact that each protein has an astronomically large number of potential structures despite folding spontaneously, known as Levinthal's paradox [3, 4].

Many experimental structure determination techniques have therefore been developed, including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryogenic electron-microscopy (cryo-EM). Most recently great advances have been made in particular for cryo-EM, which is used for a quickly growing number of determined structures [5]. Despite this, so far only the structure of about 17% of all known human proteins has been determined experimentally [6].

The main reason is the time consuming and resource-intensive nature of the three main structure determination techniques used today, but also other limitations are important. For instance, crystallography requires the protein to be crystallized, which for many proteins is difficult to impossible [7]. NMR is limited to small proteins of typically less than 30 kDa [8], although recent advances have softened this limit [9–11], and requires substantial quantities of the sample

molecule in solution, with up to millimolar concentrations that may be difficult or impossible to achieve [9]. Cryo-EM, in turn, requires larger proteins of typically more than 50 kDa and suffers from low signal-to-noise ratios [12–14]. Finally, both cryo-EM and X-ray crystallography are fundamentally limited by the required unphysiological conditions of freezing to cryogenic temperatures or crystallization [15].

Even more challenging than single structure determination is to obtain additional ensemble information or even a time-resolved structure. The ultimate goal here is akin to a 'molecular movie' of the transitions between conformational states or of the movement involved in the protein function [16]. An important step towards this goal is to determine all or most of a proteins conformational states with their corresponding populations, that is, its structural ensemble or distribution. If the resolution of this ensemble is high enough, it can be used in combination with molecular dynamics simulations to construct conformational motions [17].

However, the established experimental structure determination approaches generally do not provide an entire structural ensemble, only averages or superpositions of it [18]. X-ray crystallography is particularly limited by this averaging, being fundamentally based on collective observation of many particles. NMR has been successfully used to resolve ensembles, but is also based on ensemble-averaged information [9]. In contrast, cryo-EM is based on many individual observations of single particles, such that structural ensembles can be inferred [19, 20].

As an alternative or extension to experimental structure determination techniques, recently great advances have been made in the prediction of protein structures from their sequence alone. Particularly the recent development of neural network models for structure prediction, such as AlphaFold2 [21] or RoseTTAFold [22], has been successful. Indeed, it is estimated that the fraction of known human protein structures would increase to up to 76% with the inclusion of such predictions [6]. However, these predicted structures are less reliable than experimentally determined ones [23].

While this success even lead some to state that structural biology is now 'solved' [24], such artificial intelligence approaches have so far not been successfully applied to predict dynamics or structural ensembles [24, 25]. One important reason for this limitation is the so far too small amount of experimentally known protein structural ensembles [25]. The further development of methods to determine structural ensembles from established approaches such as cryo-EM, but also the development of new experiments will be essential for the further understanding of protein dynamics [25].

The development of brighter X-ray sources in the form of X-ray free electron lasers has opened the way towards one class of such novel experiments. In these 'diffraction before destruction' experiments [26], a high intensity X-ray pulse is scattered on a small sample, and the ultrashort pulse duration ensures that all scattering occurs before the sample is destroyed by the Coulomb explosion due to the high intensity [27].

Most current experiments use the high intensity of these XFELs to collect useful scattering data from much smaller nano-crystals. Termed 'serial femtosecond crystallography', these scattering experiments have been very successful, for both static [28, 29] and time-resolved

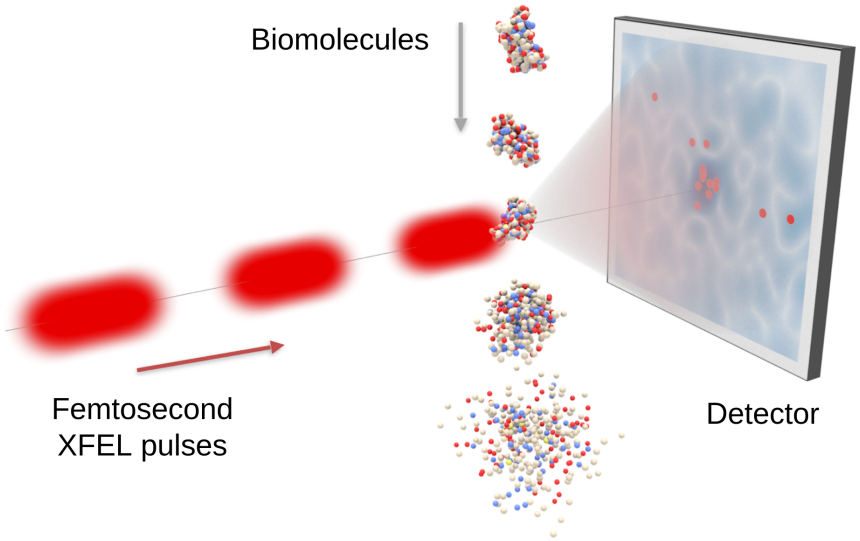

Figure 1.1: Single-molecule X-ray scattering experiment. A stream of single molecules is hit by femtosecond X-ray free electron laser pulses, and for each pulse the scattered photons are recorded as images. Reproduced from von Ardenne et al. [36].

structures [30, 31]. However, these nano-crystals still average over many molecules, such that for time-resolution, all or most molecules in one crystal must be in the same conformational state. Therefore, synchronization is required for molecular movies, typically by optical laser pulses [32]. Consequently, most studies focus on light-induced conformational changes in light-sensitive samples, for example the receptor protein rhodopsin [30].

Taking things to the extreme, it was proposed to instead perform X-ray scattering experiments on single molecules, foregoing the crystal altogether [27]. In such experiments (Figure 1.1), a stream of single copies of the molecule to be studied enters the pulsed XFEL beam, and for each pulse, the scattered photons are recorded as a scattering image [27]. The high repetition rates of current XFELs (for example 27,000 Hz for the European XFEL [33]) allow the collection of millions of such images in reasonable times, despite low hit rates of less than 1 % [34]. Because the scattering happens not on a crystal but on a single molecule, this means that each scattering image is a snapshot of the molecule in one single state with femtosecond exposure time [35].

With these advantages over other methods, single-molecule X-ray scattering experiments are therefore, in principle, in an ideal position for the determination of protein structural ensembles and the construction of molecular movies [16, 37, 38]. They are, however, much more challenging, both from an experimental point of view and due to the lack of sufficient analysis methods, such that the vast majority of current experiments focuses on nano-crystals [28, 32, 39–44].

The analysis of single-molecule scattering experiments is challenging mainly because the molecular orientation is random and unknown for each scattering image [27]. Further, the number of scattered photons is small, the hit rates are low, and the signal-to-noise ratio is very low in this extreme Poisson regime [27]. In contrast, for scattering on nano-crystals, the orientation is also random but not unknown because in this case it can be determined from the much stronger scattering signal [29]. Indeed, whereas the feasibility of this approach has been demonstrated by a number of proof-of-principle experiments, successful refinement of structures or electron

densities has so far been limited to large specimen like entire viruses [45–47]. For such large specimen, up to 10^7 photons are coherently scattered for each hit, as opposed to only $10 - 50$ expected photons for typical proteins.

Because single-molecule X-ray scattering images are inaccessible to conventional analysis methods, many new methods have been developed. Most of these methods apply the same general principle, aiming first to determine the molecular orientation for each image from the positions of the scattered photons and subsequently to assemble the properly oriented images in Fourier space into a full three-dimensional scattering intensity [48–55]. From this intensity, the electron density is then determined using a phase retrieval algorithm [56–58]. Notably, in contrast to crystallography, here a continuous intensity distribution is obtained as opposed to a pattern of Bragg-peaks, such that this phase retrieval is indeed possible [59].

Multiple different instances of these orientation determination methods exist. The common line method, first proposed by Huldt et al. [60], utilizes the fact the any two scattering images share, as the name suggests, a common line (or, rather, curve on the Ewald sphere) on which they intersect in Fourier space [48, 60, 61]. By identifying this common lines for any two or three images, their relative orientation is estimated, and these relative orientations are combined to determine the orientation for each image. Bortel et al. proposed a variation of the common line method that reduces the number of required pattern comparisons and first demonstrated orientation determination with realistic target parameters [61]. However, to determine this common line for two diffraction patterns, an average photon count of at least 10 per pixel is required [48], corresponding to more than 16,000 photons per image. It was therefore proposed to classify the images into similar groups and average them to increase the photon counts; this, however, also requires many photons per pixel [60].

Because of the exceedingly high photon count requirements of the common line method, alternative approaches have been developed. Loh and Elser proposed the EMC algorithm [49], which iteratively fits orientation estimates to a model for the intensity function. Each iteration involves first an expansion step, which generates all possible diffraction patterns of the current intensity model, second a maximization step, which estimates the orientations with their likelihoods given these diffraction patterns and updates the diffraction patterns to maximize the likelihood, and third a compression step, which determines a new estimate for the intensity function from these orientations [49]. For each image, not a single orientation, but the likelihood distribution of the orientation conditioned on the current model for the intensity function is determined [49].

The EMC algorithm requires much lower photon counts than the common line algorithm, and is, in fact, the current method of choice for orientation determination [34, 55, 62], and successful structure determination has been demonstrated both on synthetic data and on experimental data. For instance, it was used to determine the structure of GroEL molecules at 2 nm resolution from 106 simulated images [49], and to determine the structure of a coliphage PR772 virus at 9 nm resolution from experimental images, although the latter with explicitely imposed icosahedral symmetry [62]. However, this approach still requires a few hundred photons per image and suffers from low convergence rates for low photon counts [49]. While a successful application of

the EMC algorithm has been reported for less than 100 photons per image [63], these photon counts referred to photon counts outside of the central speckle, with the total photon counts per downsampled image still above 400. It has been proposed to use a low resolution structure of the studied protein as a seed model [50] to improve convergence. This seeding does indeed decrease the required photon counts to 10-100 photons per image [50], but such seed models are not always available.

Because the EMC algorithm is computationally very expensive, improvements have been proposed, for example using spherical harmonics expansions [54]. Closely related to the EMC algorithm are the correlation maximization algorithm, which, instead of averaging, selects only one correlation-maximizing orientation in the maximization step [53, 64] and the multitiered iterative phasing (M-TIP) algorithm, in which a similar iterative procedure is combined with constraints on the electron density [65].

As a further alternative, Fung et al. proposed to use a general class of so called manifold embedding algorithms [66]. These algorithms exploit correlations within the set of all scattered photons which arise from the fact that the Fourier intensity function is continuous, such that similar orientation lead to similar diffraction patterns [66]. Among these, it was first proposed to determine a maximum likelihood manifold in orientational space using generative topographic mapping, which was then used to classify the image orientations and subsequently average the diffraction patterns [66]. More recently, the diffusion map algorithm was developed, exploiting additional symmetries in the image formation [67–69]. Closely related to manifold-based approaches, Kassemeyer et al. proposed to exploit the same similarities between scattering images using routing algorithms and geodesic distances on the rotation group [51]. Whereas manifold embedding algorithms can utilize the scattering images far more efficiently than the common line method, they still require at least 100 photons per image [67–70]. Further, despite not explicitly determining the orientation, manifold embedding algorithms still function fundamentally similar to orientation determination methods; in fact, it has been argued they and the EMC algorithm are merely different implementations of the same fundamental approach [70].

In contrast, a further class of methods based on photon correlations foregoes determining the orientations altogether [36, 71–78]. These methods extract correlations between photons as a summary statistic of the full set of all images, which are used as an intermediate representation for the electron density determination. It was first proposed by Saldin et al. to use the two-photon correlation function [71], which, however, only suffices to determine the structure under symmetry assumptions [72] or if the random orientation is restricted to one axis [73–75]. Therefore, von Ardenne et al. suggested to use the three-photon correlation function, which, despite using only one additional photon, does allow determining the molecular structure without such restrictions [36]. The three-photon correlation function measures the frequency of each photon triplet, described by three distances to the detector center and two angles, among all possible triplets [36]. It was in fact already known much earlier that a degenerate three-photon correlation, which uses only those photon triplets for which two of the three photons share the same position, does, in principle, suffice for this unrestricted structure determination [79]. Notably, the three-photon correlation function can be expressed analytically from the spherical harmonics expansion of the Fourier intensity function [36, 79]. This analytic expression allows

for highly efficient computations, which were utilized by von Ardenne et al. in a Monte-Carlo simulated annealing scheme to determine a maximum likelihood Fourier intensity function from the three-photon correlation [36].

With only three photons per image the three-photon correlation method requires much fewer photons counts than other methods — in fact, because the two-photon correlation is not sufficient, three is the fundamental minimum [36]. However, photon correlation methods use only part of the available information, for instance discarding higher correlations, and thus rely on the collection of large amounts of data and averaging to address Poisson noise, detector noise, and incoherent scattering [36].

Considering the current main challenge of determining protein dynamics or structural ensembles as opposed to static structures, further limitations of these approaches become apparent. To determine a structural ensemble, images need to be classified into different sets, one for each conformer, and, so far, such classification algorithms have only been reported to be successful for more than a thousand photons per image [62, 80, 81].

A similar challenge is posed by the fact that sample delivery is successful only for a fraction of typically less than 1% of the pulses, such that 99% of the images are 'empty' but nevertheless contain noise photons [34]. Because the beam intensity at the position of the sample molecule fluctuates strongly, separating the actual scattering images ('hits') from the empty ones is not always possible, particularly at the low photon counts expected for small proteins. However, most established approaches rely on a classification into such hits and misses, with the exception of those based on photon correlations [34].

The current method of choice for classification is the diffusion map algorithm [67–69], which can straightforwardly be extended for both hit and conformer classification, but so far fails for the low signal-to-noise ratio of small proteins [34, 82]. Many other hit and conformer classification approaches exist [83, 84], including, for example, spectral clustering [85]. It has also been proposed to identify the hits by experimental means, for instance using ion Time-of-Flight spectrometers [84]. In contrast to the other established approaches, photon correlation methods do not rely on hit selection, as the correlations in the empty images should average out. It has also been suggested that it may be possible to determine conformational landscapes from the three-photon correlation function without classification, although this has not yet been demonstrated [36].

From a more general perspective, the challenges of single-molecule X-ray scattering share many similarities with cryo-EM, which also involves images of single randomly oriented particles that need to be identified ('picked'), albeit at a much lower noise level [14]. Also the spectrum of available analysis methods is similar — though much more mature — involving orientation determination and classification algorithms as parts of larger software packages [86–91] as well as sophisticated machine learning approaches [19], Bayesian approaches [92], and attempts to circumvent particle picking [93]. Here, too, the current main challenge and limiting factor is structural heterogeneity [94].

Finally, the established approaches do not systematically include noise and other experimental effects in the analysis, instead relying on, for example, averaging and subtraction of background

scattering [34]. Often, they neglect effects such as polarization, irregular detector shapes, or detector noise [36, 49, 50, 60]. I am not aware of any single-molecule scattering electron density determination method that can systematically include all experimental noise and resolve ensembles which also does not rely on classification, hit selection, and orientation determination.

A Bayesian approach would in fact allow for a systematic treatment of noise and, simultaneously, eliminate the need for classification, hit selection, and orientation determination. In such an approach, the electron density or ensemble of electron densities maximizing the Bayesian posterior probability given the whole set of typically millions of images is determined directly. The 'only' requirement is an accurate forward model, that is, a full mathematical description of the distribution of the scattering images. To account for noise and other experimental effects, they only have to included in this forward model. Importantly, the Bayesian posterior inherently contains all available structural information, such that no information of the experimental data is lost, which is particularly important considering the limited operational capacity of current XFEL facilities. The Bayesian framework can also provide error bounds and uncertainty estimates, which is particularly useful when only a limited amount of information is available.

The first and main part of this thesis is therefore the development, assessment, and application of such a rigorous Bayesian approach to determine protein structures and structural ensembles from single-molecule X-ray scattering experiments. To deal with the considerable sampling challenge due to the many degrees of freedoms of protein structures, I developed a novel specialized hierarchical Monte Carlo simulated annealing technique, which enhanced the sampling performance by a factor of at least hundreds or thousands. To further deal with the substantial computational cost of computing the Bayesian posterior probabilities, I implemented and optimized the computations on graphics processing units (GPUs). The source code is available as a Julia [95] package at `https://gitlab.gwdg.de/sschult/xfel`.

Although the most important aspect is the determination of structural ensembles, I first tested if my approach can determine single structures. Because my approach uses all available information, I expect it to require fewer scattering images to achieve a certain resolution than, for example, correlation based methods. I assessed this aspect by using synthetic scattering images generated for the same 46 amino acids comprising protein crambin [96] that was used to test the three-photon correlation method [36].

I further tested my approach on the single structure level experimental scattering images collected for single coliphage particles [81]. These were downsampled by a factor of 10^4 to mimic the scattering images expected for proteins. Notably, my approach was able to determine the structure of the virus at the same detector-limited resolution of 9 nm as reported using established methods [36, 62, 63], while using only a small fraction of the available data.

Moving from single structures to structural ensembles, I selected alanine dipeptide and the artificial mini-protein chignolin [97] as test systems to assess how well my approach can determine structural ensembles. Using molecular dynamics trajectories to generate the synthetic scattering images, my approach was indeed able to reconstruct the structural ensembles for these test systems, notably including the unfolded ensemble of chignolin. Determining these structural ensembles required much fewer scattering images than expected, and, notably, also much fewer

than would be required for a single structure with the same number of degrees of freedom. To understand this, I further analyzed the scaling behavior of this required number of images in the number of determined conformers, and, for comparison, in the number degrees of freedom in one single structure. Strikingly, the required number of images turned out to be proportional to the square of the number of conformers.

Next, I asked how well my approach performs in the presence of noise from incoherent scattering and other complicating effects. To that end, I included many of the expected experimental effects in the likelihood function, like incoherent scattering, background scattering, polarization, irregular detector shapes, Ewald curvature, hits and misses, intensity fluctuations and unknown molecular orientations. As before, I used crambin as the test structure, and demonstrated how its electron density can still be determined from highly noisy low hit rate images, albeit so far only at a lower resolution. I also analyzed if, in theory, an orientation determination approach could also be successful in this scenario, again relying on the fact that my Bayesian approach uses all the available information.

Finally, I sought to understand how the approach scales in the number of expected photons per image and the amount of noise, and, ultimately, in the size of the sample molecule. These and the above scaling results will be important to plan future experiments, particularly when considering the limited operational capacity of current XFEL facilities.

1.2 Time-lagged independent component analysis.

As already mentioned, the alternative to experiments is to study protein dynamics theoretically or by simulation. The prevalent method here is molecular dynamics simulation, which has been applied with great success to elucidate protein function, protein folding, and protein interaction [98]. They simulate the movement of every atom in a molecular system like a protein [99]. However, with the number of these atoms ranging from several hundreds to hundreds of thousands or more [100–103], the extremely complex and high-dimensional dynamics pose a formidable problem for data analysis.

It has therefore been proposed to extract essential features of these dynamics, or, in other words, to reduce the dimension of the trajectories [104]. These essential dynamics can be extracted by principal component analysis (PCA), which identifies those collective degrees of freedom (that is, directions in configuration space) that contribute most to the variance in the atomic movements [104]. More recently, time-lagged independent component analysis (tICA) has been proposed, which instead identifies those collective degrees of freedom that exhibit the strongest time-correlations for a given lag-time [105, 106], that is, along which the atomic movements appear the ‘slowest’. Both PCA and tICA are mathematically formulated by an eigenvalue problem, with PCA computing eigenvectors of the autocorrelation matrix of a given trajectory and tICA computing generalized eigenvectors of a time-lagged version of this matrix [104, 105].

Both techniques yield information on the conformational dynamics of a protein. They are particularly useful as a preprocessing step to describe the conformational dynamics as a discrete

Markov model [107–109]. Here, tICA is most widely used, and it is generally preferred over PCA because it uses time information of the input trajectory [110].

An important and natural question is how much of the output of these dimension reduction techniques is determined by the protein dynamics, as opposed to noise due to the inherently stochastic nature molecular dynamics trajectories, and how this can be quantified. To answer this question, their output when applied to molecular dynamics trajectories can be compared with their output when applied to a featureless purely random trajectory, that is, a random walk, which, by construction, does not contain any correlations or collective motions.

For PCA, the question has been answered by this comparison, with the highly unexpected finding that the principal components of such a random walk appear not at all random and featureless, instead turning out to be cosine functions [111, 112]. This finding offers a way to measure the convergence of a molecular dynamics trajectory, the so called cosine-content [111]: The more the principal components of a trajectory resemble a cosine function, the more it behaves like a random walk, and the less information on the protein dynamics it contains.

For tICA, no such analysis has so far been achieved, but similar affects seem to occur [113, 114]. In the second part of this thesis, I therefore analyzed the tICA-projections of high dimensional random walks, obtaining a semi-analytical expression for random walk tICA-projections, analogous to the cosine-function for PCA. Strikingly, the tICA-projections display a significantly more complex behavior, notably very similar for random walks and protein trajectories, such that they may serve as more sensitive tool to analyze convergence.

Thesis overview

In the first part of this thesis, consisting of Chapter 2 and 3, my Bayesian approach for single-molecule X-ray scattering is developed and assessed.

Chapter 2 introduces my approach for both single electron density and structural ensemble determination, in this instance for noise-free synthetic images and downsampled experimental images from a virus data set. Here also the hierarchical sampling technique is presented.

In Chapter 3, experimental effects are included in the approach, such as incoherent scattering, background scattering, polarization, irregular detector shapes, hits and misses, and intensity fluctuations. Here it is also analyzed how the required number of image scales in, for example, the photon count and the noise level.

In the second part of this thesis, consisting of Chapter 4, the analysis of time-lagged independent component analysis is presented. Here, the semi-analytical expression for random walk tICA-projections is derived, and the tICA-projections of random walks and protein trajectories are compared.

Finally, in Chapter 5 the results are summarized and discussed, and an outlook on future research and possible improvements is given.

Chapter 2

Bayesian structural ensemble determination

A previous version of the following text is available on arXiv as

S. Schultze and H. Grubmüller: **De novo structural ensemble determination from single-molecule X-ray scattering: A Bayesian approach** [115]. This chapter has been submitted to Nature Communications but was unfortunately rejected in the second round of review.

I carried out the research and wrote the manuscript. Helmut Grubmüller supervised the research and revised the manuscript.

Abstract

Single molecule X-ray scattering experiments with free electron lasers have opened a new route to the structure determination of biomolecules. However, structure refinement is quite challenging due to unknown molecular orientations, typically very low numbers of recorded photons per scattering image, and low signal-to-noise ratios in this extreme Poisson regime. As a further layer of complexity, many biomolecules show structural heterogeneity and conformational transitions between different distinct structures; these structural dynamics are averaged out by existing refinement methods. To overcome these limitations, here we developed and tested a rigorous Bayesian approach and demonstrate that it should be possible to determine not only a single electron density map, but an entire structural ensemble from these experiments. We tested our approach on synthetic scattering images, resolving, for example, an ensemble of eight alanine dipeptide conformers at 2 Å resolution from 10^6 images. Tests on experimental data achieved the same resolution of 9 nm for the coliphage PR772 virus as previous approaches, using only a small fraction of the available data. These findings show that X-ray scattering experiments using state-of-the-art free electron lasers should allow one to determine not only biomolecular structures, but whole structure ensembles and, ultimately, construct 'molecular movies' from these ensembles.

2.1 Introduction

Ultrashort pulse X-ray scattering experiments offer the possibility to take 'snapshots' of biomolecular structures with atomistic spatial and femtoseconds time resolution [35, 60, 116, 117]. Still, most current experiments focus on nano-crystals [28, 32, 39–44]. Like classical X-ray crystallography, these average over many molecules and, therefore, time resolved structure determination requires strict synchronization, typically by optical laser pulses [32]. Scattering on single particles or molecules avoids this limitation and should enable us to advance towards structure ensembles and, ultimately, time resolved conformational and functional motions without the need for synchronization [37, 38].

In such 'hit and destroy' experiments, a stream of single molecules is exposed to a beam of high intensity femtosecond X-ray free electron laser (XFEL) pulses (Fig. 2.1a). For each hit the positions of the scattered photons (red dots) on the detector are recorded as a scattering image [27]. Importantly, the ultra-short pulses serve to outrun the subsequent destruction of the particles due to radiation damage, but also imply that only very few photons are being recorded for each molecule [35].

The feasibility of this approach has already been demonstrated by a number of experiments [45–47], but so far only structures of relatively large specimen at low resolution have been successfully determined, for instance of entire mimivirus particles [45, 46] (450 nm in diameter) and coliphage viruses [47] (20 nm in diameter). Whereas for large specimens many photons are scattered per image, for example 10^7 for the mimivirus [45, 46], for typical proteins only 10-100 coherently scattered photons per image are expected [118, 119], which further complicates structure determination particularly for small molecules. Such images can be obtained with an intensity of 10^{12} photons per pulse at 5 keV and a 1 µm beam diameter [36], for example from the XFELs at DESY or SLAC.

Most importantly, the orientation of the molecules at the time of scattering is typically unknown, which poses an additional and substantial refinement challenge. These issues are particularly challenging for the structure refinement of small specimen such as proteins or protein complexes at near-atomic resolution. Still, the scattering images contain much more information than a simple SAXS signal, due to the fact that all scattered photons in an image arise from the same molecular orientation. A number of methods have been proposed to extract this information, such as orientation determination methods [48–55, 65], which aim to estimate the molecular orientation for each image, and manifold embedding algorithms [66–68, 70]. All these methods, however, typically require 100 to 1000 photons per scattering image. As an alternative, correlation based approaches [71, 73–75, 77] have recently allowed substantial advancements and have been shown to require, quite counterintuitively, only three photons per image [36].

Finally, many biomolecules show structural heterogeneity and conformational dynamics between different distinct structures, which, when resolved, would provide a direct view on biomolecular function. Hence, and similar to the current main challenge in cryogenic electron microscopy [94], not one but many conformations need to be extracted from the scattering images. In this scenario, in addition to the orientation, also the current conformer for each scattering image is unknown. Very much like for the unknown orientation, current approaches rest on a classification of each

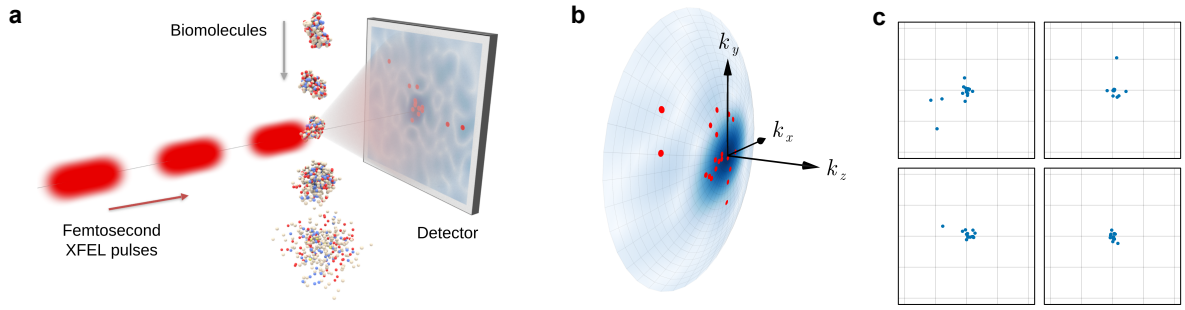

Figure 2.1: Single molecule scattering experiment. **a** A stream of single molecules is hit by femtosecond X-ray pulses, and the scattered photons are recorded as images (image reproduced from von Ardenne et al. [36]). **b** The scattered photons (red dots) are distributed on the Ewald sphere according to the 3D-intensity function I (blue). **c** Example scattering images generated for crambin. The axes each range from $-2\,\text{Å}^{-1}$ to $2\,\text{Å}^{-1}$.

single image, that is, assigning each image to a particular conformer of the molecule. Therefore, whereas both orientation determination methods and manifold-based methods have been applied to determine multiple conformations [62, 80], the required large number of photons per image, again, precludes their application to single biomolecules.

To overcome these limitations, we here developed and tested a rigorous Bayesian approach, which requires neither conformer classification nor orientation determination. Instead, a likelihood is computed for all possible structures and conformers, represented by electron density maps. This likelihood describes the probability that the resulting structure, or set of conformers, agrees with all scattering images, including a weighted average over all possible molecular orientations. Thus, rather than determining the orientation for each individual image and subsequently merging all oriented images into one 3D Fourier density, here we identify the real space electron density that is most likely to produce the observed whole set of scattering images.

This approach has two conceptual advantages. First, in contrast to classification or correlation based methods, *all* available experimental information is used. Second, the Bayesian framework only requires an accurate forward model of the experiment, that is, the probability distribution of all possible scattering images for a given structural ensemble and orientation. As a first step, we here focus at idealized experiments that only contain shot noise. In a second step, as a proof of principle demonstration using real experimental data on the coliphage PR772 virus, other noise sources or experimental uncertainties (such as incoherently scattered photons, background scattering, or detector noise) can be included within the Bayesian framework in a straightforward and rigorous manner, provided a sufficiently accurate error model is available.

Using synthetic data, our tests demonstrated that our Bayesian approach not only serves to determine a single electron density to near atomistic resolution, but also weighted ensembles of conformers described by multiple electron density maps. Remarkably, assuming the same average number of photons per image, we find that much fewer images are required for refining an ensemble of n conformers consisting of m atoms each than for refining a single structure consisting of $m \times n$ atoms. Despite identical complexity in terms of unknown positions, determining conformational ensembles seems to require less information than determining a single structure. This unexpected result should render biomolecular ensemble determination accessible to state-of-the-art experiments.

2.2 Results

2.2.1 Summary of the approach

For each scattering image, the positions of the recorded photons specify vectors $\mathbf{k}_1, \ldots, \mathbf{k}_l$ on the Ewald sphere in Fourier space (red dots in Fig. 2.1b). Here, the probability of observing a photon at a particular position on the detector is proportional to the 3D intensity function $I(\mathbf{k}) \propto |\mathcal{F}\{\mathbf{R}\rho_i\}(\mathbf{k})|^2$ at the corresponding position $\mathbf{k}$ on the Ewald sphere, which in turn is given by the Fourier transform of the electron density ρ_i of conformer i. Here, $\mathbf{R}$ is the unknown orientation of the molecule for this particular image, and $\mathbf{R}\rho_i$ denotes the electron density in this orientation.

It follows that the probability of observing an image with photon positions $\mathbf{k}_1, \ldots, \mathbf{k}_l$ is obtained by averaging over both the conformational ensemble $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_n\}$ with weights $\mathbf{w} = \{w_1, \ldots, w_n\}$ as well as over all orientations $\mathbf{R}$. Because the scattering images are statistically independent from each other, the total probability of observing the complete set of all images $\mathcal{I}$ reads

$$P(\mathcal{I} \,|\, \boldsymbol{\rho}, \mathbf{w}) \propto \prod_{(\mathbf{k}_1,\ldots,\mathbf{k}_l)\in\mathcal{I}} \sum_{i=1}^{n} w_i \int_{\mathrm{SO}(3)} P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \mathbf{R}\rho_i)\, \mathrm{d}\mathbf{R}\,. \tag{2.1}$$

This probability serves to determine either a single structure or a structural ensemble by sampling from the Bayesian posterior probability $P(\boldsymbol{\rho}, \mathbf{w} \,|\, \mathcal{I}) \propto P(\mathcal{I} \,|\, \boldsymbol{\rho}, \mathbf{w}) P(\boldsymbol{\rho}, \mathbf{w})$ using a Markov chain Monte Carlo approach. For the prior $P(\boldsymbol{\rho}, \mathbf{w})$ the orientations are assumed to be uniformly distributed. To minimize the number of required degrees of freedom, and as a means of regularization, we chose a physically motivated representation of each ρ_i in terms of a sum of Gaussian functions, which also completes the definition of the prior.

For a typical protein consisting of 50 to several hundred residues, the number of required degrees of freedom remains large and poses a formidable sampling challenge. To address this issue, we have implemented a hierarchical simulated annealing approach. Starting at very low resolution, the macromolecular structures were sampled in multiple hierarchical stages of increasing resolution. To increase the sampling efficiency, in each of these stages, for each Markov step the previous ensemble of structures of maximal posterior probability was used as a proposal density. To this end, the scattering images that would have been observed for a smoothed low resolution copy of the original molecule were obtained from the original images by rejection sampling using the convolution theorem (see the Methods section).

Further, we adapted the Bayesian formalism such that only those images are used which contain new information, that is, photons for which the magnitude of $\mathbf{k}$ is larger than the threshold of the resolution from the previous stage. With increasing resolution, the fraction of such useful images becomes very small, thus enhancing computational efficiency up to two orders of magnitude. The approach is described in detail in the Supplementary Information.

2.2.2 Sample test refinements

Because our Bayesian approach uses all available information, we expect it to require fewer scattering images to achieve a certain resolution than, for example, correlation based methods. To assess this aspect, we first tested our method on the single structure level, using the same 46-residue protein crambin [96] as in our previous study [36]. A total of 10^8 noise-free synthetic images were generated, containing a realistic average of 15 photons each (at an assumed intensity of 10^{12} photons per pulse with a beam diameter of 1 µm). From these images, the electron density was determined in five hierarchical stages (Fig. 2.2a), increasing the number of degrees of freedom by a factor of two in each stage. For the final stage, a representation of ρ consisting of 184 Gaussian functions was used, which is four times the number of residues. For more details see Supplementary Note 2.5.1. Indeed, using only half of the total number of scattered photons, a similar Fourier shell correlation resolution [120] of 4.2 Å (Fig. 2.2b) is obtained as with the previous correlation based method [36]. As explained in the Methods section, here the Fourier shell correlations serve to compare the reconstructed and reference electron density maps. As a further measure of quality, the optimal transport plan between the reconstructed and reference electron densities was computed using the Sinkhorn algorithm [121], obtaining an earth mover's distance of 1.45 Å.

Next, to demonstrate that our method can resolve not only a single protein structure, but also ensembles of multiple conformers, we used three molecular dynamics trajectories of alanine dipeptide [122] of length 250 ns each to generate 10^6 scattering images, using a randomly chosen snapshot for each image. As before, an average of 15 photons per image were generated, corresponding to a beam intensity of roughly $2.5 \cdot 10^{13}$ photons per pulse. Using our approach, a weighted ensemble of eight conformers was determined from these images (Fig. 2.3), with each conformer being described by a sum of 10 Gaussian functions. To obtain sufficient statistics, a total of 10 independent simulated annealing runs were carried out, using the same image set.

To assess the quality of the obtained conformational ensemble, for each of the eight conformers the resolution with respect to its nearest neighbor in the input trajectories was calculated using Fourier shell correlations [120] (Fig. 2.3b), resulting in a weighted average resolution of 1.8 Å. This result shows that the obtained eight conformers are indeed close to the reference ensemble. To also assess the accuracy of the entire ensemble, for each time step in the input trajectories, the resolution with respect to its nearest neighbor among all the determined conformers was calculated (Fig. 2.3d). As a main result we found that 90% of the input trajectories are within 2.1 Å Fourier shell correlation resolution of the determined conformers, and that all of the trajectory frames are within 2.5 Å resolution of the determined conformers, thus demonstrating atomistic resolution. Figure 2.3e compares the 10 obtained ensembles with the reference ensemble using a Ramachandran plot [123] showing the distribution of the torsion angles ϕ and ψ. For each of the determined conformers its nearest neighbor in the input trajectories was used to compute these angles. As can be seen, the reference density is well represented by the determined conformers. As an independent quality assessment, we calculated the optimal transport plan between reconstructed and reference ensembles of electron densities using the Sinkhorn algorithm [121]. Using the FSC resolution as the cost function, an earth mover's distance of 2.37 Å was obtained (see Supplementary Note 2.5.8).

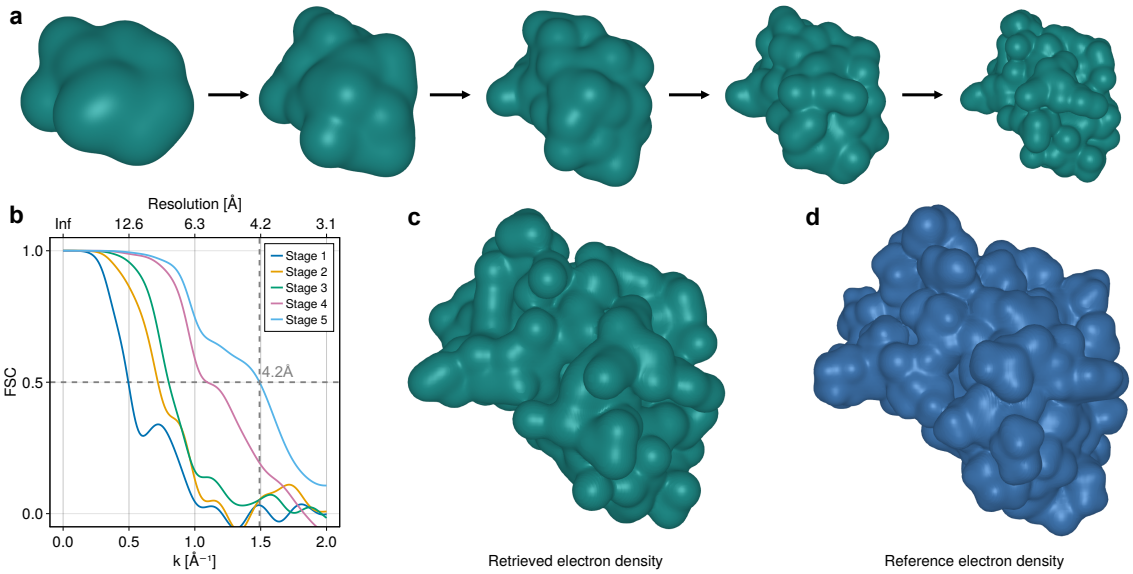

Figure 2.2: Electron density determination of Crambin. **a** Hierarchical stages of retrieved electron densities. **b** Fourier shell correlation between the retrieved densities and the reference density. **c** Retrieved electron density. **d** Reference electron density.

Next, we asked if our method is also capable of extracting structural ensembles for the larger mini-protein chignolin [97], comprising 10 residues. To that end, 50 molecular dynamics trajectories of length 10 µs were used to generate $1.2{\cdot}10^7$ images with, on average, 15 photons each, corresponding to a beam intensity of approximately $5 \cdot 10^{12}$ photons per pulse. As a further challenge, this ensemble also contained unfolded structures. From the obtained images, we determined multiple stages of weighted structural ensembles of increasing resolution and increasing number of conformers (Fig. 2.4a). As above, resolutions were computed using Fourier shell correlations (Fig. 2.4b,c), finding a weighted average resolution of 4.7 Å for the folded conformers, and 6.4 Å for the unfolded conformers. As can be seen in Fig. 2.4b, the Fourier shell correlation approaches the conservative threshold of 0.5 already for smaller values of k, which explains why, by visual inspection, some of the reconstructed conformers appear 'worse' than the above FSC resolutions. Therefore, resolutions of $4-6$ Å for the folded conformers and $6-12$ Å for the unfolded conformers seem more realistic. The independently calculated optimal transport cost between the reference and reconstructed ensembles of 6.96 Å (see also Supplementary Note 2.5.8) supports this estimate.

Interestingly, in the final stage one of the six determined weights is nearly zero, suggesting that five conformers suffice for the used number of images at this resolution level. While the current resolution level does not suffice to reliably distinguish the two compact conformers (commonly referred to as 'folded' and 'misfolded') from each other, which are both present in the reference structure ensemble at nearly equal proportion, it is worth noting that the 9% fraction of unfolded states was indeed correctly identified.

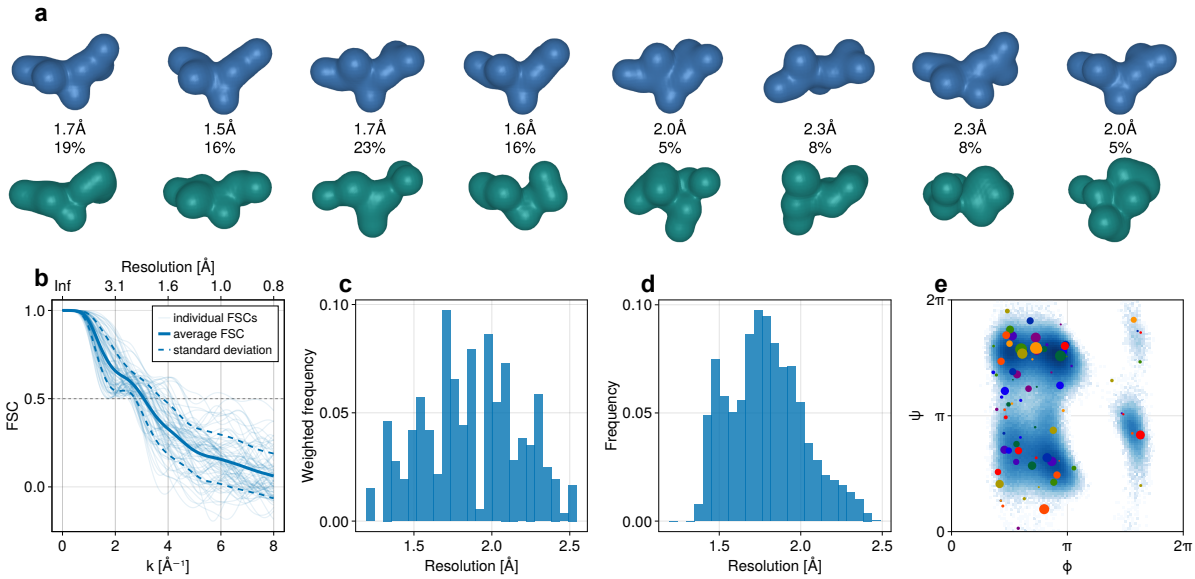

Figure 2.3: Structural ensemble determination of the alanine dipeptide. **a** Reconstructed conformers (green), the corresponding weights, and the nearest neighbors in the input trajectories (blue) with the corresponding resolutions. **b** Fourier shell correlations for the individual reconstructed conformers from the 10 independent runs, with average and standard deviation. **c** Weighted resolution distribution for the 10 independent runs from the the same data. **d** Resolution distribution over the time steps of the input trajectories relative to their nearest neighbors among the determined conformers from all 10 runs. **e** Ramachandran plot for the input trajectories (shown as a density) and the determined conformers from all 10 runs (points, the colors indicate the separate runs).

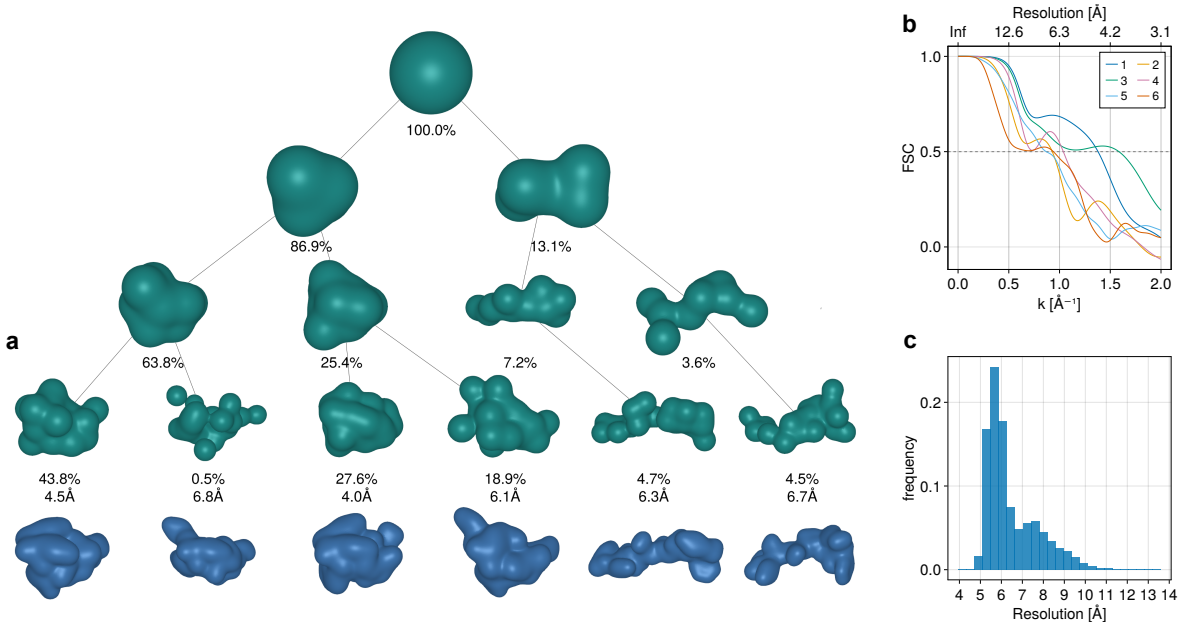

Figure 2.4: Structural ensemble determination for chignolin. **a** Hierarchical stages of retrieved densities (green) and their nearest neighbors (blue) in the input trajectories with the corresponding resolutions. **b** Fourier shell correlations of the reconstructed densities relative to their nearest neighbors (from left to right). **c** Resolution distribution over the time steps of the input trajectories relative to their nearest neighbors among the determined densities.

2.2.3 Scaling behavior

For both of the above sample applications we observed, unexpectedly, that resolving n conformers consisting of m degrees of freedom each required much fewer scattering images and photons than resolving a single $n \times m$ degree of freedom structure of the same total size and complexity — even in cases where the conformers of the ensemble are very different from each other. This result is counterintuitive, as both cases require the same amount of information, i.e., the 3D positions of $n \times m$ Gaussian functions. To investigate this result in more detail, small 'structures' consisting of randomly placed Gaussian functions were used. For each combination of parameters, eight independent structure determination runs were performed, and for each run the achieved resolution was determined. The structure weights w_i where chosen to be uniform and kept fixed during the simulated annealing runs. For a fair comparison, all electron densities were normalized such that, on average, the scattering images contained 15 photons each.

Figure 2.5c and 2.5f show for each combination of parameters the smallest number of images for which all of the replicas achieved a resolution better than a given threshold. As can be seen in Fig. 2.5c, for the structural ensemble of n conformations with m degrees of freedom each, the required number of images is approximately proportional to n^2, the square of the number of conformations. This finding is in line with a theoretical argument showing that the information content of a single image is in this case proportional to $1/n^2$ (Supplementary Note 2.5.2). In contrast, the number of images required to resolve a single structure of $n \times m$ degrees of freedom grows even much faster, approaching a power law m^c with an exponent $c \approx 5$ for increasing resolution (Fig. 2.5f). Hence, for given complexity, ensemble refinement seems to be easier than single structure determination.

2.2.4 Application to experimental data

Having assessed our method using synthetic data for which the 'ground truth' is known, we next tested it on real experimental data from the coliphage PR772 virus data set [81]. Because this virus is much larger than the molecules considered above, much more photons per image were recorded, on average about 400 000 photons each. For a fair assessment, and to mimic the more challenging low photon counts expected for single-molecule scattering experiments, we downsampled the original images by a factor of 10^4 using rejection sampling to obtain images with an average of 40 photons per image. This sample application also served to demonstrate how experimental details such as incomplete detector coverage, irregular detector shape [81], polarization effects, and intensity fluctuations can be systematically included within the Bayesian formalism via an appropriate likelihood function (see Supplementary Note 2.5.7). In addition, a normally distributed noise background of 10% with standard deviation of $\sigma = 0.2\,\mathrm{nm}^{-1}$ was assumed to account for fluctuations in the subtracted background.

The virus electron density was described by 13 Gaussian functions, adapted to the resolution set by the experimental data. Using 1510 randomly selected and downsampled images, the structure shown in Fig. 2.6a) was determined. In contrast to Hosseinizadeh et al. [62], no icosahedral symmetry was imposed. Of course, the restriction to a representation using 13 Gaussian function

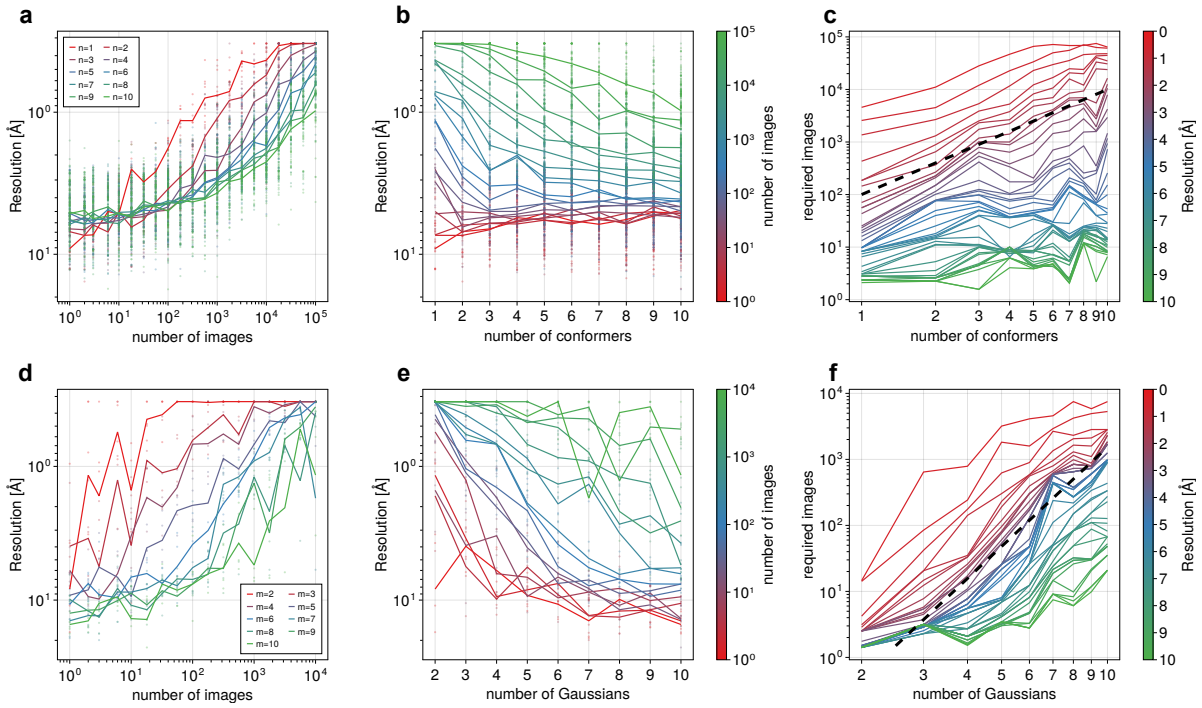

Figure 2.5: Dependence of the resolution on the number of images, the number of conformations, and the complexity of the structure. **a** Resolution as a function of the number of images for various numbers of conformations n. **b** Resolution as a function of the number of conformations for various numbers of images. **c** Required number of images to achieve various resolutions as a function of the number of conformations. For comparison, a quadratic relationship is shown (dashed line). **d** Resolution as a function of the number of images for various structure complexities (parameterized by the number of Gaussians m). **e** Resolution as a function of the number of Gaussians for various numbers of images. **f** Required number of images to achieve various resolutions as a function of the number of Gaussians. For comparison, a power law m^5 is shown (dashed line).

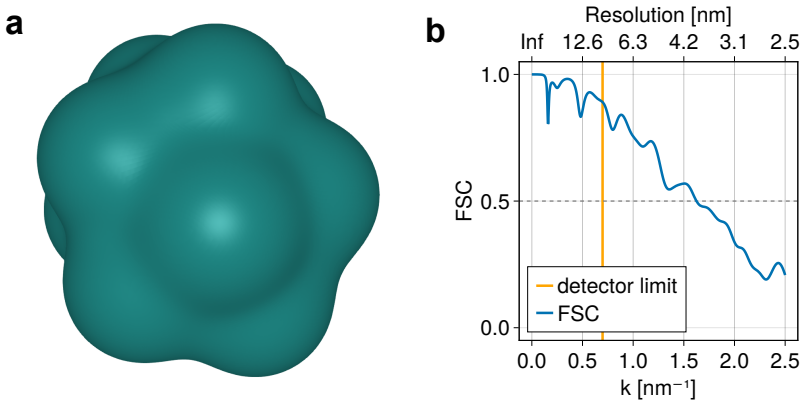

Figure 2.6: Electron density determination of coliphage PR772. **a** Electron density determined from 1510 downsampled scattering images with an average of 40 photons each. **b** Fourier shell correlation (blue) between to independently computed densities and detector limit (orange).

also implies some regularization. To ensure that the obtained icosahedral structure is not an artifact of this particular electron density representation, we determined structures using between 10 and 15 Gaussian functions, and found that 13 yielded the largest posterior probability.

An independent structure calculated from a second, different set of 1510 randomly selected and downsampled images served to calculate the Fourier shell correlation between the two structures (Fig. 2.6b). As can be seen, the resolution is still largely limited by the detector limit of 9 nm, as also reported by Hosseinizadeh et al. [62], and despite the fact that only a very small fraction (10^{-4}) of the originally recorded photons was used.

2.3 Discussion

Here we have developed a rigorous Bayesian method for determining biomolecular structures from single molecule X-ray scattering images in the extreme few photon Poisson regime. Using synthetic scattering images generated from simulated X-ray scattering experiments, we have demonstrated that both single structures as well as structural ensembles of small biomolecules can be resolved to near atomic resolution.

Our results for the globular protein crambin show that a similar resolution of 4.2 Å is obtained compared to previous correlation based methods [36] which also require very few photons per scattering image. Because such correlation based methods disregard higher correlations, whereas the full information content of each image is used in our Bayesian approach, the latter should require fewer images. This was indeed observed for the above protein, for which the number of images required to obtain near atomistic resolution was reduced from roughly $2 \cdot 10^8$ to $1 \cdot 10^8$. Assuming a pulse rate of 27,000 per second [33] and a 10% hit fraction, this would reduce required beam time from 20 to 10 hours.

Because the rather small test proteins studied here scatter very few photons, their structure determination is conceptually more challenging than for larger proteins [50]. To obtain near-atomistic resolutions for larger macromolecules, in contrast, the main bottleneck is computational cost, due to the increased number of unknown degrees of freedom. This issue will need to be addressed by improved optimization or sampling methods, or by utilizing prior structural information, either from structure databases, from AlphaFold [21], or guided by molecular dynamics force fields. Alternatively, as demonstrated for the coliphage PR772 virus data set, lowering the desired resolution reduces the number of required Gaussian functions, thus also enabling electron density determination of larger systems, albeit not at atomistic resolution.

For alanine dipeptide the full conformational ensemble generated by an atomistic simulation was extracted at atomistic resolution of on average 1.8 Å from simulated scattering experiments, in which not only the current orientation of the biomolecule but also its current conformer was unknown. For the 10 amino acid protein chignolin [97] both the folded and unfolded ensembles were resolved, albeit so far at somewhat lower resolution. Clearly, the latter two test systems pushed the limits towards very small molecules, and thus somewhat higher fluxes had to be assumed than are currently achievable. Nevertheless, these examples demonstrated that not only

a larger number of conformers were successfully reconstructed, including disordered (unfolded) states, but also the weights corresponding to the folded and unfolded conformers were accurately recovered. Further, as demonstrated by vanishing weights for incorrect structure poses, using weighted ensembles allows the number of conformers to be determined dynamically.

Unexpectedly few images were required to resolve structural ensembles. Because an ensemble of n conformers consisting of m degrees of freedom each has the same total number of degrees of freedom as a single structure of $n \times m$ degrees of freedom, a similar number of images should be required. Closer analysis suggests that roughly $O(m^5)$ images are required to a resolve a single structure with m degrees of freedom. One might therefore expect that $O(n^5 m^5)$ images are required for an ensemble of n such structures. However, our test refinements required only $O(n^2 m^5)$ images, consistent with an expected information content of $O(1/n^2)$ for a single scattering image.

This result suggests that in terms of the required number of scattering images, even determining more complex conformational ensembles should be possible with current experimental technology. This finding is consistent with previous analyses of larger specimen [124]. Assuming a linear scaling of the scattered number of photons with the sample volume, the required number of images should not depend on the particle size [124]. This translates into our finding, having assumed a constant photon count independent of the number of degrees of freedom. However, our analysis focused on the low photon count regime in which orientation determination for each image is not possible.

Our Bayesian analysis of structural heterogeneity is similar in spirit to approaches that were successfully applied in cryo-electron microscopy [19, 86–92], which, from a mathematical standpoint, shares some similarities with single molecule X-ray scattering, albeit at a much lower noise level. From a more general perspective, our Bayesian approach represents a systematic and rigorous approach to include shot noise in the extreme Poisson regime characteristic for single molecule X-ray scattering experiments. In contrast to other proposed methods, this Bayesian framework will also allow to include other sources of noise and uncertainty in a conceptually straightforward manner, such as incoherently scattered photons, background scattering, detector noise, or scattering by disordered water at the biomolecular surface.

Our analysis of the experimental coliphage PR772 data set demonstrated that and how our Bayesian approach allows for a straightforward and rigorous consideration of, for example, intensity fluctuations, polarization, and irregular detector shapes. More work will be required to also assess the performance regarding incoherent and background scattering, detector noise at high noise levels, and the effect of solvation shells. Here inclusion of calibrated forward noise models (for example, as documented by Yoon et al. [118]) will be a particularly important next step towards atomistically resolved multiple structure ensembles. If successful, and in contrast to diffraction experiments on nano-crystals, this single molecule approach might, with increasing number of resolved conformers, ultimately provide a route to the construction of time resolved structures — molecular movies — without the need for synchronization through optical laser pulses.

2.4 Methods

2.4.1 Structure and structure ensemble representation

Electron density functions of the reference structures or conformers were described by a sum of m of Gaussian functions with atomic positions $\mathbf{y}_i$, heights h_i and standard deviations σ_i,

$$\rho(\mathbf{r}) = \sum_{i=1}^{m} \frac{h_i}{\left(\sigma_i\sqrt{2\pi}\right)^3} \exp\left(-\frac{1}{2\sigma_i^2}\|\mathbf{r}-\mathbf{y}_i\|^2\right) . \tag{2.2}$$

Electron density functions of the determined structures were described similarly, with one common height $h = h_i$ and one common standard deviation $\sigma = \sigma_i$, which is treated as an unknown and determined together with the positions $\mathbf{y}_i$. Structural ensembles were represented by a weighted sum of conformers $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_n\}$ with weights $\mathbf{w} = \{w_1, \ldots, w_n\}$.

2.4.2 Synthetic data generation

For each of the synthetic scattering images, the photon positions on the detector D were drawn from a probability distribution proportional to the intensity function $I(\mathbf{k}) = |F\{\rho\}(\mathbf{k})|^2$ restricted to the appropriate Ewald sphere. Specifically, generation of each image involved the following steps:

1. A conformation of the molecule is selected randomly from the reference ensemble (for example, consisting of molecular dynamics trajectories),
2. a random orientation $\mathbf{R}$ of the molecule is drawn uniformly from the rotation group SO(3),
3. the number of scattered photons is drawn from a Poisson distribution with mean $N\int_D I(\mathbf{Rk})\,\mathrm{d}\mathbf{k}$, where N is the incoming beam intensity,
4. the position of each scattered photon is drawn from the probability distribution proportional to $(I \circ \mathbf{R})|_D$.

The last two steps were implemented using rejection sampling. To this end, a von Mises-Fisher distribution p on D was chosen with high enough standard deviation that $I(\mathbf{k}) \leq p(\mathbf{k})$ everywhere. Then, for each photon, its position $\mathbf{k}$ was drawn from p and it was accepted with probability $I(R\mathbf{k})/p(R\mathbf{k})$. The beam intensity N was chosen together with a normalization of ρ such that this procedure accurately produces a Poisson distribution of the desired expected number of photons per image.

2.4.3 Computation of likelihoods

The probability density of observing an image defined by photon positions $\mathbf{k}_1, \ldots, \mathbf{k}_l$ given an electron density function ρ with is corresponding intensity function $I(\mathbf{k})$ was computed by

averaging over all possible orientations $\mathbf{R} \in \mathrm{SO}(3)$ of the molecule,

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) = \frac{N^l}{l!} \int_{\mathrm{SO}(3)} \exp\left(-N \int_D I(\mathbf{R}\mathbf{k})\, \mathrm{d}\mathbf{k}\right) \left(\prod_{i=1}^{l} I(\mathbf{R}\mathbf{k}_i)\right) \mathrm{d}\mathbf{R}, \tag{2.3}$$

where for each orientation, the probability is a product of the Poisson distribution for the number of photons l in the image and a factor depending on the photon positions. These integrals were approximated by averaging over a discrete set of typically $r \approx 10^3$ to $r \approx 10^5$ rotations $\mathbf{R}_i$ with weights s_i,

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) \approx \frac{N^l}{l!} \sum_{i=1}^{r} s_i \exp\left(-N \int_D I(\mathbf{R}_i\mathbf{k})\, \mathrm{d}\mathbf{k}\right) \prod_{j=1}^{l} I(\mathbf{R}_i\mathbf{k}_j). \tag{2.4}$$

The rotations $\mathbf{R}_i$ and their weights s_i are constructed by combining a Lebedev quadrature rule on S^2 with a uniform quadrature rule on S^1 via the Hopf map [125, 126] (Supplementary Note 2.5.3).

2.4.4 Simulated annealing and hierarchical sampling

A Monte Carlo simulated annealing approach with the energy function $-\log P$ was used to sample from or maximize the Bayesian posterior probability, as described in detail in the Supplement. To enhance convergence, Bayesian sampling and maximization were performed in multiple hierarchical resolution stages. Starting from a low resolution representation of ρ with correspondingly few degrees of freedom, the number of Gaussian functions was doubled in each stage and the reduced resolution electron density determined by the previous stage was used as a proposal density (see Supplement). To calculate likelihoods for the reduced resolution structures, lower resolution scattering images were generated from the original images by rejection sampling, that is, by removing each photon in the original images with probability $1 - \exp(-\sigma^2 k^2/2)$. By construction, this rejection scheme samples from a Fourier transformed density $I_\rho \cdot \exp(-\sigma^2|\mathbf{k}|^2/2)$ which, by the convolution theorem, corresponds to a smoothed real space density $\tilde{\rho} = \rho * \mathcal{N}(\sigma)$ obtained as the convolution of ρ with a Gaussian kernel with width (resolution) σ. Computational efficiency was further increased substantially by selecting only those original images for the likelihood computations that actually contain useful information at the respective resolution. As described in the Supplement, the Bayesian formalism allows for removing this selection bias.

2.4.5 Structure alignment and resolution estimate

Because the orientations of the obtained structures are irrelevant, these were rotationally aligned to each other by minimizing the cost function

$$d(\mathbf{S}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1}^{m} \|\mathbf{y}_i - \mathbf{S}\mathbf{y}'_j\| + \frac{1}{m} \sum_{j=1}^{m} \min_{i=1}^{n} \|\mathbf{y}_i - \mathbf{S}\mathbf{y}'_j\|. \tag{2.5}$$

Here, the positions $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\mathbf{y}'_1, \ldots, \mathbf{y}'_m$ define two structures per equation (2.2) and $\mathbf{S}$ is a rotation matrix $\mathbf{S} \in \mathrm{O}(3)$. Both rotations and reflections were included, as X-ray scattering images do not distinguish between mirror images.

The resolution of the aligned structures was estimated using Fourier shell correlations [120],

$$\mathrm{FSC}(k) = \frac{\int_{\|\mathbf{k}\|=k} \hat{\rho}_1(\mathbf{k})^* \hat{\rho}_2(\mathbf{k}) \, \mathrm{d}\mathbf{k}}{\sqrt{\int_{\|\mathbf{k}\|=k} |\hat{\rho}_1(\mathbf{k})|^2 \, \mathrm{d}\mathbf{k}} \sqrt{\int_{\|\mathbf{k}\|=k} |\hat{\rho}_2(\mathbf{k})|^2 \, \mathrm{d}\mathbf{k}}}, \tag{2.6}$$

where ρ_1 are ρ_2 the structures to be compared and $\hat{\rho}$ denotes the Fourier transform of ρ. Accordingly, the achieved resolution was determined as $2\pi/k_{\mathrm{fsc}}$, where k_{fsc} is the threshold at which the Fourier shell correlation drops below 1/2, providing a conservative estimate [120].

2.4.6 Molecular dynamics simulations

All atomistic simulation trajectories were generated using the GROMACS 2018 software package [127] with the Charmm36mm force field [128] and the OPC water model [129]. For chignolin, the starting structure was taken from the Protein Data Bank [130], entry 5AWL [97]. All hydrogen atoms were described by virtual sites [131]. Each protein was placed within a triclinic water box, such that the smallest distance between protein surface and box boundary was larger than 1.5 nm. Sodium and chloride ions were added to neutralize the system, corresponding to a physiological concentration of 150 mmol/l. Energy minimization was performed using steepest descent for $5 \cdot 10^4$ steps. Each system was subsequently equilibrated for 0.5 ns in the NVT ensemble, and subsequently for 1.0 ns in the NPT ensemble at 1 atm pressure and temperature 300 K using an integration time step of 2 fs. The velocity rescaling thermostat [132] and Parrinello-Rahman pressure coupling [133] were used with coupling coefficients of $\tau = 0.1$ ps and $\tau = 1$ ps, respectively. All bond lengths of the solute were constrained using the LINCS algorithm [134] with an expansion order of 6, and the geometry of the water molecules was constrained using the SETTLE algorithm [135]. Electrostatic interactions were calculated using PME [136], with a real space cutoff of 10 Å and a Fourier spacing of 1.2 Å. For all production runs, a 4 fs integration was used, and the atom coordinates were saved every 100 ps, such that 10^5 snapshots were available for each trajectory. The trajectories for alanine dipeptide were taken from mdshare [137]. The structure for crambin was taken from PDB entry 1EJG [96].

2.5 Supplementary Notes

2.5.1 Parameters

The parameters used for the test cases are shown in Table 2.1. The Lebedev precision and the number of angular rotations S_j are chosen such that the expected angular distance between nearest neighbors in the resulting grid is smaller than the length scale corresponding to the desired relative resolution divided by the approximate radius of the molecule. The parameters

for image selection (r_i and m_i) were chosen such that the radial distribution of photons in the selected images was close to uniform up to the desired resolution level.

Name	Stage	total images	selected images	n_i	r_i [1/Å]	σ [Å]	$t_{1/2}$	m	n	Lebedev precision	angular rotations
Crambin	1	$8.96 \cdot 10^3$	1,000	(4)	$(0.25, \infty)$	2.0	$1 \cdot 10^3$	12	1	23	32
	2	$1.00 \cdot 10^7$	19,315	(3, 2)	$(0.33, 0.5, \infty)$	1.5	$1 \cdot 10^4$	23	1	47	32
	3	$3.04 \cdot 10^6$	50,000	(1, 1, 2)	$(0.35, 0.5, 0.65, \infty)$	1.2	$2 \cdot 10^4$	46	1	47	64
	4	$1.00 \cdot 10^8$	204,447	(1, 2, 2)	$(0.35, 0.5, 0.8, \infty)$	0.9	$1 \cdot 10^5$	92	1	89	64
	5	$1.00 \cdot 10^8$	634,032	(1, 1, 3)	$(0.4, 0.65, 0.9, \infty)$	0.5	$1 \cdot 10^5$	184	1	89	64
Dipeptide	1	$1.00 \cdot 10^6$	3,965	(4, 4)	$(0.9, 1.3, \infty)$	0.5	$1 \cdot 10^3$	10	2	23	32
	2	$1.00 \cdot 10^6$	-	-	-	0.0	$5 \cdot 10^3$	10	8	35	32
Chignolin	1	$1.00 \cdot 10^4$	-	-	-	2.5	$1 \cdot 10^3$	5	2	23	32
	2	$1.09 \cdot 10^7$	100,000	(2, 2)	$(0.4, 0.6, \infty)$	1.5	$5 \cdot 10^3$	10	4	23	32
	3	$1.24 \cdot 10^7$	100,000	(2, 3)	$(0.4, 0.6, \infty)$	1.2	$1 \cdot 10^4$	20	6	47	64
Coliphage	1	$1.51 \cdot 10^3$	-	-	-	70.0	$1 \cdot 10^4$	13	1	35	50

Table 2.1: Parameters for the test cases.

2.5.2 Expected information content of scattering images

The information content of scattering image on structural ensembles can be estimated analogous to an argument for mixtures of normal distributions [138], as follows. Consider an ensemble of two structures ρ_1 and ρ_2 with weights w and $1 - w$, respectively. The probability of observing an image x is then a mixture of the two single distributions,

$$p(x; \rho_1, \rho_2) = wp(x; \rho_1) + (1 - w)p(x; \rho_2). \tag{2.7}$$

By the Bayesian central limit theorem, in the limit of many scattering images the posterior becomes a multivariate normal distribution with covariance $N^{-1}I^{-1}$,

$$P(\rho_1, \rho_2 \,|\, \mathcal{I}) \approx \mathcal{N}(\rho_1, \rho_2; N^{-1}I^{-1}), \tag{2.8}$$

where N is the number of images, and I the Fisher information matrix. The first diagonal element of this matrix is approximately proportional to the weight squared,

$$I_{\rho_1\rho_1} = \mathrm{E}\left[\left(\frac{\partial}{\partial\rho_1}\log p(x; \rho_1, \rho_2)\right)^2\right] = \mathrm{E}\left[\left(\frac{w\frac{\partial}{\partial\rho_1}p(x; \rho_1)}{p(x; \rho_1, \rho_2)}\right)^2\right] = w^2\ \mathrm{E}\left[\left(\frac{\frac{\partial}{\partial\rho_1}p(x; \rho_1)}{p(x; \rho_1, \rho_2)}\right)^2\right]. \tag{2.9}$$

Therefore, under the assumption that the off-diagonal elements are small, the limiting variance for ρ_1 becomes $1/(Nw^2)$. An similar argument can be carried out for more than two distinct structures. In the special case of uniform weights $w = 1/n$ the limiting variance becomes n^2/N, consistent with the quadratic scaling observed in Fig. 5.

2.5.3 Computation

The integral over SO(3) is approximated by a finite sum over rotations $\mathbf{R}_i$ with weights s_i,

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_n \,|\, \rho) \approx \frac{N^n}{n!} \sum_i s_i \exp\left(-N \int_D I(\mathbf{R}_i \mathbf{k}) \,\mathrm{d}\mathbf{k}\right) \prod_{j=1}^{n} I(\mathbf{R}_i \mathbf{k}_j) \tag{2.10}$$

Computing this sum involves evaluating the intensity function I at all points of the form $\mathbf{R}_i \mathbf{k}_j$. Since this has to be done for all the images, this leads to a very large number of evaluations of I. It is therefore efficient to first discretize the images. To that end, the detector is pixelated, that is, partitioned into a grid of cells with centers $\mathbf{x}_k$ and areas a_k. Each image $\mathbf{k}_1, \ldots, \mathbf{k}_n$ is replaced with a set of indices $k_1, \ldots, k_n$, such that for each $\mathbf{k}_i$ the closest point in the grid is $\mathbf{x}_{k_i}$. In this setting, the probability distribution becomes

$$P(k_1, \ldots, k_n \,|\, \rho) \approx \frac{N^n}{n!} \sum_i s_i \exp\left(-N \sum_k a_k I(\mathbf{R}_i \mathbf{x}_k)\right) \prod_{j=1}^{n} a_{k_j} I(\mathbf{R}_i \mathbf{x}_{k_j}) \tag{2.11}$$

To construct the quadrature rule for SO(3), we proceed as follows. First, we choose a Lebedev grid as a uniform grid of points $\mathbf{v}_i$ in the 2-sphere S^2. For each one of these, we find a rotation $Q_i \in \mathrm{SO}(3)$ such that $Q_i \mathbf{v}_i \parallel \mathbf{k}_0$. In addition, let S_j be uniformly spaced rotations around the axis defined by $\mathbf{k}_0$. The set of products $S_j Q_i$ is then a uniform grid in SO(3). Equation (2.11) becomes

$$P(k_1, \ldots, k_n \,|\, \rho) \approx \frac{N^n}{n!} \sum_{i,j} s_i \exp\left(-N \sum_k a_k I(Q_i S_j \mathbf{x}_k)\right) \prod_{m=1}^{n} a_{k_m} I(Q_i S_j \mathbf{x}_{k_m}) \tag{2.12}$$

Choosing the pixel grid $\mathbf{x}_k$ such that it is rotationally symmetric allows further simplification. We reindex it as $\mathbf{x}_{k,l}$, such that $S_j \mathbf{x}_{k,l} = \mathbf{x}_{k+j,l}$. Here, the first index is considered cyclic, that is, if, say, k ranges from 1 to $k_{\max}$, then $\mathbf{x}_{k+j,l}$ is to be interpreted as $\mathbf{x}_{(k+j \mod k_{\max}),l}$.

The corresponding areas $a_{k,l}$ only depend on l, so we write $a_l = a_{k,l}$. The images now also consist of these new indices. Plugging this in, we get

$$P(k_1, l_1, \ldots, k_n, l_n \,|\, \rho) \approx \frac{N^n}{n!} \sum_{i,j} w_i \exp\left(-N \sum_{k,l} a_l I(Q_i S_j \mathbf{x}_{k,l})\right) \prod_{m=1}^{n} a_{l_m} I(Q_i S_j \mathbf{x}_{k_m,l_m}) \tag{2.13}$$

$$= \frac{N^n}{n!} \sum_i w_i \exp\left(-N \sum_{k,l} a_l I(Q_i \mathbf{x}_{k,l})\right) \sum_j \prod_{m=1}^{n} a_{l_m} I(Q_i \mathbf{x}_{k_m+j,l_m}) \tag{2.14}$$

$$= \frac{N^n}{n!} \sum_i w_i P_i \sum_j \prod_{m=1}^{n} I_{i,k_m+j,l_m} \tag{2.15}$$

The values $I_{i,k,l} \coloneqq a_l I(Q_i \mathbf{x}_{k,l})$ and $P_i \coloneqq \exp(-N \sum_{k,l} I_{i,k,l})$ can be computed in advance and reused for each image.

Due to limited floating point precision, a number of adjustments must be made. Due to the large value of N, computing P_i results in underflow. Therefore, we write

$$\tilde{P}_i = P_i/\bar{P}, \qquad \bar{P} = \left(\prod_{i'=1}^{i_{\max}} P_{i'}\right)^{\frac{1}{i_{\max}}}. \tag{2.16}$$

Further, $I_{i,k,l} \ll 1$, so if the images contain enough photons the product over m will underflow. Since the magnitude of $I_{i,k,l}$ depends mostly on l, we define

$$\tilde{I}_{i,k,l} = I_{i,k,l}/\bar{I}_l, \qquad \bar{I}_l = \frac{1}{i_{\max}k_{\max}} \sum_{i'=1}^{i_{\max}} \sum_{k'=1}^{k_{\max}} I_{i',k',l} \tag{2.17}$$

Both $\bar{P}$ and $\bar{I}_l$ do not depend on the rotation index i and factor out,

$$P(k_1, l_1, \ldots, k_n, l_n \,|\, \rho) \approx \frac{N^n}{n!}\bar{P}\left(\prod_{m=1}^{n} \bar{I}_{l_m}\right) \sum_i w_i \tilde{P}_i \sum_j \prod_{m=1}^{n} \tilde{I}_{i,k_m+j,l_m} \tag{2.18}$$

Taking the logarithm,

$$\log P(k_1, l_1, \ldots, k_n, l_n \,|\, \rho) \approx \log \frac{N^n}{n!} + \log \bar{P} + \sum_{m=1}^{n} \log \bar{I}_{l_m} + \log \sum_i w_i \tilde{P}_i \sum_j \prod_{m=1}^{n} \tilde{I}_{i,k_m+j,l_m}, \tag{2.19}$$

we see that only $\log \bar{P}$ and $\log \bar{I}_l$ appear, which can be computed without overflow.

2.5.4 Monte Carlo Simulated Annealing

Let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$ denote vectors of electron densities and weights, respectively. A Markov chain of structural ensembles $\boldsymbol{\rho}_t$ with weights $\mathbf{w}_t$ was constructed iteratively using a Metropolis-within-Gibbs algorithm. This algorithm works as follows. For each step t, first a Metropolis step for the structures is performed, that is, new candidate structures $\boldsymbol{\rho}'$ are drawn from a proposal distribution $g(\boldsymbol{\rho}'|\boldsymbol{\rho}_t)$, and this candidate is accepted ($\boldsymbol{\rho}_{t+1} = \boldsymbol{\rho}'$) or rejected ($\boldsymbol{\rho}_{t+1} = \boldsymbol{\rho}_t$) with probability

$$1 \wedge \exp\left(\frac{\log P(\boldsymbol{\rho}', \mathbf{w}_t \,|\, \mathcal{I}) - \log P(\boldsymbol{\rho}_t, \mathbf{w}_t \,|\, \mathcal{I}) + \log g(\boldsymbol{\rho}_t|\boldsymbol{\rho}') - \log g(\boldsymbol{\rho}'|\boldsymbol{\rho}_t)}{T(t)}\right), \tag{2.20}$$

adopting the notation $1 \wedge x = \min(1, x)$. The temperature $T(t)$ is determined according to an exponential annealing schedule $T(t) = T_0 \exp(-\lambda t)$ for some constant λ. The proposal density g is an isotropic normal distribution $\mathcal{N}(\boldsymbol{\rho}_t, d)$ around $\boldsymbol{\rho}_t$, that is, to obtain the candidate, the position of each Gaussian in the structure representation is perturbed by a normally distributed amount; or it is given by our hierarchical sampling method as described in the next section. The step size d is determined iteratively such that the acceptance rate is the optimal 23% [139], by increasing or decreasing it after a successful or unsuccessful step, respectively.

Second, a separate Metropolis step for the weights is performed. To correctly sample from the n-simplex of weights w_i such that $w_i \leq 0$ and $\sum_i w_i = 1$, we introduce variables $s_j \leq 0$ such

that $w_i = s_i / \sum_j s_j$. For these variables, the proposals are drawn from a Gamma distribution of mean s_j and standard deviation given by the current step size. Note that this is not a proposal distribution in the sense of equation (2.20), as it does not appear in the acceptance probability. If one of the weights w_i becomes zero during the sampling process, the corresponding structure ρ_i does no longer affect the posterior probability, hindering convergence. To prevent this, a delayed acceptance scheme is used as follows. Each proposal $\mathbf{s}'$ with corresponding weights $\mathbf{w}'$ generated by the above procedure is accepted with probability

$$g^*(\mathbf{w}' \,|\, \mathbf{w}_t) = 1 \wedge \exp\left(\frac{1}{2\nu}\|\mathbf{w}' - \mathbf{c}\|^2 - \frac{1}{2\nu}\|\mathbf{w}_t - \mathbf{c}\|^2\right), \tag{2.21}$$

where $\mathbf{c} = (1/n, \ldots, 1/n)$ and ν is sufficiently small to ensure that the weights remain non-zero. Finally, the proposal is accepted with probability

$$1 \wedge \exp\left(\frac{\log P(\boldsymbol{\rho}_{t+1}, \mathbf{w}' \,|\, \mathcal{I}) - \log P(\boldsymbol{\rho}_{t+1}, \mathbf{w}_t \,|\, \mathcal{I}) + \log g^*(\mathbf{w}_t|\mathbf{w}') - \log g^*(\mathbf{w}'|\mathbf{w}_t)}{T(t)}\right). \tag{2.22}$$

The metropolis step for the weights has little computational cost, as the computationally costly parts of equation (2.5.3) are unaffected. Therefore, it is repeated multiple times in each iteration.

2.5.5 Proposal density for hierarchical sampling

In each hierarchical sampling stage, the number of Gaussian functions was doubled, and the reduced resolution structure determined by the previous stage was used as a proposal density to improve convergence in the simulated annealing, as follows. Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be the positions of the Gaussian functions from the previous stage, and $\mathbf{z}_1, \ldots, \mathbf{z}_{2n}$ those of the current stage. Then the proposal density was, up to normalization, given by

$$g(\mathbf{z}'_1, \ldots, \mathbf{z}'_{2n} \,|\, \mathbf{z}_1, \ldots, \mathbf{z}_{2n}) \propto \prod_{i=1}^{2n} \exp\left(-\frac{\|\mathbf{z}'_i - \mathbf{z}_i\|^2}{2\sigma^2}\right) \prod_{i=1}^{n} \exp\left(-\frac{\|\mathbf{z}'_{2i} - \mathbf{y}_i\|^2 + \|\mathbf{y}'_{2i+1} - \mathbf{y}_i\|^2}{2w^2}\right), \tag{2.23}$$

where w is the width of the Gaussians from the previous stage. For ensembles of structures, the proposal density becomes a product over the single structures ρ_i with separate intermediates for each ρ_i,

$$g(\boldsymbol{\rho}' \,|\, \boldsymbol{\rho}_t) = \prod_{i=1}^{n} g(\rho'_i \,|\, \rho_i), \tag{2.24}$$

where $g(\rho'_i \,|\, \rho_i)$ is the proposal density from equation (2.23).

2.5.6 Image selection

In our hierarchical sampling scheme, images containing only photons with $|\mathbf{k}|$ below a threshold are no longer useful, and the computations were sped up by removing these images. To achieve this, numbers (r_i) and integers m_i were chosen, and only the subset $\mathcal{I}_C$ of those images was used that fulfilled the condition $C(I)$ that for each i the image I contains at least m_i photons with $r_i < |\mathbf{k}| < r_{i+1}$. To ensure that the posterior was not biased by this filtering, it was taken

into account in the Bayesian formalism by dividing by the probability $P(C \,|\, \boldsymbol{\rho}, \mathbf{w})$ that an image fulfills C. In other words, the original posterior probability was replaced with $P(\boldsymbol{\rho}, \mathbf{w} \,|\, \mathcal{I}_C, C) \propto P(\mathcal{I}_C \,|\, \boldsymbol{\rho}, \mathbf{w}) / P(C \,|\, \boldsymbol{\rho}, \mathbf{w})$. The probability that an image fulfills C depends on both the orientation $\mathbf{R}$ and the conformer i. Therefore, $P(C \,|\, \boldsymbol{\rho}, \mathbf{w})$ was obtained by averaging over both,

$$P(C \,|\, \boldsymbol{\rho}, \mathbf{w}) = \sum_j w_j \int_{\mathrm{SO}(3)} \prod_i \left(1 - Q\left(m_i - 1, N \int_{D_i} |F\{\rho_j\}(\mathbf{Rk})|^2 \,\mathrm{d}\mathbf{k}\right)\right) \mathrm{d}\mathbf{R}, \tag{2.25}$$

where $Q(x, \lambda)$ is the cumulative distribution function of a Poisson distribution with mean λ and $D_i = \{\mathbf{k} \in D \,|\, r_i < \|\mathbf{k}\| < r_{i+1}\}$ is the relevant slice of the Ewald sphere.

2.5.7 Intensity fluctuations and noise

The Bayesian framework allows for a straightforward inclusion of effects like intensity fluctuations, noise, polarization, and the irregular detector shape. For example, to include a simple model of background scattering, the 3D intensity function I gets an additional normal distribution,

$$I(\mathbf{k}) \propto |\mathcal{F}\{\rho\}(\mathbf{k})|^2 + C \exp\left(\frac{\mathbf{k}^2}{2\sigma^2}\right). \tag{2.26}$$

where σ sets the standard deviation of the background scattering and C is a normalization constant setting the amount of noisy photons. As this only changes the 3D intensity function, the other parts of our method are unaffected. Other noise distributions can be included similarly. To account for the irregular detector shape, we introduce an additional factor $f(\mathbf{k})$, into equation (3) from the main text, describing how likely a photon with scattering vector $\mathbf{k}$ is to be detected,

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) = \frac{N^l}{l!} \int_{\mathrm{SO}(3)} \exp\left(-N \int_D I(\mathbf{Rk}) f(\mathbf{k}) \,\mathrm{d}\mathbf{k}\right) \left(\prod_{i=1}^{l} I(\mathbf{Rk}_i) f(\mathbf{k}_i)\right) \mathrm{d}\mathbf{R}. \tag{2.27}$$

Nothing that $f(\mathbf{k}_i)$ is constant (it neither depends on the structure ρ nor the orientation $\mathbf{R}$), this becomes

$$\begin{aligned} P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) &= \left(\prod_{i=1}^{l} f(\mathbf{k}_i)\right) \frac{N^l}{l!} \int_{\mathrm{SO}(3)} \exp\left(-N \int_D I(\mathbf{Rk}) f(\mathbf{k}) \,\mathrm{d}\mathbf{k}\right) \left(\prod_{i=1}^{l} I(\mathbf{Rk}_i)\right) \mathrm{d}\mathbf{R} \\ &\propto \frac{N^l}{l!} \int_{\mathrm{SO}(3)} \exp\left(-N \int_D I(\mathbf{Rk}) f(\mathbf{k}) \,\mathrm{d}\mathbf{k}\right) \left(\prod_{i=1}^{l} I(\mathbf{Rk}_i)\right) \mathrm{d}\mathbf{R}. \end{aligned} \tag{2.28}$$

This was implemented into equation (5) by modifying the weights a_k. The polarization of the XFEL-beam is handled in the same way, with f encoding the additional factor due to the polarization. For the analysis of the coliphage data set, we fitted a very simple polarization model $f(\mathbf{k}) = 1 - c \cdot k_y^2$ to the data.

To account for intensity fluctuations, one could, in principle, integrate equation (3) from the main text over the incoming intensity N. However, as the correct distribution is unknown, we

instead opted for a 'normalized' version independent of the intensity,

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) \propto \frac{1}{l!} \int_{\mathrm{SO}(3)} \left(\int_D I(\mathbf{R}\mathbf{k}) f(\mathbf{k}) \, \mathrm{d}\mathbf{k} \right)^{-l} \left(\prod_{i=1}^{l} I(\mathbf{R}\mathbf{k}_i) \right) \mathrm{d}\mathbf{R}. \tag{2.29}$$

This corresponds of observing the image $\mathbf{k}_1, \ldots, \mathbf{k}_l$ given an already known number of photons l.

2.5.8 Optimal transport resolutions

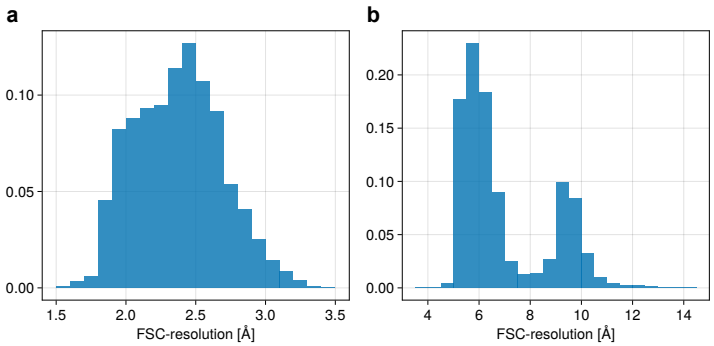

Figure 2.7: Cost distribution of optimal transport plan between reconstructed and reference ensembles for alanine dipeptide (**a**) and chignolin (**b**)

Chapter 3

Scaling behavior and noise tolerance

The following text is in preparation for submission as

S. Schultze and H. Grubmüller: **Noise tolerance and scaling behavior of Bayesian electron density determination from single-molecule X-ray scattering**.

I carried out the research and wrote the manuscript. Helmut Grubmüller supervised the research and revised the manuscript.

Abstract

Single molecule X-ray scattering experiments using free electron lasers hold the potential to resolve both single structures and structural ensembles of biomolecules. However, structure refinement is exceedingly challenging due to complications such as low photon counts, high noise levels and low hit rates. Furthermore, for each scattering image the molecular orientation is random and unknown.

Here we developed and assessed a Bayesian framework with a quite realistic forward model, accounting for experimental effects such as intensity fluctuations, hits vs. misses, beam polarization, irregular detector shapes, incoherent scattering and background scattering. Importantly, our approach does not rely on hit selection or orientation determination. We demonstrate that it should be possible to determine electron densities of small proteins in this extreme low hit rate and high noise Poisson regime. We show that in this scenario it is indeed not possible to determine hit vs. miss or the orientation for each image. Further, we found that the structural information per scattering image scales with the square of the number of coherently scattered photons, and that already a small amount of noise strongly decreases the achieved resolutions. Extrapolating our scaling analyses, we estimate that 10^{12} to 10^{14} scattering images should be required to resolve the protein Crambin at atomistic resolution, but already slightly reduced resolutions should require substantially fewer images

3.1 Introduction

Single-molecule X-ray scattering experiments using ultrashort X-ray free electron laser (XFEL) pulses offer a new route for the structure determination of biomolecules [27, 60, 116, 117]. They also hold the potential to extract the entire structural ensemble of a molecule [115] without the need for synchronization, presenting a promising alternative to the established approaches using nano-crystals [28, 32, 39–44].

In such single-molecule scattering experiments, single copies of the molecule enter a high intensity femtosecond pulsed XFEL beam, and for each pulse, a scattering image is recorded (as shown in Figure 3.1). The ultrashort pulse duration is used to outrun the Coulomb explosion due to the extensive radiation damage. Importantly, for small specimen like single molecules, this image does not consist of a continuous intensity distribution, but only the positions of a few scattered photons on the detector. In addition to photons from coherent scattering, each image also contains incoherently scattered photons arising from Compton scattering and interaction via the photo effect and subsequent Auger decay. Furthermore, photons also scatter on stray gas particles in the beam line is, which originate mostly from carrier gas and evaporated solvent.

Only the photons from coherent scattering on the sample molecule carry structural information, the remaining photons are noise. The relative intensities of these scattering effects depend strongly on the molecule size and the photon energy [140]. It is, however, estimated that up to 90% of the scattered photons are noise, and only 10% signal photons [118]. We here assume that coherent and incoherent photons are indistinguishable, because their small energy difference can typically not be detected [118].

Several proof-of-principle experiments [45–47] have demonstrated the feasibility of the approach. However, successful density determination has so far been limited to much larger specimen, for instance of entire mimiviruses [45, 46] and coliphage viruses [47]. For such large specimen the number of coherent photons per image is much higher than for single molecules — specifically, it is 10^7 for the mimivirus [45, 46] as opposed to 10-100 expected photons per image for average sized proteins [118, 119]. In combination with the high noise level, this lower estimated photon count poses substantial challenges for structure refinement, especially due to two further complications.

First, the molecular orientation at the time of scattering is typically unknown. Although there have been attempts to determine the orientation from the images [48–55, 65], they all require several hundreds to thousands of photons per image — and even more for higher noise levels. In fact, as we will demonstrate, for lower photon counts the orientations can generally not be determined from the images alone, even with knowledge of the true molecular structure. As an alternative, manifold based approaches (including, for example, diffusion map) have been developed, which aim to determine the manifold of orientations [66–70]. They are, however, limited by similarly high photon count requirements.

Second, most proteins also show structural heterogeneity and undergo conformational transitions [1], such that also the current conformational state is unknown. Delivering a single molecule into the beam line is successful for only a small fraction of typically less than 1% of the pulses, such that 99% of the images are ‘empty’ but nevertheless contain noise photons [34]. Because,

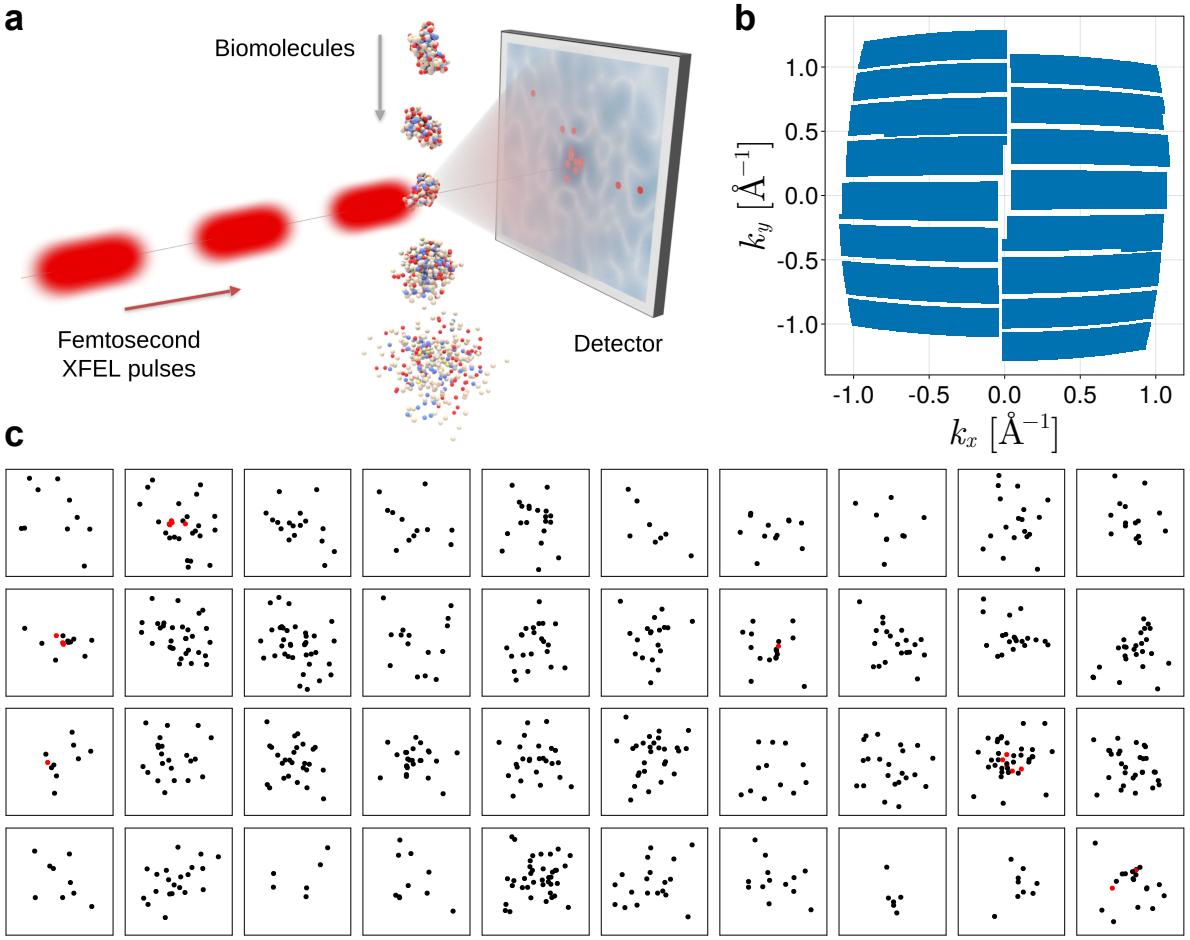

Figure 3.1: **a** Single molecule scattering experiment (image reproduced from von Ardenne et. al. [36]). **b** Irregular detector shape used for our simulated scattering experiments, modelled after the AGIPD [141]. **c** Example simulated scattering images, showing both hits and misses. Only the photons from coherent scattering on the sample molecule (red) carry structural information, all others (black) do not. All axes show the range $q \leq 1.4\,\text{Å}^{-1}$. These images were generated as described in Section 3.4.1.

further, the intensity of the beam at the position of the scattering molecule fluctuates, separation of the actual scattering images ('hits') from the empty ones is a non-trivial task. In addition to the unknown conformational state is therefore also unknown if each image is a hit or a miss. Similar to the molecule orientations, most established approaches rely on such a classification into multiple conformers and hits and misses. Some exemplary scattering images are shown in Figure 3.1c.

Because most conventional refinement methods rely on both accurate hit selection as well as orientation determination, correlation based approaches have been developed as an alternative [36, 71, 73–75, 77], which, notably, require only 3 photons per image [36]. However, as is also the case for the other approaches discussed so far, they use only part of the available information and thus require the collection of large amounts of data and averaging to address shot noise, detector noise and incoherent scattering. Further, we are not aware of any single molecule scattering electron density determination method that allows for systematic inclusion of all types of occurring noise.

In contrast, a Bayesian approach would in fact offer such rigorous inclusion of noise and uncertainties in the form of a forward error model. Also, and equally importantly, such an approach holds the promise to circumvent the need of orientation determination, conformation sorting, or maybe even hit selection. Rather, the electron density that maximizes the Bayesian posterior probability given the whole set of typically millions of images is determined directly. For this, the only requirement is the forward model, that is, the full mathematical description of the distribution of the scattering images including all effects of noise and experimental uncertainties. Importantly, the Bayesian posterior inherently contains all structural information available from the images and thus no information of the experimental data is discarded, such that a minimum number of images should be required to achieve a given resolution level.

In our previous study [115] we have developed such a Bayesian method for electron density determination and have demonstrated that using such a Bayesian method allows for the determination of electron densities at high resolutions, albeit so far only from synthetic noise-free scattering images. Further, we have demonstrated that not only single densities but entire conformational ensembles can be determined from single-molecule scattering images, importantly requiring much fewer images than expected.

Here we present an implementation and assessment our Bayesian framework for much more realistic experimental data, accounting for intensity fluctuations, hits and misses, polarization, irregular detector shapes, incoherent scattering and background scattering. Importantly, our forward model is realistic in that it considers observed signal-to-noise ratios [34, 142] and photon counts [36, 119]. Our approach should therefore provide realistic estimates of the required number of images to achieve a given resolution level.

Considering the limited operational capacity of current XFEL facilities, such as the European XFEL or the LCLS, such realistic estimates are crucial for future experiments. In addition to such an estimate, we used our approach to systematically analyze the scaling behavior of the required number of images in dependence on parameters such as the expected photon count and the signal-to-noise ratio. Unexpectedly, we find that the amount of structural information per scattering image is proportional to the square of the number of scattered photons.

3.2 Theory

Here we first recall the basic scattering theory from which we will derive the noise-free forward model for the Bayesian structure determination [115]. Subsequently, we will expand and refine the forward model by including the most relevant noise sources within this framework. In particular we will consider (1) incoherently scattered photons, (2) background scattering on gas molecules, (3) beam polarization, (4) the irregular shape of the detector, (5) intensity fluctuations, and (6) hits vs. misses. For each step we will first describe the noise model and then implement it within the Bayesian likelihood function.

3.2.1 Basic theory and noise-free forward model

In the experiments, single copies of the sample molecule enter a pulsed femtosecond XFEL-beam, and for each pulse, the positions of the scattered photons are recorded on the detector as a scattering image. Only the coherently scattered photon contain structural information. Each location on the detector corresponds to a specific scattering vector $\mathbf{k} = \mathbf{k}_\mathrm{i} - \mathbf{k}_\mathrm{s}$ on the Ewald sphere E in Fourier space. Here, $\mathbf{k}_\mathrm{i}$ is the incident wave vector and $\mathbf{k}_\mathrm{s}$ the wave vector after scattering. Each scattering image is, therefore, given by a list of scattering vectors $\mathbf{k}_1, \dots, \mathbf{k}_l$. Their probability distribution is given by 3D-intensity function $I_\rho(\mathbf{Rk}) = |\mathcal{F}\{\rho\}(\mathbf{Rk})|^2$, which in turn is given by the Fourier transform of the electron density ρ. Here, $\mathbf{R} \in \mathrm{SO}(3)$ is a rotation matrix describing the orientation of the molecule.

The likelihood that an image $\mathbf{k}_1, \dots, \mathbf{k}_l$ is observed for a given electron density ρ is obtained by averaging the conditional likelihood over all possible orientations $\mathbf{R}$. This conditional likelihood is given by the product of a Poisson distribution for the number of photons l and, because the photons are conditionally independent given $\mathbf{R}$, a product of the intensity function evaluated at the scattering vectors of the scattered photons,

$$P(\mathbf{k}_1, \dots, \mathbf{k}_l \,|\, \rho) \propto \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R}\; I_0^l \exp\left(-I_0 \int_\mathrm{E} I(\mathbf{Rk})\,\mathrm{d}\mathbf{k}\right) \prod_{i=1}^{l} I(\mathbf{Rk}_i). \tag{3.1}$$

Because each scattering image is an independent event, the likelihood that a whole set of images $\mathcal{I}$ is observed for a given electron density ρ is the product of the likelihood of each single image,

$$P(\mathcal{I} \,|\, \boldsymbol{\rho}) = \prod_{(\mathbf{k}_1,\dots,\mathbf{k}_l)\in\mathcal{I}} P(\mathbf{k}_1, \dots, \mathbf{k}_l \,|\, \rho). \tag{3.2}$$

Note that here and subsequently we omit all normalization factors and constants such as the electron radius; instead the normalization is chosen at the end such that the correct photon counts are obtained.

This likelihood, given by equation (3.2), represents the complete noise-free forward model, which forms the basis for the subsequent inclusion of error models [115].

3.2.2 Incoherent and background scattering

In addition to the coherent photons, also incoherently scattered photon from, for example, Compton scattering and Auger decay, are observed. They make up up to 90% of the total scattered photons, but are distributed uniformly on the Ewald sphere. They therefore spread over a much larger solid angle than the coherent photons, such that the effective amount of noise due to this incoherent scattering is smaller. For this reason, the noise due to incoherent scattering is larger for increasing resolution, whereas at the lower resolutions of about 10 nm that have been demonstrated for viruses it can be neglected.

A second source of noise is scattering from other molecules, such as water molecules attached to the sample in aerosol delivery [143], bulk water for liquid beam [144] or sheet [145] delivery,

or remaining gas molecules in the beam volume. These molecules scatter both coherently and incoherently, but, due to the random positions and orientations of these particles, incoherent summation to I_ρ is a good approximation.

Neglecting beam polarization for a moment, the distribution of the photons from incoherent and background scattering is radially symmetric. For simplicity, we here assume a uniform distribution on the Ewald sphere for the incoherently scattered photons and a Gaussian distribution centered at the origin of reciprocal space for the background scattering. Other radial distributions, for example from measurements, can of course be readily implemented

To include incoherent and background scattering within the likelihood function, their distributions are added to the intensity function, replacing I_ρ by

$$I_\mathrm{n}(\mathbf{k}, \rho) = I_\rho(\mathbf{k}) + I_\mathrm{b}(\mathbf{k}) + u_\mathrm{s} + u_\mathrm{b}. \tag{3.3}$$

in equation (3.1). Here, the constants u_s and u_b describe the uniform incoherent scattering on sample molecule and background gas, respectively, and

$$I_\mathrm{b}(\mathbf{k}) = \frac{c_\mathrm{b}}{2\pi\sigma^2} \exp\left(-\frac{\mathbf{k}^2}{2\sigma^2}\right) \tag{3.4}$$

is the Gaussian distribution of the 'coherent' background scattering.

3.2.3 Polarization

To additionally include the linear polarization of the XFEL beam, the scattering intensity needs to be changes by the factor $f_\mathrm{p}(\mathbf{k}) = \cos^2\theta + \cos^2\phi \sin^2\theta = 1 - k_y^2\lambda/2\pi$, where θ is the scattering angle and ϕ the azimuthal angle relative to the direction of polarization [146]. As a result for each scattering vector k the expected number of photons from coherent and Compton scattering is reduced by $f_\mathrm{p}(\mathbf{k})$. In contrast, the distribution of photons arising from Auger decay is unaffected. For our forward model, we assume therefore that the Gaussian noise I_b is multiplied by this factor while the uniform noise is not.

Accordingly, I_n is replaced by

$$I_\mathrm{np}(\mathbf{R}, \mathbf{k}, \rho) = f_\mathrm{p}(\mathbf{k})(I_\rho(\mathbf{R}\mathbf{k}) + I_\mathrm{b}(\mathbf{k})) + u_\mathrm{s} + u_\mathrm{b}. \tag{3.5}$$

, which now also depends on the molecular orientation. As result, rotating the molecule around the beam axis does no longer rotate the scattering images, because the polarization orientation is stationary. The likelihood becomes

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) \propto \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R}\; I_0^l \exp\left(-I_0 \int_E \mathrm{d}\mathbf{k}\; I_\mathrm{np}(\mathbf{R}, \mathbf{k}, \rho)\right) \prod_{i=1}^{l} I_\mathrm{np}(\mathbf{R}, \mathbf{k}_i, \rho). \tag{3.6}$$

3.2.4 Irregular detector shape

Most X-ray detectors have irregular shapes. For example, the AGIPD detector [141] used at the European XFEL is composed of 16 separate modules arranged as shown in Figure 3.1b. Note that because the detector is flat, the distortion of the detector in k-space results from the projection onto the Ewald sphere. In our forward model, the shape of the detector is encoded in the detection probability $p_\mathrm{d}(\mathbf{k})$ that a photon with scattering vector $\mathbf{k}$ is registered by the detector. This allows for the inclusion of any detector shape. Also, this formalism can be used include individual detection probabilities per pixel.

The resulting likelihood function is a straightforward extension similar to the above polarization, the only difference being that here all photons are affected. As a consequence, the factors $p_\mathrm{d}(\mathbf{k}_i)$ in the product over i factor out and can be omitted because they do not depend on the images,

$$\begin{aligned} P(\mathbf{k}_1, \dots, \mathbf{k}_l \,|\, \rho) &\propto \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R}\; I_0^l \exp\left(-I_0 \int_E \mathrm{d}\mathbf{k}\; p_\mathrm{d}(\mathbf{k}) I_\mathrm{np}(\mathbf{R}, \mathbf{k}, \rho)\right) \prod_{i=1}^{l} p_\mathrm{d}(\mathbf{k}_i) I_\mathrm{np}(\mathbf{R}, \mathbf{k}_i, \rho) \\ &\propto \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R}\; I_0^l \exp\left(-I_0 \int_E \mathrm{d}\mathbf{k}\; p_\mathrm{d}(\mathbf{k}) I_\mathrm{np}(\mathbf{R}, \mathbf{k}, \rho)\right) \prod_{i=1}^{l} I_\mathrm{np}(\mathbf{R}, \mathbf{k}_i, \rho). \end{aligned} \tag{3.7}$$

3.2.5 Intensity fluctuations

The fluctuations of the incoming beam intensity I_0 are described by a Gamma distribution $I_0 \sim \langle I_0 \rangle \Gamma(\alpha, \beta)$, where $\langle I_0 \rangle$ is the average intensity [147–149]. The shape and rate parameters α and β depend on the specific free electron laser. For our forward model, we use $\alpha = \beta = 4$, which as has been measured for an XFEL operating at 32 nm wavelength [149].

To include these fluctuations within the likelihood function, an additional integral over I_0 weighted by the probability density of the Gamma distribution is required,

$$\begin{aligned} &P(\mathbf{k}_1, \dots, \mathbf{k}_l \,|\, \rho) \\ &\quad\propto \int_0^\infty \mathrm{d}I_0\, I_0^{\alpha-1} \exp\left(-\frac{\beta I_0}{\langle I_0 \rangle}\right) \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R}\; I_0^l \exp\left(-I_0 \int_E \mathrm{d}\mathbf{k}\; p_\mathrm{d}(\mathbf{k}) I_\mathrm{np}(\mathbf{R}, \mathbf{k}, \rho)\right) \prod_{i=1}^{l} I(\mathbf{R}, \mathbf{k}_i, \rho). \end{aligned} \tag{3.8}$$

Conveniently, this integral can be carried out analytically, and the likelihood reads

$$P(\mathbf{k}_1, \dots, \mathbf{k}_l \,|\, \rho) \propto \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R}\; \left(\frac{\beta}{\langle I_0 \rangle} + \int_E \mathrm{d}\mathbf{k}\, p_\mathrm{d}(\mathbf{k}) I_\mathrm{np}(\mathbf{R}, \mathbf{k}, \rho, x)\right)^{-l-\alpha} \prod_{i=1}^{l} I_\mathrm{np}(\mathbf{R}, \mathbf{k}_i, \rho). \tag{3.9}$$

Note that this likelihood does not include fluctuations due to the relative position of the sample molecule within the beam; these will be included in the next section.

3.2.6 Hits and misses

Finally, we take into account that most pulses actually miss the sample molecule and, hence, the resulting scattering image contains only noise. In fact, typical hit rates are below 1 to 10 percent [142, 150]. Unfortunately, in the very low signal to noise regime considered here, most of the scattering images resulting from hits are indistinguishable from misses. Further, due to the non-uniform beam profile, the beam intensity at the molecule position is reduced by an approximately log-uniformly distributed factor $\eta \sim \mathcal{LU}(\eta_{\min}, 1)$. To see this, assume that for each pulse exactly one sample molecule is placed at a position $\mathbf{r}$ uniformly distributed on a disc of radius R. The squared norm of $\mathbf{r}$ is uniformly distributed, as seen by the calculation $p(\mathbf{r}^2{=}c) = p(\|\mathbf{r}\|{=}\sqrt{c}) \left(\frac{\mathrm{d}}{\mathrm{d}r} r^2|_{r=\sqrt{c}}\right)^{-1} \sim \sqrt{c}/\sqrt{c} = 1$. Assuming a Gaussian beam profile with standard deviation σ, the logarithm of the relative intensity at $\mathbf{r}$ is $\log \eta = \log(\exp(-\mathbf{r}^2/(2\sigma^2))) = -\mathbf{r}^2/(2\sigma^2)$, which due to the previous calculation is uniform between 0 and $\eta_{\min} = -R^2/(2\sigma^2)$.

In the intensity function, the scattering intensities corresponding to scattering on the sample are scaled by the relative intensity η wile those corresponding to scattering on the gas background remain unchanged,

$$I_{\mathrm{np}}(\mathbf{R}, \mathbf{k}, \eta, \rho) = f_{\mathrm{p}}(\mathbf{k})(\eta I_\rho(\mathbf{R}\mathbf{k}) + I_{\mathrm{b}}(\mathbf{k})) + u_{\mathrm{s}}\eta + u_{\mathrm{b}}. \tag{3.10}$$

The likelihood function is obtained by integrating over η as a nuisance parameter, weighted by the probability of the log-uniform distribution $p(\eta) \sim 1/\eta$,

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) \propto \int_{\eta_{\min}}^1 \frac{1}{\eta} \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R} \, \left(\frac{\beta}{\langle I_0 \rangle} + \int_E \mathrm{d}\mathbf{k}\, p_{\mathrm{d}}(\mathbf{k}) I_{\mathrm{np}}(\mathbf{R}, \mathbf{k}, \eta, \rho)\right)^{-l-\alpha} \prod_{i=1}^{l} I_{\mathrm{np}}(\mathbf{R}, \mathbf{k}_i, \eta, \rho). \tag{3.11}$$

The likelihood function in equation (3.11) represents the so far complete forward model, including incoherently scattered photons, background scattering, beam polarization, the irregular detector shape, intensity fluctuations, and hits vs. misses. It will be used subsequently to determine electron densities from synthetic scattering images.

3.3 Methods

3.3.1 Structure representation

Electron density functions of the reference structures were described by a sum of one Gaussian bead per atom with position $\mathbf{y}_i$, heights h_i and standard deviations σ_i assigned depending on the atom type,

$$\rho(\mathbf{r}) = \sum_{i=1}^{m} \frac{h_i}{\left(\sigma_i\sqrt{2\pi}\right)^3} \exp\left(-\frac{1}{2\sigma_i^2}\|\mathbf{r} - \mathbf{y}_i\|^2\right). \tag{3.12}$$

Electron density functions of the determined electron densities were described similarly, with one common height $h = h_i$ and one common standard deviation $\sigma = \sigma_i$ for all Gaussian function in the above sum, which is treated as an unknown and determined together with the positions $\mathbf{y}_i$.

3.3.2 Simulated scattering experiments

Synthetic scattering images were calculated using the forward model described in Section 3.2.1 as follows.

1. Draw the pulse intensity $I_0 \sim \langle I_0 \rangle \Gamma(\alpha, \beta)$, the relative intensity $\eta \sim \mathcal{L}U(\eta_{\min}, 1)$, and a random orientation $\mathbf{R} \sim \mathcal{U}(\mathrm{SO}(3))$,
2. draw $\bar{l} \sim \mathrm{Pois}\left(I_0 4\pi (2\pi/\lambda)^2 I_{\mathrm{np}}(\mathbf{R}, \mathbf{0}, \eta, \rho)\right)$ with the intensity function I_{np} from eq. (3.10),
3. draw photon positions $\mathbf{k}_1, \ldots, \mathbf{k}_{\bar{l}}$ uniformly distributed on the Ewald sphere, and accept each with probability $p_{\mathrm{d}}(\mathbf{k}_i) I_{\mathrm{np}}(\mathbf{R}, \mathbf{k}_i, \eta, \rho) / I_{\mathrm{np}}(\mathbf{R}, \mathbf{0}, x, \rho)$.

This procedure was repeated for each scattering image with different seeds. Note that this rejection sampling works correctly because the intensity function $I_{\mathrm{np}}(\mathbf{R}, \mathbf{k}, \eta, \rho)$ has its maximum at $\mathbf{k} = \mathbf{0}$.

3.3.3 Computation of likelihoods

Likelihoods were computed according to equation (3.11) or (3.9). The integral over $\mathbf{R}$ was approximated by a weighted average over a discrete set of orientations as described in detail in our previous study [115]. Here, we used a precision of 23 for the Lebedev grid and 32 angular orientations, resulting in a total of 6208 orientations. Similarly, the integral over η was approximated by averaging over a discrete set $\eta_1, \ldots, \eta_n$ of $n = 10$ values uniformly spaced between $\eta_1 = 0$ and $\eta_n = 1$, with weights $w_i = p((\eta_{i-1} + \eta_i)/2 < \eta < (\eta_{i+1} + \eta_i)/2)$, using the notation $\eta_0 = 0$ and $\eta_{n+1} = 1$.

3.3.4 Monte Carlo simulated annealing

The positions $\mathbf{y}_i$ of the Gaussian beads, were determined using a hierarchical Monte Carlo simulated annealing approach. As described in detail in our previous study [115], the sampling challenge due to the high number of degrees of freedom at high resolutions is circumvented by determining the structure in multiple hierarchical stages with an increasing number of Gaussian beads. In each step, the density from the previous stage is used as a proposal density, greatly increasing the sampling performance.

To resolve the electron density at reduced resolutions, a regularization procedure is used. To that end, consider a smoothed version of the true electron density function ρ obtained by a convolution with a Gaussian kernel, $\tilde{\rho} = \rho * \mathcal{N}(\sigma)$. The intensity function corresponding to this smoothed version is, due to the Fourier convolution theorem, given by the pointwise product of the original intensity function and the squared absolute value of Fourier transform of the smoothing kernel, $I_{\tilde{\rho}}(\mathbf{k}) = I_{\rho}(\mathbf{k}) \cdot \exp(-\sigma^2 \mathbf{k}^2)$. This means that scattering images corresponding to the smoothed density $\tilde{\rho}$ could be created from those arising from ρ by rejection sampling. These images could then be used to determine the smoothed density.

In consequence, however, much of the available data would be lost. While this could be prevented by averaging over all possible outcomes of this rejection procedure, this would come great computational cost. Therefore, we developed a procedure to approximate this average. To that end, the likelihood function from equation (3.11) was modified by shifting some of the photons originating from the high resolution structure ρ to the background part of the intensity functions. For this, let $s(k)$ be the radial average of $|\mathcal{F}\{\rho\}|^2$ (the powder spectrum of the sample molecule). Then for each k the smoothed structure $\tilde{\rho}$ scatters, on average, $r(k) = s(k)(1-\exp(-\sigma^2 k^2))$ fewer photons with $\|\mathbf{k}\| = k$ than the original structure ρ. By adding this function into the likelihood function as an additional noise-like term,

$$P(\mathbf{k}_1, \ldots, \mathbf{k}_l \,|\, \rho) \propto \int_{\eta_{\min}}^{1} \mathrm{d}\eta \frac{1}{\eta} \int_{\mathrm{SO}(3)} \mathrm{d}\mathbf{R} \left(\frac{\beta}{\langle I_0 \rangle} + \int_E \mathrm{d}\mathbf{k}\, p_{\mathrm{d}}(\mathbf{k}) (I_{\mathrm{np}}(\mathbf{R}, \mathbf{k}, \rho, \eta) + r(\|\mathbf{k}\|)) \right)^{-l-\alpha} \times \prod_{i=1}^{l} (I_{\mathrm{np}}(\mathbf{R}, \mathbf{k}_i, \rho, \eta) + r(\|\mathbf{k}_i\|)), \tag{3.13}$$

the posterior is biased towards lower resolution densities. The radial distribution $s(k)$ is straightforwardly obtained from the scattering images by radial averaging and subtraction of the expected contributions from the noise sources. For the subsequent density determinations for Crambin we used the approximation $s(k) = 2.5\exp(-6k^2) + 0.15\exp(-2k^2) + 0.13\exp(-0.6k^2)$.

3.3.5 Structure alignment and resolution estimate

Because the orientations of the obtained electron densities are random and irrelevant, they were aligned to each other by minimization of the cost function

$$d(\mathbf{S}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1}^{m} \|\mathbf{y}_i - \mathbf{S}\mathbf{y}'_j\| + \frac{1}{m} \sum_{j=1}^{m} \min_{i=1}^{n} \|\mathbf{y}_i - \mathbf{S}\mathbf{y}'_j\|. \tag{3.14}$$

Here, the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\mathbf{y}'_1, \ldots, \mathbf{y}'_m$ define the representation of two electron densities per equation (3.12) and $\mathbf{S}$ is an orthogonal matrix $\mathbf{S} \in \mathrm{O}(3)$, describing both rotations and reflections, as X-ray scattering images do not distinguish between mirror images.

The resolution of the aligned densities was calculated using Fourier shell correlations [120],

$$\mathrm{FSC}(k) = \frac{\int_{\|\mathbf{k}\|=k} \hat{\rho}_1(\mathbf{k})^* \hat{\rho}_2(\mathbf{k})\, \mathrm{d}\mathbf{k}}{\sqrt{\int_{\|\mathbf{k}\|=k} |\hat{\rho}_1(\mathbf{k})|^2\, \mathrm{d}\mathbf{k}} \sqrt{\int_{\|\mathbf{k}\|=k} |\hat{\rho}_2(\mathbf{k})|^2\, \mathrm{d}\mathbf{k}}}, \tag{3.15}$$

where ρ_1 are ρ_2 the densities to be compared and $\hat{\rho}$ denotes the Fourier transform of ρ. The achieved resolution was determined as $2\pi/k_{\mathrm{fsc}}$, where a conservative threshold of k_{fsc} is the value at which the Fourier shell correlation drops below the conservative threshold of ½ [120].

3.4 Results and Discussion

Because our Bayesian approach does not require to determine for each single image the orientation of the molecule or whether it was a hit or not, it should also be able to extract electron densities from a set of images where the number of photons per image is so small that indeed neither is possible. We therefore tested if our method is able to extract electron densities in such a low hit rate low signal-to-noise scenario. Further, we used our Bayesian framework to check that orientation determination and hit classification are indeed not possible in this scenario. Finally, because the Bayesian framework extracts all available information and is therefore particularly well-suited for this, in this section we also analyze the scaling behavior with respect to the expected number of photons per image and the amount of noise to allow one to estimate the required number of images for given resolutions.

3.4.1 Density determination from noisy low hit-rate images

To test our method in this low hit rate low signal-to-noise regime, we selected the same 46-residue protein Crambin [96] as in our previous studies [36, 115], and generated 10^7 synthetic scattering images. Figure 3.1c shows a sample of these super-noisy images, where only the red-colored photons are coherently scattered and thus carry structural information. The same average of 15 coherently scattered photons per central hit ($\eta = 1$) as in our previous study [115] was chosen such that (as will be demonstrated below) the orientations for each image cannot be determined. Further, on average 15 normally distributed photons from elastic scattering on background gas were added, and for inelastic scattering on both the sample molecule and the background gas, on average 135 (90 %) uniformly distributed photons were added each, as should be achievable from experiments [34, 118, 142]. The relative radial distributions of these photon sources are shown in Figure 3.2. Note that these photon numbers refer to all scattered photons, in any direction. Therefore, only a fraction of these photons actually arrive at the detector. The hit rate was chosen to be about 2%, in the sense that 2% of the images contained more than 5 signal photons, with an even smaller fraction of only 0.1% of the images containing more than 15 signal photons. For each image, the distribution of the relative intensity η as in Section 3.2.6 was $\eta \sim \mathcal{LU}(10^{-10}, 1)$. As a result of this setup, for the whole set of 10^7 simulated images, 2% of all the recorded photons were signal photons, and for the central hits 25% were signal photons.

Despite this super-low signal-to-noise ratio, Figure 3.3 shows that still some electron density was recovered. Although at this relatively low resolution the electron density is not accurately recovered, the main features of the molecule are clearly visible. This finding is supported and quantified by the Fourier shell correlation (Figure 3.3c), which provides a conservative resolution estimate of 10.2 Å. Here, two hierarchical stages were used, with density representations consisting of 12 and 23 Gaussian functions, respectively. Although no atomistic resolution was achieved, the obtained resolution is markedly higher than the FSC-resolution of the reference density relative to a perfectly spherical object (described by just one Gaussian bead) of about 18 Å.

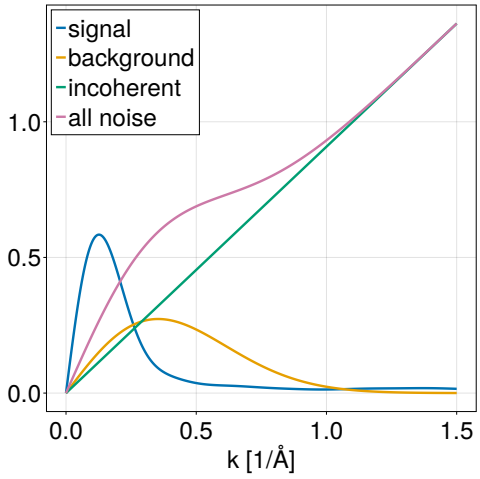

Figure 3.2: Radial photon distribution for a central hit. Shown are coherent scattering on the sample molecule hit (blue), Gaussian background scattering (orange) and incoherent uniformly distributed scattering (green). For the vast majority of images the relative amount of signal photons is much smaller than shown here.

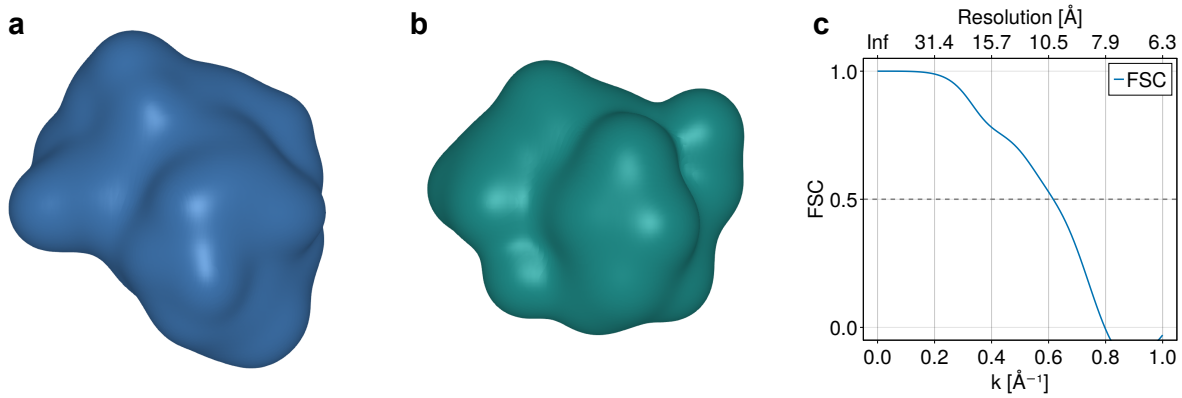

Figure 3.3: Electron density determination for Crambin. **a** Reference electron density. The reference density was smoothed to allow for a better comparison at this resolution level. **b** Reconstructed electron density. **c** Fourier shell correlations show the achieved resolution of 10.2 Å.

3.4.2 Hit classification and orientation determination of single images

To test if in the above scenario it is indeed not possible to accurately identify which of the images are hits, we generated 10^4 scattering images with the same parameters as above, and for each image compared the true value for the relative intensity η at the molecule position with the corresponding maximum likelihood estimates. For this, equation (3.11) served as the likelihood function for the relative intensity η, excluding the integral over η. Note that in order to calculate the likelihood via equation (3.11), one has to assume that the true structure is known. Therefore, if hit selection is not possible in this ideal case, it will certainly not be possible in the realistic case where no structure would be known a priori.

Figure 3.4a shows for each image the true value for η and the corresponding maximum likelihood estimate η_{ml}. Note that the visible vertical 'line' at $\eta = 0$ is a consequence of the log-normal distribution of η, and the horizontal 'lines' at $\eta_{\mathrm{ml}} = 1$ and arises from the fact that, statistically, some images will appear as if η was higher than one, but only values for $\eta_{\mathrm{ml}} \leq 1$ were tested.

The maximum likelihood estimates alone do indeed not suffice to uniquely select the hits, here defined as images with $\eta > 0.01$. For instance, while selecting only the images with a maximum likelihood estimate of $\eta_{\mathrm{ml}} = 1$ would result in set of nearly exclusively hits (true positive rate 90 %), much information would be lost as only a small fraction of the true hits would be selected (true discovery rate 5 %). Opting for a smaller threshold of, say, $\eta_{\mathrm{ml}} > 0.1$ would result in an increased true positive rate of 50% but the true discovery rate would fall to 30%. This is further shown in 3.4b, which compares the true discovery rate with the true positive rate for thresholds for η_{ml} from 0 (upper left) to 1 (lower right). It is clearly visible how a higher true positive rate comes at the cost of a reduced true discovery rate. Importantly, this suggests that no other classification algorithm could achieve significantly more accurate classifications, because the likelihood function captures all available information.

A similar approach was used to test to what extent it is possible to determine the molecular orientation for each image. To this end we used equation (3.9) as a likelihood function for $\mathbf{R}$, but without the integral over $\mathbf{R}$. Again 10^4 images were generated. To simplify the analysis, only central hits (images with $\eta = 1$) were considered. Figure 3.4c shows the obtained distribution of the rotation angle φ between the maximum likelihood orientations and the true orientations for these images. As can be seen, for most of the images the maximum likelihood estimate is off by over 90°. Comparison with a uniform distribution on the rotation group SO(3) (orange line) shows that the maximum likelihood estimates are only marginally more accurate than random guesses. In a realistic scenario without prior knowledge of the structure the estimate would be even worse, even more so for the vast majority of non-central hits. To assess the significance of this result, consider the above protein crambin. To achieve even the low resolution of 10 Å by conventional superposition of oriented images, an orientational accuracy of 60° would be required [70].

Interestingly, the distribution for the estimated orientation shows a pronounced peak at $\varphi = \pi$. The reason for this is is that, due to Friedel's law, the intensity function is antipodally symmetric ($I_\rho(\mathbf{k}) = I_\rho(-\mathbf{k})$), such that $I_\rho(\mathbf{k}) \approx I_\rho(\mathbf{R}_z(\pi)\mathbf{k})$, where $\mathbf{R}_z(\pi)$ is a rotation by 180° around the beam axis and $\mathbf{k}$ is a scattering vector on the Ewald sphere. This would in fact be an equality in the limit of $\lambda = 0$ for which Ewald curvature becomes negligible. No such peak is visible at $\varphi = 0$ due to the vanishing amount of rotations with small rotation angles in SO(3), as is also seen in the rotation angle distribution for the uniform distribution which vanishes near $\varphi = 0$.

3.4.3 Required number of images and scaling behavior

Finally, we asked how many images are required to achieve given resolutions and, further, sought to understand how the required number of images depends on parameters like the expected number of photons per image or the amount of noise. To that end, a large number of independent refinement runs were performed for different values of these parameters (50 to 100 replicas per value), with independently generated synthetic images each. To reduce the substantial computation cost of this parameter scan, we here considered only central hits (images with $\eta = 1$) and used the corresponding likelihood function from (3.9).

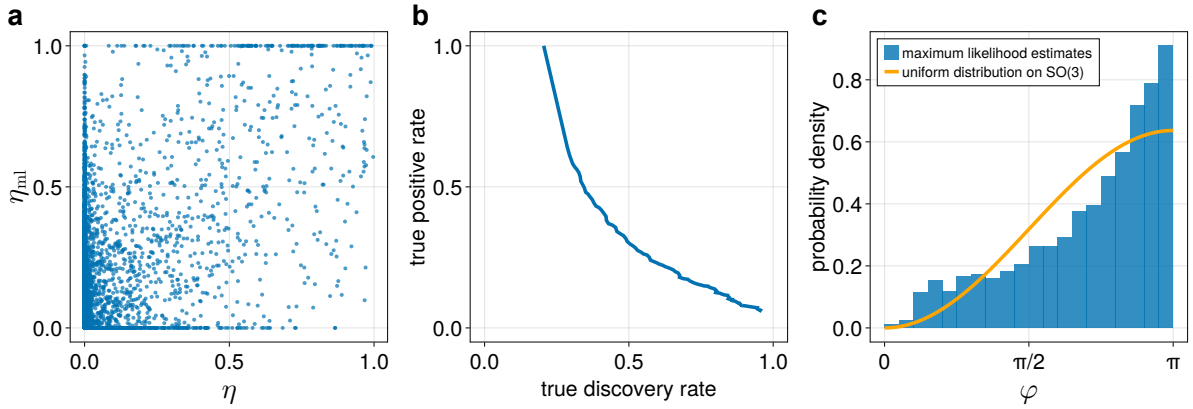

Figure 3.4: **a** True relative intensities η and corresponding maximum likelihood estimates for 10^4 scattering images of Crambin. **b** True discovery rate vs true positive rate for thresholds for the maximum likelihood estimate η_{ml} from $\eta_{\mathrm{ml}} > 0$ (upper left) to $\eta_{\mathrm{ml}} \geq 1$ (lower right). **c** Rotation angle φ between true orientation and maximum likelihood orientation for 10^4 scattering images of Crambin, calculated using 10^4 test orientations. For reference, the distribution of rotation angles for a uniform distribution on SO(3) is shown, given by $p(\varphi) = (1 - \cos(\varphi))/\pi$.

Some of these runs resulted in implausible resolutions on the order of 100 Å, particularly for very low numbers of images. Inspecting the corresponding densities showed that this was caused by some Gaussian beads diverging to infinity. For the subsequent analysis, these values were set to the 'worst case' resolutions of the sample molecule relative to the density described by just one Gaussian function.

Figure 3.5a and 3.5c show the achieved FSC-resolution for each of these independent runs (dots), and the average achieved resolution (lines) for each parameter value. Interestingly, in both cases there seem to be two regions with different relationships between the resolution and the number of images, with a change in slope at around 13 Å resolution. We attribute this to the fact that, as shown in Figure 3.2, the amount of signal photons (and also the signal-to-noise ratio) is much higher for k below the threshold $2\pi/(13\,\text{Å}) \approx 0.48\,\text{Å}^{-1}$ corresponding to this resolution than above.

In both Figure 3.5a and 3.5c, the resolution increases only very slowly with the increasing number of images. This also means that the required number of images increases exceedingly quickly with increasing resolution. For example, for resolutions higher than 13 Å, to increase the resolution by only 1 Å seems to require roughly a 10-fold increase in the number of images. Extrapolating to higher numbers of images suggests that 10^8 to 10^{10} hits would be required to resolve crambin at a resolution of 6 Å, and 10^{12} to 10^{14} hits for 3 Å resolution (assuming log-linear behavior outside of the sampled range). Considering that the proportion of signal photons decreases further with increasing resolutions and that we have here only considered central hits, this is a most conservative estimate, and the required number of images may turn out to be even higher.

To calculate how the required number of images depends on the expected number of photons per image, we calculated for each number of photons and a few resolution thresholds the number of images at which the average obtained resolution becomes lower than that threshold. Figure 3.5b shows these required numbers of images as thick lines. As can be seen, the required number of images is well described by an inverse quadratic relationship $N_{\mathrm{req}} = O(1/n^2)$ (thin lines), where n is the expected number of photons per image. Indeed, only the lines corresponding to 16 Å

and 17 Å differ much from this quadratic relationship, likely because these resolutions are barely below the worst case resolution of about 18 Å between the reference structure and a perfectly spherical object. For these two thresholds the required number of images also becomes close to one, and for obvious reasons always at least one image is required, such that the quadratic relationship cannot hold universally.

This finding strongly suggests that the amount of structural information per scattering image is proportional to the square of the number of signal photons n^2. This is somewhat unexpected, as one might expect that each additional photon should contain the same amount of information, which would imply a linear relationship. Despite this, the quadratic relationship can be explained by the following argument mimicking an orientation determination algorithm. Suppose that, instead of the scattering images, we are given a list of vectors $\mathbf{k}_1, \ldots, \mathbf{k}_N$, where the $\mathbf{k}_i$ drawn independently from the distribution defined by $|\mathcal{F}\{\rho\}(\mathbf{k})|^2$, and known with an uncertainty σ. Because these are independent data points, the uncertainty of the estimate for the electron density ρ should be roughly proportional to $\sigma/\sqrt{N}$. To relate this to the problem at hand, consider that we may estimate the molecule orientation $\mathbf{R}$ for each image $\mathbf{k}_1, \ldots, \mathbf{k}_n$, with an uncertainty roughly given by $\sigma_{\mathbf{R}} \sim 1/\sqrt{n}$. The images can then be combined into a list of vectors as above, each image contributing the vectors $\mathbf{R}\mathbf{k}_1, \ldots, \mathbf{R}\mathbf{k}_n$. The uncertainty of these vectors should then largely be given by the uncertainty of the orientation estimate, $\sigma \sim \sigma_{\mathbf{R}}$. While this does not result in completely independent data points, the uncertainty of the electron density estimated from this should still be $\sigma_\rho \sim \sigma_{\mathbf{R}}/\sqrt{nm} = 1/(n\sqrt{m})$, where m is the number of scattering images. Finally, solving this equation for m yields the quadratic scaling $m \sim 1/(n^2\sigma_\rho)$.

Having understood the scaling behavior in the number of photons, we next looked at the effect of the noise photons, using a similar method as above. For this, the intensity of all sources of noise photons was scaled linearly, with a value of 0 corresponding to zero noise and 1 corresponding to the model described in Section 3.4.1. For each noise amount and for a few resolution thresholds we determined the number of images at which the average obtained resolution (shown in Figure 3.5c) becomes lower than that threshold. Figure 3.5d shows these required numbers of images divided by the number required in the noise-free case as well as corresponding 90%-confidence intervals computed using bootstrapping. While no simple relationship is apparent, there is a very steep increase in the number of images required for 12 Å and 14 Å resolution already for very small amounts of noise. This behavior can be explained in the using the same point-cloud model as above, in which the additional noise photons act both as additional points and by reducing the accuracy of the orientation estimate $\sigma_{\mathbf{R}}$, and these effects combine to cause the steep increase in the required number of images. It is also clearly visible in Figure 3.5d that the effect of the noise becomes stronger for higher resolutions. This is plausible, as the number of photons noise photons increases with increasing magnitudes of the scattering vector.

Because the accuracy of orientation estimates seems to be an essential factor, we investigated it more closely. Figure 3.6 shows how this accuracy depends on the expected number of photons per image and the amount of noise. It is indeed clearly visible that the addition of small amounts of noise does affects the accuracy of the maximum likelihood estimates for the orientation. From Figure 3.6a, it can extrapolated how many photons per image are required for successful orientation determination. For instance, to estimate the orientation to within $0.3\pi = 54^\circ$ about

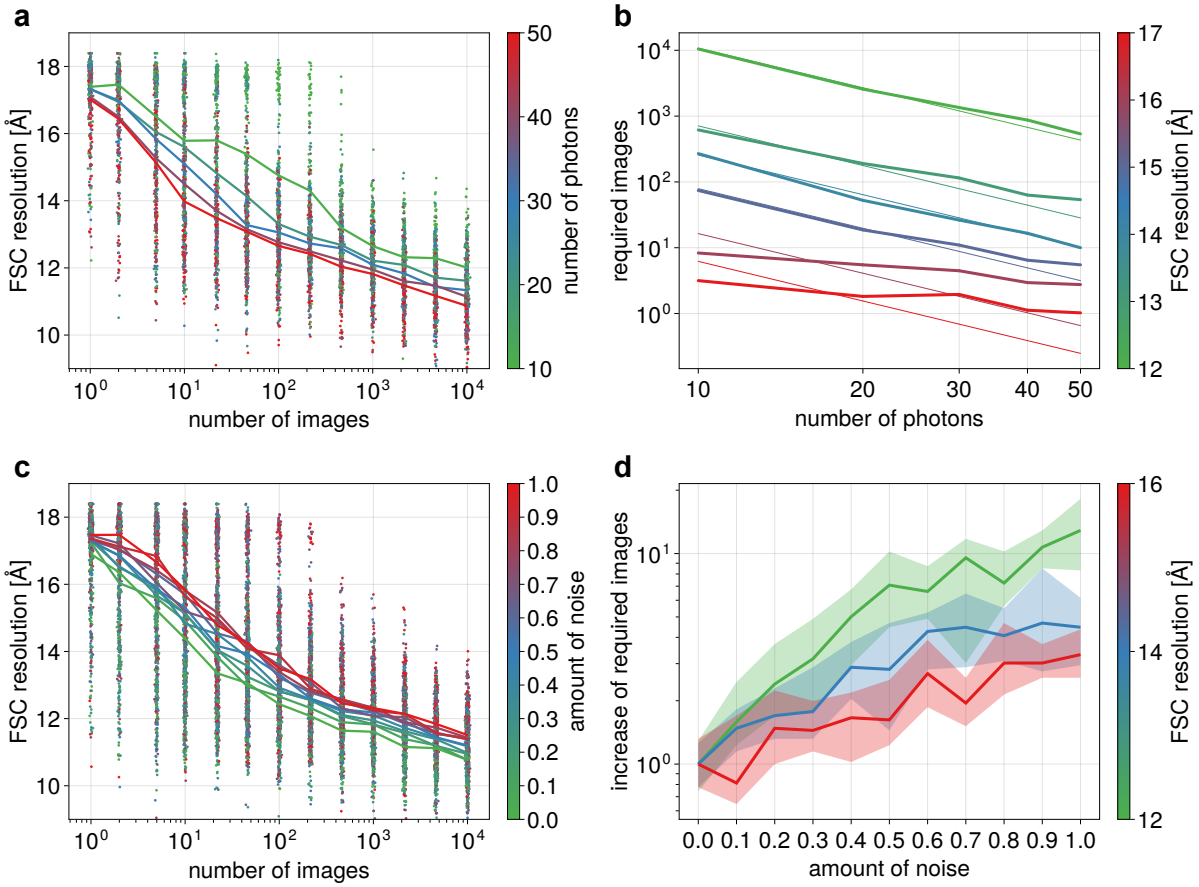

Figure 3.5: Scaling behavior in the number of photons and amount of noise. **a** Achieved resolution for 50 independent density determination runs for each number of images and each expected photon count (dots) and the corresponding averages (lines). **b** Required number of images as a function of the expected number of photons and the resolution with quadratic relationships for comparison (thin lines). **c** Achieved resolution for 100 independent density determination runs for each number of images and each amount of noise (dots) and the corresponding averages (lines). Here, a value of 0 corresponds to noise-free images and a value of 1 to the amount of noise as described in Section 3.4.1. **d** Factor by which the required number of images increases relative to zero noise as a function of the amount of noise for different resolution thresholds (lines) with 90%-confidence intervals calculated using bootstrapping (shaded areas).

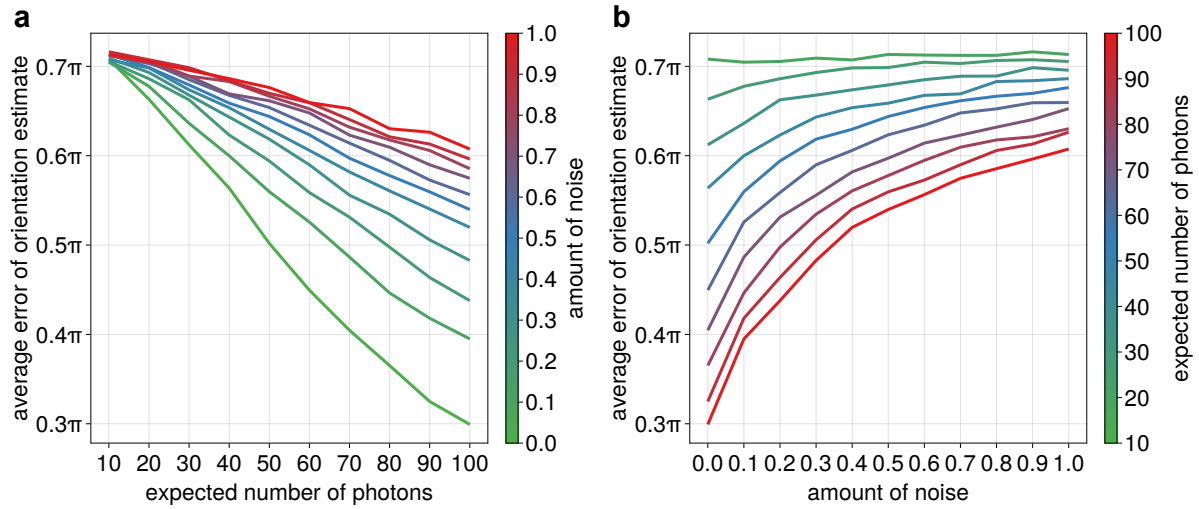

Figure 3.6: Average error of the maximum likelihood orientation estimate as a function of **a** the expected photon count and **b** the amount of noise, calculated using 10^4 synthetic scattering images and 1000 random test orientations each for each value.

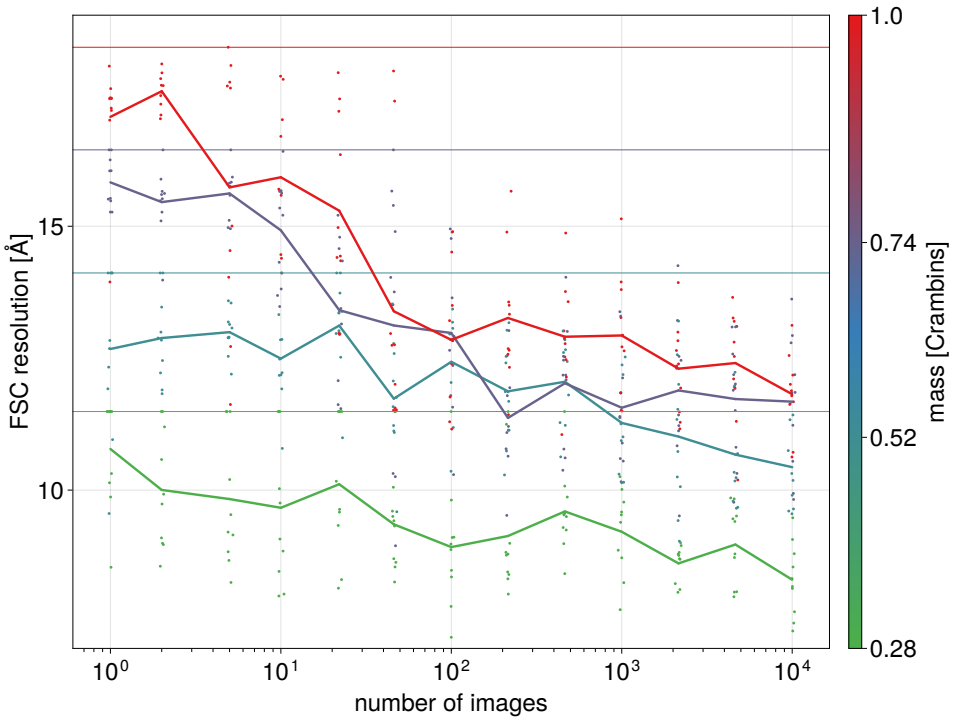

Figure 3.7: Scaling behavior in the molecule size. **a** Achieved resolutions for 10 independent density determination runs for 4 different test proteins (PDB entries 5AWL, 1JLZ, 1ARE and 1EJG) and each number of images (dots) and the corresponding averages (lines), with the corresponding 'worst case' resolutions of the reference densities relative to perfectly spherical objects (horizontal lines). Note that, in particular for low numbers of images, some of the runs resulted in implausible resolutions on the order of 100 Å, caused by one or more Gaussian beads diverging to infinity. For this analysis, these values were set to the corresponding 'worst case' resolutions, visible as the dots directly on the horizontal lines.

100 photons per image are required at zero noise, but about 400 photons are required with the full amount of noise photons. Note however, that as discussed above, these are estimates using the true protein structure, such that in a realistic setting were the structure is not known a priori these numbers will be higher.

Finally, we asked how the size of the sample molecule affects the achieved resolutions. To answer this question, we selected 4 protein structures (5AWL, 1JLZ, 1ARE and 1EJG [96, 97, 151, 152]) from the Protein data bank [130], and for each performed a large number of independent density determination runs (Figure 3.7). The forward model was chosen as above, with on average 15 coherent photons per image for crambin (1EJG) and appropriately fewer for the smaller proteins, down to an average of 2–3 for chignolin (5AWL). Figure 3.7 shows the achieved resolution of a large number of independent density determination runs depends on the number of images. As can be seen, for the two smallest proteins the image numbers were not sufficient to achieve resolutions higher than the respective 'worst case' values. Unfortunately, while for the two largest proteins considered here a small increase in the required number of images for the same resolution is seen, the data are so far insufficient to determine the exact relationship between molecule size and required images.

3.5 Conclusion

Here we have demonstrated electron density determination from highly noisy low hit rate single molecule X-ray scattering images using a Bayesian approach. This approach takes into account many experimental effects such as intensity fluctuations, hits and misses, polarization, irregular detector shapes, incoherent scattering, and background scattering. Our simulated scattering experiments show that electron densities can be reliably determined even in this extreme low hit rate and high noise regime. Furthermore, our results show that our Bayesian approach does not require image classification into hits and misses, which is particularly important because, as we have demonstrated, such a classification as well as orientation determination becomes impossible at the low signal-to-noise ratios considered here.

Despite the super-high noise-level, out approach was able to determine the electron density of the small globular protein Crambin at a resolution of 10.2 Å, which, due to its small size is a particularly challenging test case. While this is still a relatively low resolution, higher resolutions should be achievable, albeit requiring a larger number of images. This is further supported by our analysis of the scaling behavior. While extrapolation of our data suggests that for an atomistic resolution of 3 Å an unrealistic number of 10^{12} to 10^{14} images would be required, already slightly reduced resolutions of, say, 6 Å require substantially fewer images and should therefore within reach.

Our scaling results suggest that for larger proteins even more images would be required for the same absolute resolutions. However, for a similar relative resolution, that is, a similar protein size to resolution ratio, our quadratic scaling result for the number of photons per image implies that much fewer images would be required. Notably, our finding that the required number of images increases with the molecule size may at first glance seem to contradict the previous finding that the achievable resolution should increase for larger molecules [50]. However, in the latter case, this refers to the resolution that can be achieved using an orientation determination approach in the limit of infinitely many images, and does therefore not contradict our finding.

This scaling behavior in the molecule size is also compatible with our previous finding that about $O(m^5)$ images are required to resolve a structure consisting of m Gaussian functions [115], because here we additionally took into account the increase in the number of signal photons and the signal-to-noise ratio for larger proteins. Similarly, this result should also be compatible with other analyses of the size dependence that have come to the conclusion that the achievable resolution should be independent of the molecule [124, 153], because these focused on larger specimen where orientation determination is possible for each image. The scaling behavior clearly deserves further analysis. It will be very interesting to see how the achievable resolutions depend on other parameters, like, for instance, the distribution of the background noise.

While we have here only considered single structures or electron densities, our Bayesian method also allows one to extract structural ensembles [115]. Although not explicitely tested, this should also be possible from the noisy images considered here, albeit requiring more images. In our previous study [115] we further found that extracting such a structural ensemble consisting of n conformers requires only $O(n^2)$ images. For instance, determining a structural ensemble of, say,

2 conformers at 10–12 Å resolution would only require 4 times as many images, which should be within reach of current experiments. This is particularly noteworthy because such structural ensembles already offer valuable insights at much lower resolutions than single electron densities.

The main bottleneck of our approach is its high computational cost, particularly for high relative resolutions and for larger molecules. To address this, improved optimization or sampling methods will be essential, in combination with the use of prior structural information from structure databases, AlphaFold [21], or molecular dynamics force fields. Alternatively, reduced resolutions require much fewer degrees of freedom to represent and present a substantially smaller computational problem. In particular in combination with structural ensembles this will be a worthwhile route to pursue.

While our results show how our Bayesian approach should be well-suited for structure determination from noisy single-molecule X-ray scattering images, we have so far only assessed its performance and accuracy on synthetic scattering images or on preprocessed images of much larger virus specimen [115]. Although the forward model presented in this work is already quite realistic, it will still need to be calibrated and expanded for a future application to experimental data. In particular, some further experimental effects will have to be included.

For instance, we have so far neglected detector noise. The reason is that it presents an unexpected computational challenge. To see why this is the case, consider a simple model in which, in addition to the scattered photons, each detector pixel has a certain probability to be false positive. The amount of the resulting false positive photons would be Poisson distributed, but independent of the beam intensity I_0. Therefore, the distribution of the photons would no longer be independent of I_0 (for example, at high I_0 each photon would have a much lower probability to be false positive than for low I_0). Consequently, I_0 would also appear in the product over i in equation (3.8), such that the integral over I_0 would no longer have a simple analytical expression. In fact, it would have to be reevaluated for each image, at great computational cost.

We have also not yet accounted for a possible solvation shell around the molecule. While fully disordered water would have the same effect as the background scattering and fully ordered water would simply be included in the resulting electron density, partially ordered water poses a challenge. The two main routes to its inclusion are treating is as additional conformers, or as an additional nuisance parameter (in principle, by integrating over all possible water configurations in dependence on the electron density). Here, molecular dynamics approaches will be particularly helpful.

We have also not accounted for multi-hits, in which more than one sample molecule would be present in the beam focus at the same time. An exact treatment in terms of an additional nuisance parameter would require a sum over the number of molecules and computationally expensive integrals over their relative orientations. For cases where the fraction of multi-hits is not too high, an alternative could be to treat the multi-hits as additional conformers, discarding the structural information provided by the respective images.

Finally, the effect of the Coulomb explosion of the sample molecule is noteworthy. While this effect should be negligible at the resolutions considered here, as the atoms in the molecule move

only by up to 2 Å during exposure [118], it is remarkable that to include their motion within the likelihood would come at almost no computational cost. The reason for this is that, as long as this motion is approximately deterministic, the value of the corresponding nuisance parameter (the time of scattering within the pulse duration) is independent for each scattered photon instead of for each image, in contrast to, say, the molecular orientation. Therefore, this effect can be taken into account simply by appropriately modifying the intensity function. It should be possible to include realistic simulations of the effects of radiation damage in a very similar way [154–156]

In the long run, and in addition to the effects analyzed and implemented here, other effects will becomes relevant and will have to be included. This will likely be possible in a similarly straightforward and systematic manner as demonstrated here, again highlighting the great advantage of the Bayesian approach. The calibration of the forward model considered here and the inclusion of these additional effects will be crucial for future protein structure and structural ensemble determination from single-molecule X-ray scattering experiments.

Chapter 4

Time-lagged Independent Component Analysis of Random Walks and Protein Dynamics

The following text has been published as

S. Schultze and H. Grubmüller: **Time-lagged Independent Component Analysis of Random Walks and Protein Dynamics**, *Journal of Chemical Theory and Computation*, 2021 [157].

This project was initiated after observations by Nicolai Kozlowski, Andreas Volkhardt and Helmut Grubmüller. I carried out the research and wrote the manuscript. Helmut Grubmüller supervised the research and revised the manuscript.

Abstract

Time-lagged independent component analysis (tICA) is a widely used dimension reduction method for the analysis of molecular dynamics (MD) trajectories and has proven particularly useful for the construction of protein dynamics Markov models. It identifies those 'slow' collective degrees of freedom onto which the projections of a given trajectory show maximal autocorrelation for a given lag time. Here we ask how much information on the actual protein dynamics and, in particular, the free energy landscape that governs these dynamics the tICA-projections of MD-trajectories contain, as opposed to noise due to the inherently stochastic nature of each trajectory. To answer this question, we have analyzed the tICA-projections of high dimensional random walks using a combination of analytical and numerical methods. We find that the projections resemble cosine functions and strongly depend on the lag time, exhibiting strikingly complex behaviour. In particular, and contrary to previous studies of principal component projections, the projections change non-continuously with increasing lag time. The tICA-projections of selected 1 μs protein trajectories and those of random walks are strikingly similar, particularly for larger proteins, suggesting that these trajectories contain only little information on the energy landscape that governs the actual protein dynamics. Further, the tICA-projections of random walks show clusters very similar to those observed for the protein trajectories, suggesting that clusters in the tICA-projections of protein trajectories do not necessarily reflect local minima in the free energy

landscape. We also conclude that, in addition to the previous finding that certain ensemble properties of non-converged protein trajectories resemble those of random walks, this is also true for their time correlations.

4.1 Introduction

The atomistic dynamics of proteins, protein complexes, and other biomolecules is exceedingly complex, covering time scales from sub-picoseconds to up to hours [158, 159]. It is governed by a similarly complex high-dimensional free energy landscape or funnel [160], characterized by a hierarchy of free energy barriers [1], and has been widely studied computationally by molecular dynamics (MD) simulations [99]. With particle numbers ranging from several hundreds to hundreds of thousands or more [100–103], the correspondingly high-dimensional configuration space of the system poses considerable challenges to a fundamental understanding of biomolecular function, e.g., of the conformational motions of these biological 'nano-machines' [17, 161], protein folding [98], or specific binding.

Several attempts to reduce the dimensionality of the dynamics have addressed this issue. Most notable approaches are principal component analysis (PCA) to extract the essential dynamics [104] of the protein that contributes most to the atomic fluctuations, and time-lagged independent component analysis (tICA), which identifies those collective degrees of freedom that exhibit the strongest time-correlations for a given lag-time [105, 106]. Both dimension reduction techniques can yield information on the conformational dynamics of a protein, i.e., how the protein moves through several conformational substates, which can be defined as metastable conformations characterized by local free energy minima [162].

This property also renders these dimension reduction techniques highly useful as a pre-processing step to describing the conformational dynamics of macromolecules in terms of a discrete Markov process [107–109]. Currently tICA is most widely used, and it is preferred over PCA for this purpose [110] because it additionally uses time information of the input trajectory.

In this context, both PCA and tICA rely on MD trajectories as input, which raises the question how much of these analyses is determined by actual information on the protein dynamics, as opposed to noise due to the inherently stochastic nature of each trajectory, and, importantly, how these two can be quantified.

For PCA, this question has been answered by analysis of the principal components of a high-dimensional random walk in a flat energy landscape [111, 112]. Unexpectedly, these turned out to approximate cosine functions, thus providing a very powerful criterion for the convergence of MD trajectories: The more an MD trajectory resembles a cosine, quantified by the cosine content [111], the more it resembles a random walk, and the less information it contains on the actual protein dynamics or the underlying free energy landscape.

These analyses [111, 112] have also suggested that clusters observed in low-dimensional PCA projections do not necessarily imply the existence of conformational substates and, instead, may also be a stochastic and/or projection artefact. Particularly the latter finding is highly relevant

for the use of PCA for the construction of Markov models [109], which thus may also in part reflect the randomness of one or several trajectories. Note that this holds also true — albeit probably to a lesser extent — for the construction of Markov models from several or many trajectories, as these have to be spawned from a seeding trajectory or from starting structures generated from other advanced sampling methods [162–165].

For tICA, no such analysis is available, but inspection of several examples suggests that similar effects may also be at work [113, 114]. To address this issue, here we will therefore analyze the tICA-projections of high dimensional random walks, and subsequently compare them to tICA-projections of selected protein trajectories. In particular, we will semi-analytically derive an expression for random walk tICA-projections, which will prove analogous to the PCA cosine functions and thus can also serve as a criterion for convergence as well as for the quality of derived Markov models. Unexpectedly, and contrary to the regular behaviour of random walk PCA projections, tICA-projections turn out to display much more complex behaviour. In particular, we observed critical lag times at which the random walk projections change drastically and — for high dimensions — even discontinuously. The resulting much richer and more intricate structure of random walk projections renders the proper interpretation of tICA-projections of protein dynamics trajectories particularly challenging, and has profound implications for the proper constructions of Markov models.

4.2 Theoretical Analysis and Methods

4.2.1 Definition of tICA

To establish notation, we briefly summarize the basic principle of tICA; for a more comprehensive treatment with particular focus on molecular dynamics applications, see Ref. [166].

Consider a d-dimensional trajectory $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))^T \in \mathbb{R}^d$ with Cartesian coordinates $x_1, \ldots, x_d$, which for compact notation we assume to be mean-free, that is, the time average $\langle \mathbf{x}(t) \rangle_t$ is zero. TICA determines those 'slowest' independent collective degrees of freedom $\mathbf{v}_k \in \mathbb{R}^d$, $k = 1, \ldots, d$, onto which the projections $y_k(t) = \mathbf{v}_k \cdot \mathbf{x}(t)$ have the largest time-autocorrelation

$$\frac{\langle y_k(t) y_k(t+\tau) \rangle_t}{\langle y_k(t)^2 \rangle_t},$$

where τ is a chosen lag time. Equivalently, using the time-lagged covariance matrix

$$\mathbf{C}(\tau) = \left(\langle x_i(t) x_j(t+\tau)_t \rangle \right)_{ij} \in \mathbb{R}^{d \times d},$$

each degree of freedom $\mathbf{v}_k$ maximizes

$$\frac{\mathbf{v}_k^T \mathbf{C}(\tau) \mathbf{v}_k}{\mathbf{v}_k^T \mathbf{C}(0) \mathbf{v}_k}$$

under the constraint that it is orthogonal to all previous degrees of freedom. Hence, the $\mathbf{v}_k$ are the solutions of the generalized eigenvalue problem

$$\mathbf{C}(\tau)\mathbf{v}_k = \lambda_k \mathbf{C}(0)\mathbf{v}_k. \tag{4.1}$$

We will use the term 'tICA-eigenvector' for the $\mathbf{v}_k$ and 'tICA-projection' for the projections y_k onto the tICA-eigenvectors. In the literature, the term 'tICA-component' is often used, but it is somewhat ambiguous and we will therefore avoid it.

For an infinite trajectory of a time-reversible system the matrices in this eigenvalue problem are symmetric. However, for the finite trajectories considered here, with time steps $t = 1, \ldots, n$, the matrix $\mathbf{C}(\tau)$ is usually not symmetric. There are two slightly different symmetrization methods that circumvent this problem. The more popular one, which we denote the 'main' method, uses an estimator that replaces the simple time-lagged averages above by averages over all pairs $(\mathbf{x}_t, \mathbf{x}_{t+\tau})$ and $(\mathbf{x}_{t+\tau}, \mathbf{x}_t)$, following e.g. Noé [166] and the popular software package PyEMMA [122]. As a result, on the left hand side of equation (4.1) $\mathbf{C}(\tau)$ is replaced with

$$\mathbf{C}_{\mathrm{sym}}(\tau) = \frac{1}{2}\left(\mathbf{C}(\tau) + \mathbf{C}(\tau)^T\right) = \left(\frac{1}{2}\frac{1}{n-\tau}\left(\sum_{t=1}^{n-\tau} x_i(t)x_j(t+\tau) + \sum_{t=1}^{n-\tau} x_i(t+\tau)x_j(t)\right)\right)_{ij}$$

and on the right hand side $\mathbf{C}(0)$ with

$$\mathbf{\Sigma} = \left(\frac{1}{2}\frac{1}{n-\tau}\left(\sum_{t=1}^{n-\tau} x_i(t)x_j(t) + \sum_{t=1}^{n-\tau} x_i(t+\tau)x_j(t+\tau)\right)\right)_{ij},$$

yielding a symmetrized version of equation (4.1) with real eigenvalues,

$$\mathbf{C}_{\mathrm{sym}}(\tau)\mathbf{v}_k = \lambda_k \mathbf{\Sigma}\mathbf{v}_k. \tag{4.2}$$

The second 'alternative' symmetrized version of equation (4.1) only differs on the right hand side, where $\mathbf{C}(0)$ is not replaced with $\mathbf{\Sigma}$,

$$\mathbf{C}_{\mathrm{sym}}(\tau)\mathbf{v}_k = \lambda_k \mathbf{C}(0)\mathbf{v}_k. \tag{4.3}$$

Our analysis is very similar for both versions, though with unexpectedly different results.

4.2.2 Theory

To render this symmetrized generalized eigenvalue problem more amenable to analysis, and following Ref. [167], we define a matrix formed from the trajectory

$$\mathbf{X} = \begin{pmatrix} | & | & & | \\ \mathbf{x}(1) & \mathbf{x}(2) & \ldots & \mathbf{x}(n) \\ | & | & & | \end{pmatrix}$$

as well as a shorter time-lagged matrix

$$\mathbf{X}_{\text{lag}} = \begin{pmatrix} | & | & & | \\ \mathbf{x}(\tau+1) & \mathbf{x}(\tau+2) & \dots & \mathbf{x}(n) \\ | & | & & | \end{pmatrix}$$

and one that is cut off at the end

$$\mathbf{X}_{\text{cut}} = \begin{pmatrix} | & | & & | \\ \mathbf{x}(1) & \mathbf{x}(2) & \dots & \mathbf{x}(n-\tau) \\ | & | & & | \end{pmatrix}.$$

The latter two matrices serve to re-write the above left and right hand sides,

$$\mathbf{C}_{\text{sym}}(\tau) = \frac{1}{2}\frac{1}{n-\tau}\left(\mathbf{X}_{\text{cut}}\mathbf{X}_{\text{lag}}^T + \mathbf{X}_{\text{lag}}\mathbf{X}_{\text{cut}}^T\right)$$

and

$$\mathbf{\Sigma} = \frac{1}{2}\frac{1}{n-\tau}\left(\mathbf{X}_{\text{lag}}\mathbf{X}_{\text{lag}}^T + \mathbf{X}_{\text{cut}}\mathbf{X}_{\text{cut}}^T\right),$$

and, hence, also the symmetrized tICA-equation,

$$\left(\mathbf{X}_{\text{cut}}\mathbf{X}_{\text{lag}}^T + \mathbf{X}_{\text{lag}}\mathbf{X}_{\text{cut}}^T\right)\mathbf{v}_k = \lambda_k\left(\mathbf{X}_{\text{lag}}\mathbf{X}_{\text{lag}}^T + \mathbf{X}_{\text{cut}}\mathbf{X}_{\text{cut}}^T\right)\mathbf{v}_k\,. \tag{4.4}$$

This defining equation (4.4) for tICA can be converted into a more convenient form using the matrices

$$\mathbf{A} = \begin{pmatrix} 0 & \overset{\tau}{\cdots} & 0 & 1 & & & \\ \vdots & & & \ddots & \ddots & & \\ 0 & & & & \overset{n-\tau}{\ddots} & & \\ 1 & \ddots & & & & 1 & \\ & \ddots & \ddots & & & 0 & \\ & & 1 & 0 & \cdots\cdots & 0 & \end{pmatrix}$$

and

$$\mathbf{B} = \operatorname{diag}\big(\underbrace{1,\dots,1}_{\tau},\underbrace{2,\dots,2}_{n-2\tau},\underbrace{1,\dots,1}_{\tau}\big).$$

Noting that

$$\left(\mathbf{X}_{\text{cut}}\mathbf{X}_{\text{lag}}^T + \mathbf{X}_{\text{lag}}\mathbf{X}_{\text{cut}}^T\right) = \mathbf{X}\mathbf{A}\mathbf{X}^T, \qquad \left(\mathbf{X}_{\text{lag}}\mathbf{X}_{\text{lag}}^T + \mathbf{X}_{\text{cut}}\mathbf{X}_{\text{cut}}^T\right) = \mathbf{X}\mathbf{B}\mathbf{X}^T,$$

equation (4.4) reads

$$\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{v}_k = \lambda_k\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{v}_k\,. \tag{4.5}$$

This can be transformed into a normal eigenvalue problem using the AMUSE-algorithm [168, 169] as follows. First diagonalize the right hand side by an orthogonal matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{\Lambda}$ such that

$$\mathbf{Q}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{Q} = \mathbf{\Lambda}.$$

Substituting $\mathbf{v}_k = \mathbf{W}\mathbf{u}_k$, with $\mathbf{W} = \mathbf{Q}\mathbf{\Lambda}^{-1/2}$, and assuming all diagonal elements of $\mathbf{\Lambda}$ are nonzero, yields

$$\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{W}\mathbf{u}_k = \lambda_k\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{W}\mathbf{u}_k\,.$$

Note that this assumption is actually not necessarily true here, but since we are only interested in the nonzero eigenvalues and their eigenvectors the end results will still be correct. Since $\mathbf{W}$ is invertible, this equation is equivalent to

$$\mathbf{W}^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{W}\mathbf{u}_k = \lambda_k\mathbf{W}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{W}\mathbf{u}_k\,,$$

where the matrix on the right hand side turns out to be the unit matrix,

$$\mathbf{W}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{W} = \mathbf{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{Q}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} = \mathbf{1}\,.$$

Hence equation (4.5) simplifies to

$$\mathbf{W}^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{W}\mathbf{u}_k = \lambda_k\mathbf{u}_k\,. \tag{4.6}$$

Now consider the following 'swapped' version [167]:

$$\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{A}\mathbf{y}_k = \lambda_k\mathbf{y}_k. \tag{4.7}$$

Notably, for each $\mathbf{y}_k$ satisfying equation (4.7) there exists a corresponding eigenvector that solves equation (4.6). Indeed, choosing $\mathbf{u}_k = \mathbf{W}^T\mathbf{X}\mathbf{A}\mathbf{y}_k$ yields

$$\mathbf{W}^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{W}\mathbf{u} = \mathbf{W}^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{A}y = \mathbf{W}^T\mathbf{X}\mathbf{A}\lambda_k\mathbf{y}_k = \lambda_k\mathbf{u}_k.$$

Finally, up to normalization, $\mathbf{y}_k$ is the projection of the trajectory onto the corresponding $\mathbf{v}_k = \mathbf{W}\mathbf{u}_k$,

$$\mathbf{X}^T\mathbf{v}_k = \mathbf{X}^T\mathbf{W}\mathbf{u}_k = \mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{A}\mathbf{y}_k = \lambda_k\mathbf{y}_k.$$

In other words, the tICA-projections of the trajectory are the eigenvectors (with non-zero eigenvalues) of the matrix $\mathbf{M} = \mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{A}$. We will use this reformulation of the tICA defining equation to calculate the tICA-projections of random walks of given finite dimension and length.

4.2.3 Random Walks

For the numerical and semi-analytical evaluation of tICA components, random walk trajectories $\mathbf{x}(t) \in \mathbb{R}^d$ of dimension d were generated by carrying out n steps according to

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \mathbf{r}(t), \qquad \mathbf{r}(t) \sim \mathcal{N},$$

where $\mathcal{N}$ is a d-dimensional univariate normal distribution centered at 0. Each trajectory was centered to zero before further processing. We verified empirically that other fixed probability distributions with mean 0 and finite variance yield similar results.

4.2.4 Molecular Dynamics Simulation

For two proteins a 1 μs molecular dynamics trajectory each was analyzed (Andreas Volkhardt, private communication). Both were generated using the GROMACS 4.5 software package [170] with the Amber ff99SB-ILDN force field [171] and the TIP4P-Ew water model [172]. The starting structures were taken from the PDB [130] entries 11AS [173] and 2F21 [174], respectively. From the latter, only a part of the structure (the WW-domain) was used. Energy minimization was performed using steepest descent for $5 \cdot 10^4$ steps. The hydrogen atoms were described by virtual sites. Each protein was placed within a triclinic water box using gmx-solvate, such that the smallest distance between protein surface and box boundary was larger than 1.5 nm. Natrium and chloride ions were added to neutralize the system, corresponding a physiological concentration of 150 mmol/l. Each system was first equilibrated for 0.5 ns in the NVT ensemble, and subsequently for 1.0 ns in the NPT ensemble at 1 atm pressure and temperature 300K, both using an integration time step of 2 fs. The velocity rescaling thermostat [132] and Parrinello-Rahman pressure coupling [133] were used with coupling coefficients of $\tau = 0.1$ ps and $\tau = 1$ ps, respectively. All bond lengths of the solute were constrained using LINCS with an expansion order of 6, and water geometry was constrained using the SETTLE algorithm. Electrostatic interactions were calculated using PME [136], with a real space cutoff of 10 Å and a fourier spacing of 1.2 Å. The integration time step was 4 fs, and the coordinates of the alpha carbons were saved every 10 ps, such that 10^5 snapshots were available for each trajectory. Of these we discarded the first 10^4 steps, leading to trajectories of length $n = 9 \cdot 10^4$.

4.3 Results and Discussion

To characterize the tICA components and projections of random walks, we will proceed in two steps. We will first analyse a special case, for which some analytical results can be obtained. Second, we will use the obtained insights to generalize this result to random walks of arbitrary length n and dimension d using a combined analytical/numerical approach. Subsequently, we will compare the obtained random walk projections to tICA analyses of biomolecular trajectories.

4.3.1 A Special Case

To gain first insight into the tICA components of a random walk, first consider the special case $d = n$, which allows for an almost fully analytical approach. In this case, all matrices in equation (4.7) are square and, assuming that $\mathbf{X}$ is invertible,

$$\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X} = \mathbf{X}^T(\mathbf{X}\mathbf{B}\mathbf{X}^T)^{-1}\mathbf{X} = \mathbf{X}^T\mathbf{X}^{-T}\mathbf{B}^{-1}\mathbf{X}^{-1}\mathbf{X} = \mathbf{B}^{-1},$$

such that equation (4.7) becomes independent of $\mathbf{X}$,

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{y}_k = \lambda_k\mathbf{y}_k. \tag{4.8}$$

Note that the assumption that $\mathbf{X}$ is invertible is not strictly correct, as it has one zero-eigenvalue associated to the eigenvector given by $\mathbf{y}_0 = (1, \ldots, 1)^T$. This is also an eigenvector of $\mathbf{B}^{-1}\mathbf{A}$, but instead with eigenvalue 1. Therefore all the eigenvectors and all but one eigenvalue of equation (4.7) are identical to those of equation (4.8), and the analysis can proceed using equation (4.8).

In the limit of large n, and using the above definitions for $\mathbf{A}$ and $\mathbf{B}$, the matrix $\mathbf{B}^{-1}\mathbf{A}$ approaches a circulant matrix with the property that each of its columns is a cyclic permutation of the preceding one. It differs from a circulant matrix only at the four 'corners' (of size τ) of the matrix, and for large $n = d$ these 'corners' become small relative to the size of the matrix. More precisely, $\mathbf{B}^{-1}\mathbf{A}$ and the circulant matrix are asymptotically equivalent as in defined in Ref. [175].

Circulant matrices are diagonalized by the Fourier transform [176], yielding eigenvectors are

$$\tilde{\mathbf{y}}_k = \left(1, \omega_k, \omega_k^2, \ldots, \omega_k^{n-1}\right), \qquad \omega_k = \exp\left(2\pi\mathrm{i}\frac{k}{n}\right).$$

and eigenvalues

$$\lambda_k = \frac{\omega_k^\tau + \omega_k^{n-\tau}}{2} = \cos\left(2\pi\frac{\tau k}{n}\right). \tag{4.9}$$

These eigenvectors are complex, but since $\lambda_k = \lambda_{n-k}$ and $\tilde{\mathbf{y}}_k = \tilde{\mathbf{y}}_{n-k}^*$, the real and imaginary part of $\tilde{\mathbf{y}}_k$ (cosine and sine) are real eigenvectors for the same eigenvalues. Depending on τ and n, many of these eigenvalues are equal, since they only depend on τk mod n.

This result implies that for large $n = d$ the eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$ approach those of the circulant matrix. More precisely, their eigenvalues asymptotically equally distributed [175]. In contrast, the eigenvectors are only preserved in limits or under small perturbations if the respective adjacent eigenvalues are well-separated from each other [177]. For the case at hand, however, this eigenvalue separation very quickly approaches zero for small k and large n (and for other k with $|\cos(2\pi\tau k/n)| \approx 1$). As a result, the eigenvectors of $\mathbf{B}^{-1}\mathbf{A}$ for small k (and other k as before) differ from those of the circulant matrix even in this limit. Rather, they need to be represented as approximate linear combinations of those eigenvectors of the circulant matrix with similar eigenvalues.

This subtlety contributes to the complexity of the problem as well as of the solution, and has so far prohibited us from proceeding further purely analytically both for finite $d = n$ as well as for $d = n \to \infty$. Nevertheless, the eigenvalue problem equation (4.8) provides a good starting point for a numerical approach. Still, the degeneracy discussed above needs to be taken properly into account, as the numerical eigenvectors are essentially arbitrarily chosen from the eigenspaces.

Inspecting the Fourier transforms of the numerical eigenvectors suggests that the eigenspaces of equation (4.8) for small k each contain an eigenvector that resembles a cosine function

$$y_k(t) \approx \cos\left(\pi \frac{tk}{n}\right),$$

with increasing accuracy for increasing n.

Another effect of the poor separation of the eigenvalues is that the above results are very sensitive to small changes to the matrix in equation (4.8). E.g., using the alternative symmetrization method defined by equation (4.3), the analysis in Section 4.2.2 is unchanged, except that all diagonal entries of B become 2, and equation (4.8) reads

$$\frac{1}{2}\mathbf{A}\mathbf{y}_k = \lambda_k \mathbf{y}_k\,.$$

For $n = d \to \infty$, the same circulant matrix is obtained, such that the eigenvalues, equation (4.9), are unchanged. The numerical solution however reveals that the first few eigenspaces instead contain eigenvectors given by

$$y_k(t) \approx \sin\left(2\pi \frac{tk}{n}\right).$$

This result is indeed strikingly different, in that the cosine functions are replaced by sine functions with twice the frequency.

4.3.2 General Solution

Next, we will consider the general case, i.e., a random walk of length n in $d < n$ dimensions. Unfortunately, we were unable to find analytical solutions similar to the above; however, the results of Section 4.2.2 permit an elegant way for a numerical approach by computing the expectation value of the matrix $\mathbf{M}$. To this aim, $\mathbf{M}$ was computed for a sample of 20000 random walks of given fixed dimension d and number of time steps n, from which an average matrix $\langle\mathbf{M}\rangle$ was computed. The eigenvectors of $\langle\mathbf{M}\rangle$ served as the semi-analytical solution for the general case. We note that this does not necessarily produce the same results as averaging the individual tICA-projections directly. We have, however, tested that the eigenvectors of $\langle\mathbf{M}\rangle$ are very similar to the averages of the tICA-projections. An exception to this is that averaging the tICA-projections can produce artefacts arising from to the fluctuating order of the eigenvectors, and these artefacts are not present in the eigenvectors of $\langle\mathbf{M}\rangle$.

As an illustration, Figure 4.1 shows the first two resulting tICA-projections for random walks with $n = 1000$ and $d = 50$, revealing a strong dependence on the lag time τ. For short lag times τ, $y_1(t) \approx \cos(\pi t/n)$ and $y_2(t) \approx \cos(2\pi t/n)$. With increasing τ, this low-frequency cosines are

gradually replaced by higher-frequency components, first in $\mathbf{y}_2$ (starting at about $\tau = 90$) and for further increasing $\tau > 150$ also in $\mathbf{y}_1$. From then on, the frequencies of both $\mathbf{y}_1$ and $\mathbf{y}_2$ slowly decrease, maintaining a π phase shift.

In contrast to the special case considered above (Section 4.3.1), our numerical studies suggest that for large lag times the averaged projections do not approach exact cosines for large n. Rather, 'cosine like' functions appear, as can be seen for the high lag-times shown in Figure 4.1, where the circular shape that would be expected for exact cosines is noticeably distorted, even if n is further increased. In contrast, for short lag times, where the higher frequency components have not yet appeared (e.g. $\tau < 90$ in Figure 4.1), the projections do seem to approach exact cosines with increasing n.

For the alternative symmetrization method, equation (4.3), the same method can be applied, and the obtained projections are shown in Figure 4.2. Indeed, comparing the two Figures, even more dramatic differences are seen as a result of this very small change. In particular, for short τ values, the cosine-like functions seem to be replaced by sine-like functions of twice the frequency, just like we have already seen for the special case $d = n$. Also, for increasing τ a much richer and complex behavior is seen. Finally, the onset of higher frequencies occurs for somewhat smaller τ values (at $\tau \approx 100$) compared to Figure 4.1 (at $\tau \approx 110$). This abrupt emergence of higher frequencies deserves closer inspection.

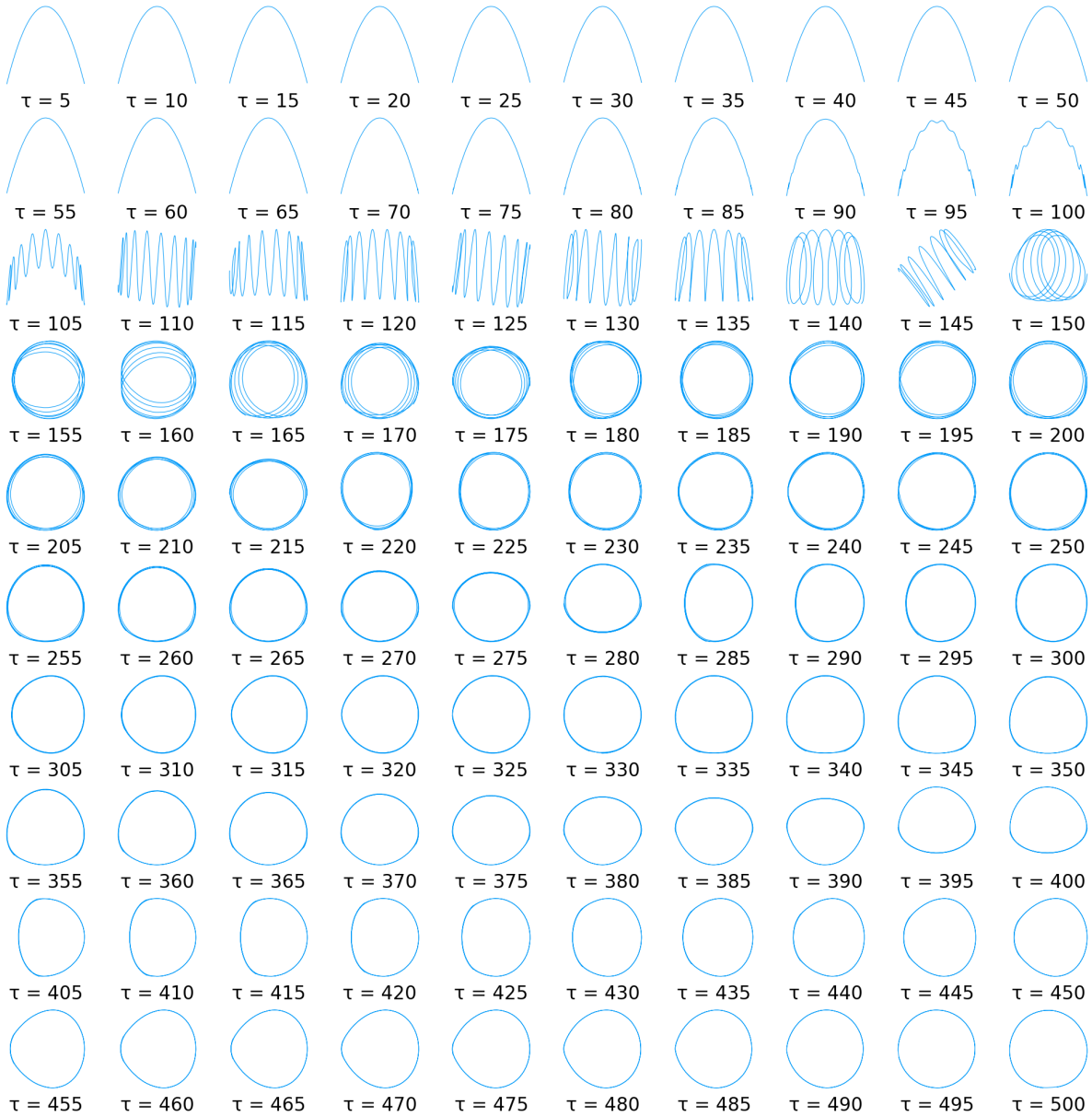

Figure 4.1: The first two 'expected' tICA-projections of random walks of dimension $d = 50$ with $n = 1000$ time steps for varying lag time τ, computed with the averaging method from Section 4.3.2 using a sample of 20000 random walks. For each τ, the first tICA-projection is shown on the x-axis and the second one on the y-axis.

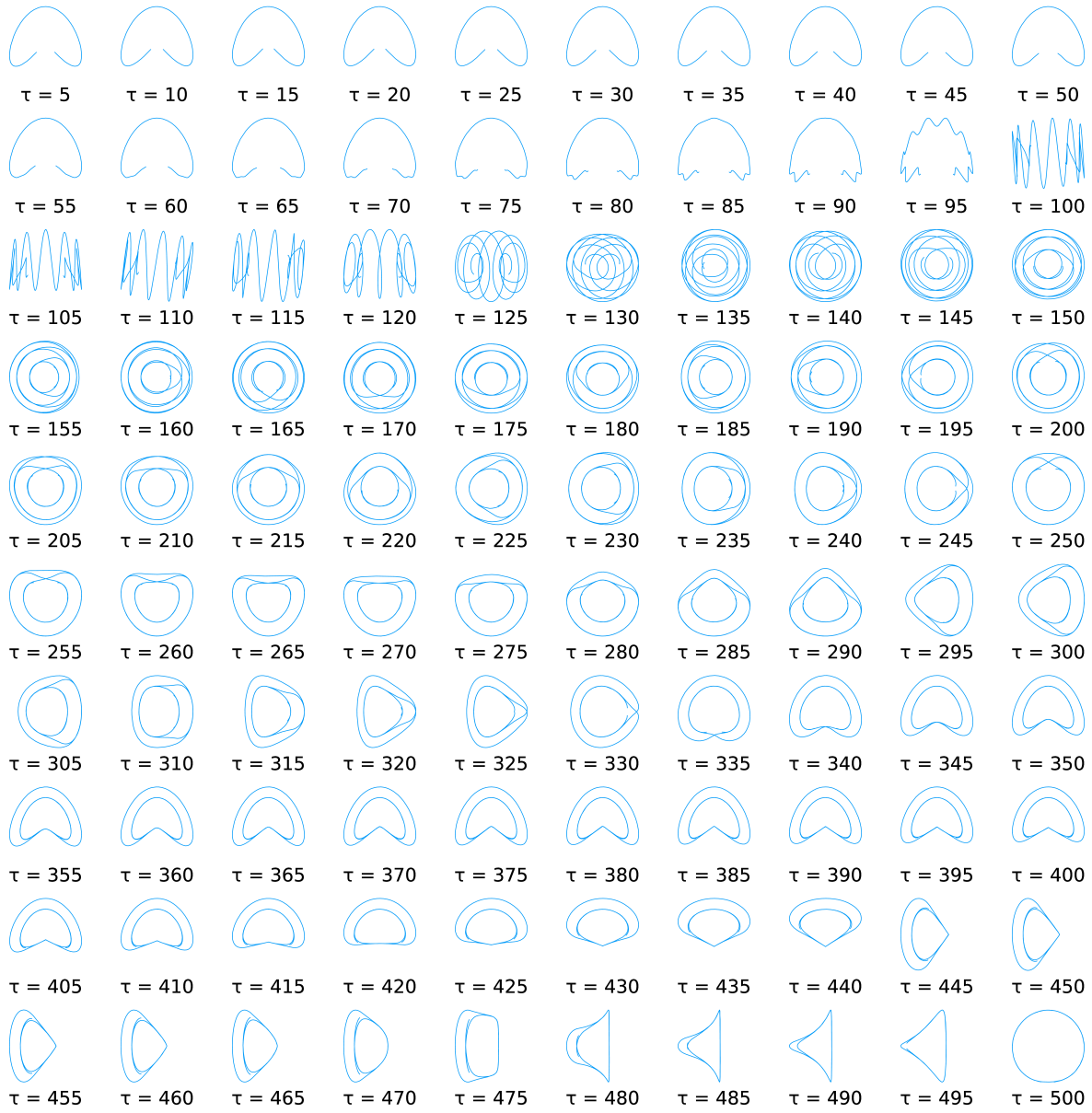

Figure 4.2: The first two 'expected' tICA-projections, for the alternative symmetrization method, of random walks of dimension $d = 50$ with $n = 1000$ time steps for varying lag time τ, computed with the averaging method from Section 4.3.2 using a sample of 20000 random walks. For each τ, the first tICA-projection is shown on the x-axis and the second one on the y-axis.

4.3.3 Abrupt Changes

To gain more insight into why these abrupt changes occur, Figure 4.3 (A) shows the eigenvalues of $\langle \mathbf{M} \rangle$ as a function of τ for dimension $d = 30$, revealing a strikingly complex pattern. For small lag times τ all eigenvalues decrease with τ, with associated cosine-shaped eigenvectors of period lengths $2n, 2n/2, 2n/3, \ldots$, as annotated in the Figure. The decrease of these curves reflects the sampling of the cosine-shaped eigenvectors with increasing lag time τ and, hence, the respective autocorrelations also resemble cosine functions.

Also visible are several curves that monotonically increase with τ, each starting at zero for small τ. These curves represent two eigenvalues each, with cosine-shaped and sine-shaped eigenvectors of period lengths $\tau, 2\tau, 3\tau, \ldots$, respectively, as also annotated in the Figure. Their increase is less obvious, as one might expect the autocorrelation of a τ-periodic function at lag time τ to be unity and, therefore, constant. Note, however, that the eigenvalue of $\langle \mathbf{M} \rangle$ does not strictly represent this autocorrelation; rather, it represents the average of the autocorrelations of many instances of this eigenvector for each single random walk — each of which is not strictly periodic. For increasing period lengths, the eigenvectors approach cosines or sines, such that their average autocorrelation increases and so do the corresponding eigenvalues of $\langle \mathbf{M} \rangle$.

At the intersections of these two sets of curves (black circles) the respective eigenvalues are degenerate and their order changes, which causes abrupt changes of the eigenvectors and, therefore, also of the projections onto these eigenvectors, the first two of which were discussed above.

For larger dimensions d, e.g., for $d = 50$ as shown in Figure 4.3 (B), one would expect that the tICA-projections resemble cosine or sine functions increasingly closely, also also at increasingly higher frequencies. As a result, the eigenvalues corresponding to the eigenvectors with period lengths $\tau, 2\tau, 3\tau, \ldots$ should increase with d at any given lag time τ, whereas the decreasing eigenvalue curves on the left side should remain unchanged. Therefore, the respective intersections should occur at smaller lag times τ. Comparison of the black circles in the two panels of Figure 4.3 shows that this is indeed the case. To illustrate this effect, Figure 4.4 shows the first two tICA-projections of random walks with dimensions ranging from 50 (top row) to 500 (bottom row) for increasing τ.

To quantify this behaviour, we generated a large number of random walks and determined the lag times τ at which the abrupt changes occur. Figure 4.5 shows the first and second of these critical lag times as a function of dimension d and for n ranging from 1000 to 5000 (colors). To enable direct comparison, the lag times τ have been normalised by n. As can be seen, for d between ca. 150 and $n/2$ both the first (upper curves) and second (lower curves) approximate power laws $n/\tau \propto d^b$, as indicated by the respective fits (solid lines, the colors correspond to the values of n). For each fit, only dimensions d within the above range have been used.

The inset of Figure 4.5 shows the power law exponents b for varying n and for the first and second abrupt change, both of which apparently approach $b = -1/2$ for large n (also represented by the black lines in the main Figure). Although we were unable to find a rigorous proof, this finding suggests that in the limit of large n and d, with d markedly smaller than n, the first few lag times at which abrupt changes occur scale as $\tau \propto n/\sqrt{d}$.

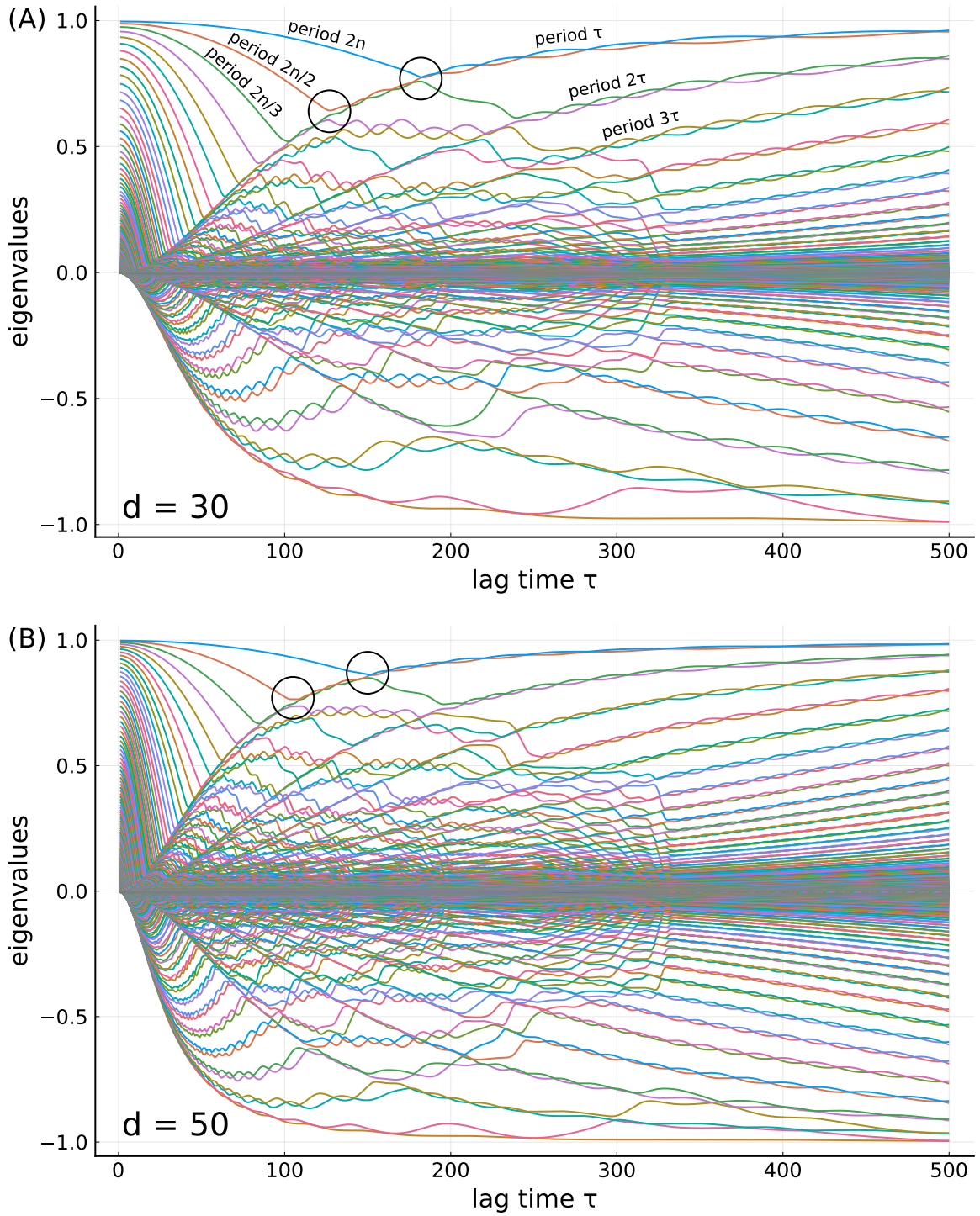

Figure 4.3: The eigenvalues of the averaged matrix $\langle \mathbf{M} \rangle$ as a function of the lag time τ at (A) dimension $d = 30$ and (B) dimension $d = 50$. The two abrupt changes are indicated using black circles. The colors indicate the order of the eigenvalues.

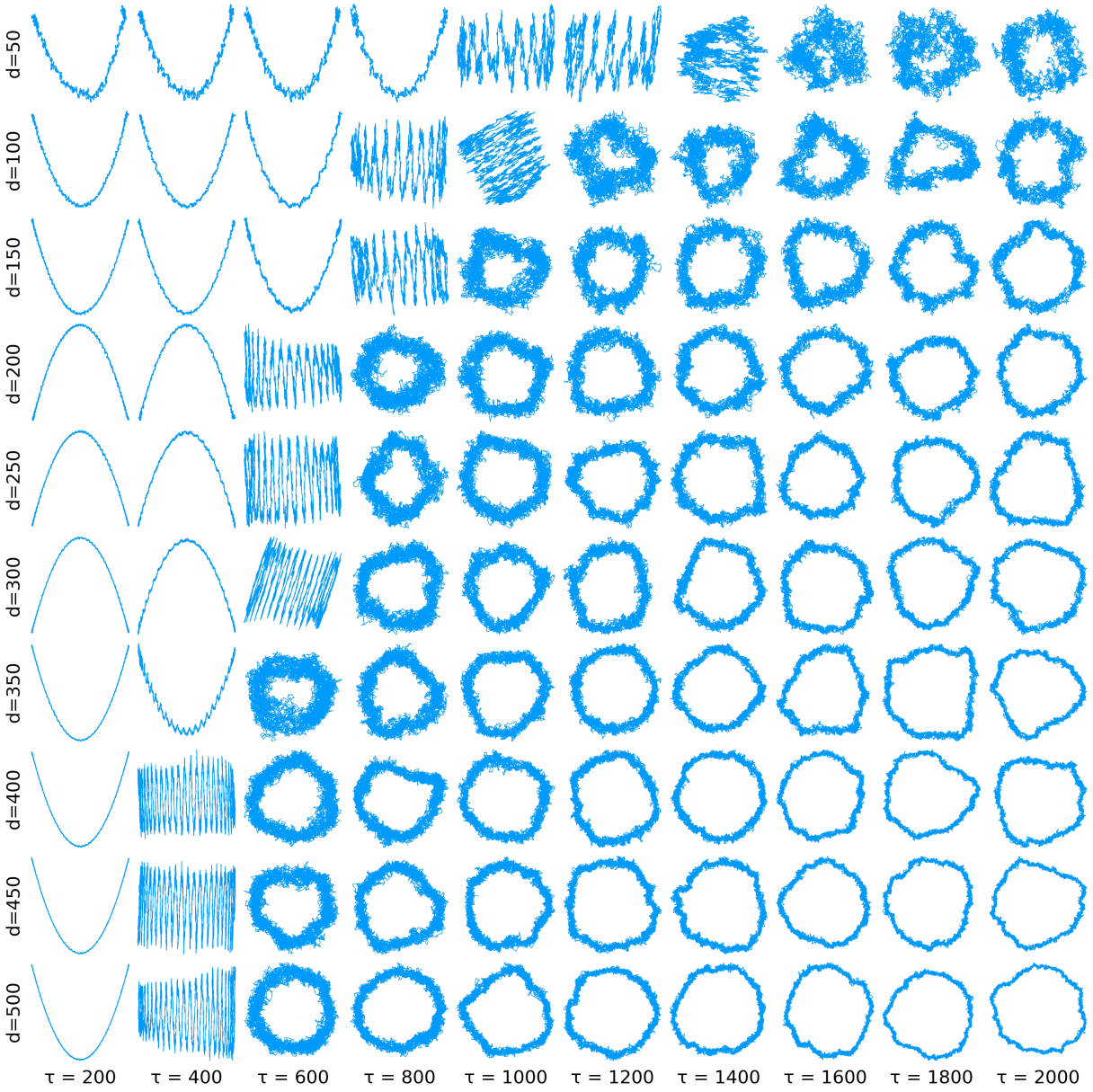

Figure 4.4: The first two tICA-projections of random walks with varying dimensions d, each with $n = 10000$. The lag times of the abrupt changes decrease with increasing dimension.

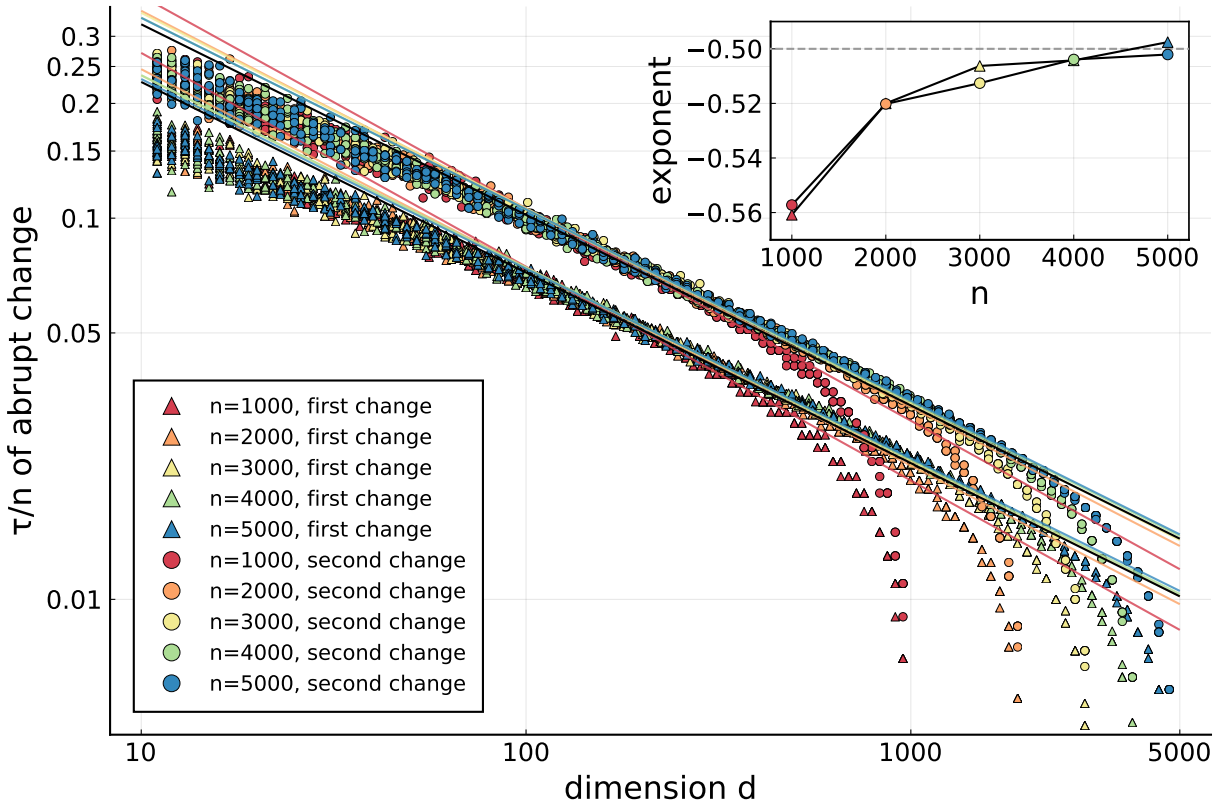

Figure 4.5: The lag time at which the abrupt changes occur in dependence of the dimension for various n. Each dot represents an independently generated random walk. Also shown are the power law fits $n/\tau = a \cdot d^b$ (colored lines), their exponents (inset), and the lines corresponding to $b = -0.5$ (black lines).

4.3.4 Comparison of Random Walks and MD-trajectories

We next compared the tICA-projections of random walks with those of molecular dynamics trajectories of proteins in solution. To that end, we used two MD-trajectories of length 1 μs each (generated as described in Section 4.2.4), one of a comparatively large protein (PDB 11AS, 330 amino acids) [173] and one of a smaller protein (WW-domain of PDB 2F21, 34 amino acids) [174].

As can be seen in Figure 4.6, the tICA-projections of the larger protein (top group) are indeed spectacularly similar to those of a random walk (bottom group). Even the strong dependence on the lag time is very similar, as are the abrupt changes discussed above.

Note that this striking similarity was obtained for a particular choice of $d = 40$ for the random walk; other dimensionalities yield less similar projections. Intriguingly, this finding thus suggests a new method of estimating an 'effective' dimensionality of MD trajectories.

It is also worth noting that both the MD-trajectory and the random walk projections show apparent 'clusters', e.g. for $\tau = 500$ and $\tau = 8000$, which also look quite similar. The fact that such clusters are also seen for the random walk strongly suggests that these are mostly stochastic artefacts and do not point to minima of the underlying free energy landscape.

Closer inspection of the random walk projections offers an additional possible explanation for some of the clusters, which may also apply to the MD trajectory projections. Focusing, e.g., at the averaged tICA-projections in Figure 4.1 immediately before the first abrupt change, one can see that the projection becomes overlayed with a cosine of higher frequency. Particularly at the ends of the curves, and in the presence of noise typical for single trajectories, this high frequency component can also produce apparent 'clusters'.

In contrast, for the smaller protein (Figure 4.7) no similarity to the tICA-projections of random walks is observed. In fact, the tICA-projections of the trajectory of the smaller protein show no resemblance to a cosine-like function at all. In light of the above analysis, this finding suggests that this trajectory is sufficiently long to explore one or several minima of the underlying free energy landscape, thereby deviating from a random walk. Further, one may infer that the three clusters seen in the Figure actually point to conformational substates and, hence can serve as proper Markov states.

It is an intriguing question whether or not, for given trajectory length, larger or more flexible proteins tend to more closely resemble random walks.

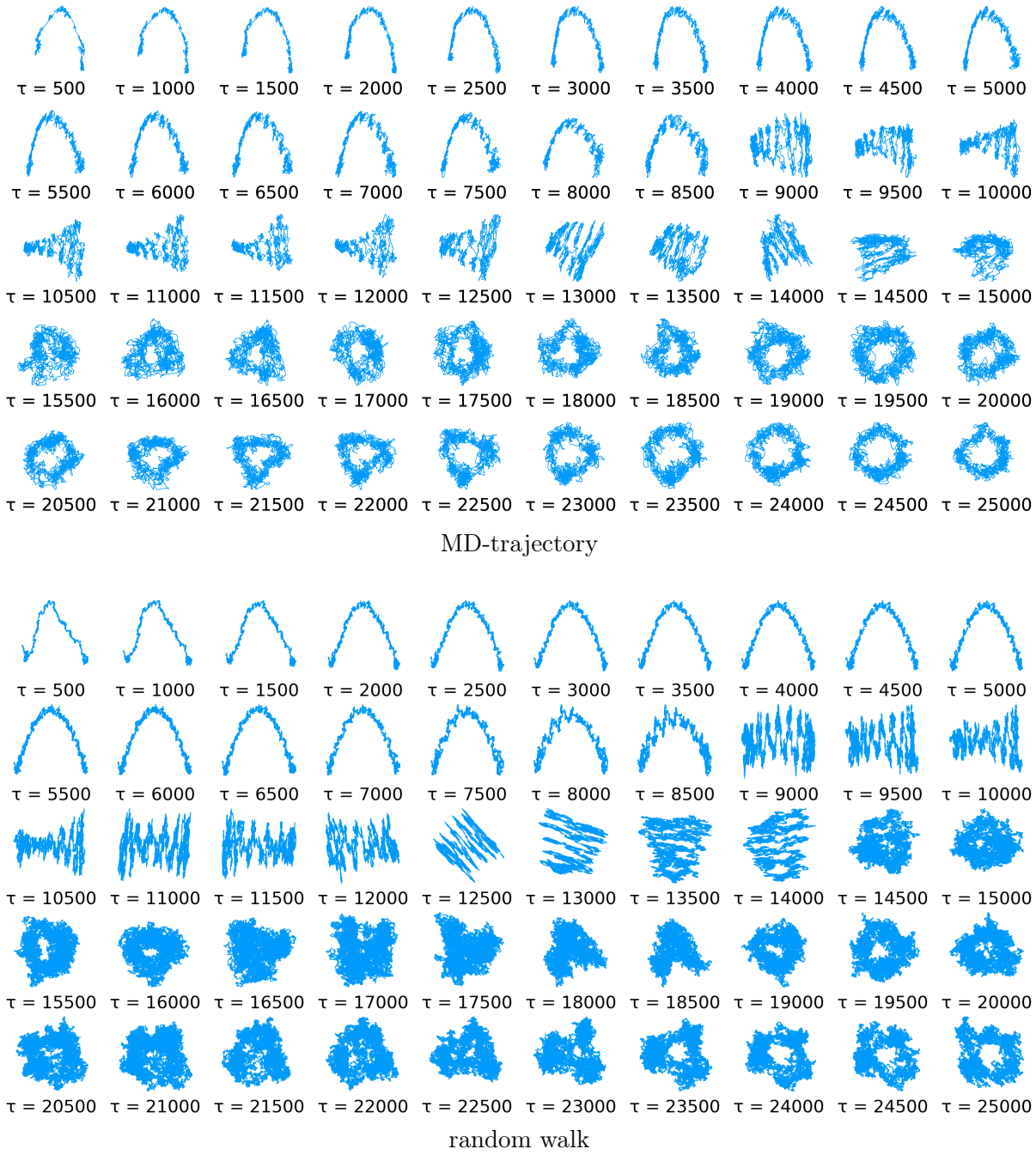

Figure 4.6: The first two tICA-projections of an MD-trajectory of PDB-entry 11AS (upper group) and those of a 40-dimensional random walk (lower group) for varying lag time τ. In this plot those of the MD-trajectory are smoothened using a moving average to improve readability.

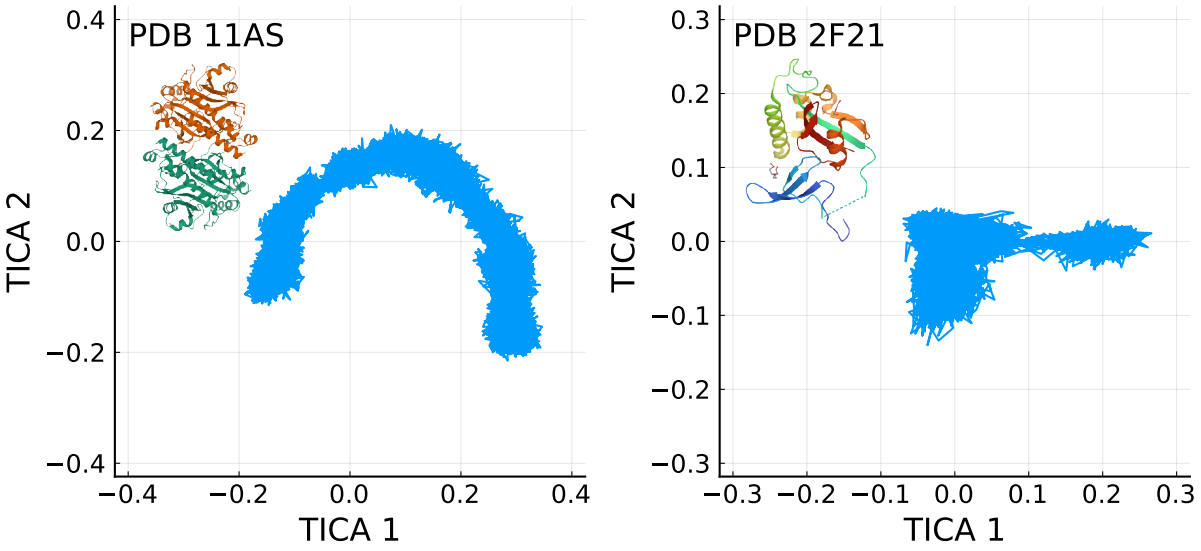

Figure 4.7: The first two tICA-projections of trajectories of the PDB-entries 11AS (on the left) and 2F21 (on the right). The larger protein (11AS) produces a cosine-like shape while the smaller one does not.

4.4 Conclusions

Here we have analysed projections of random walks on tICA subspaces and subsequently compared those to tICA-projections of molecular dynamics trajectories of proteins. Our combined analytical and numerical study revealed a staggering complexity of the random walk tICA-projections, which showed a much richer mathematical structure than projections of random walks on principal components (PCA) [111, 112].

We attribute this complexity primarily to the fact that, in contrast to PCA, tICA components encode time information of the trajectory and, therefore, extract and process significantly more information. Mathematically, the complex behavior originates from the non-continuous switch of the order of eigenvalues for increasing lag time τ, when passing through points of eigenvalue degeneracy. At these points, the associated eigenvectors change abruptly, and so do the corresponding projections of both random walks and molecular dynamics simulations. We also find that tICA can be very sensitive to very small changes in the definitions of the involved matrices. In particular, the projections of random walks are very different for the two discussed symmetrization methods.

A particularly striking example is the first abrupt change of the projections onto the two largest eigenvalues. Here, a closer inspection revealed an approximate square root relationship between the lag times at which this occurs and the dimensionality of the random walk. A similar square root law is already known for PCA: Approximately the first $\sqrt{d}$ principal components of random walks resemble cosines [111].

Comparison of tICA-projections of random walks with those of a large protein (PDB 11AS) revealed striking similarities. This remarkable finding suggests that not only the ensemble properties of the finite protein trajectory resemble those of a random walk, as has been shown earlier via PCA [111], but also the time correlations of the underlying protein dynamics. Here, the appearance of cosine-like functions in the projections onto the tICA-vectors associated with

the longest correlation times clearly points to a non-converged trajectory. For the comparatively small lag times typically used, the tICA-projections of random walks almost exactly resemble cosine functions, such that the cosine-content [112] of the tICA-projections should serve as a good quantifier of this.

In contrast, no resemblance to a random walk was seen for the second, smaller protein studied here, indicating that the projection reflects actual features of the underlying conformational dynamics of the protein.

The example in Figure 4.6 also illustrates the risk of over-interpreting apparent 'clusters' seen in the tICA-projections as actual conformational substates [1, 162], which are defined as local minima of the protein free energy landscape that are sufficiently deep for the system to stay there for a certain amount of time [162]. Clearly, it is tempting to also see 'clusters' in the random walk projections, which, however, by the definition of the random walk as a diffusion on a flat energy landscape, cannot represent conformational substates. This finding raises concerns for using automated clustering algorithms to identify, e.g., folding intermediates or to characterize conformational motions from tICA-projections [178].

Because the additional parameter of a varying lag time provides a much richer structure and many instead of only one projection (as is the case for PCA), we speculate that tICA resemblance to a random walk offers a much more sensitive tool to detect lack of convergence in MD trajectories of large biomolecules. Further, by adjusting the dimension of the random walk such as to maximise the similarity to a given MD trajectory, one can estimate the effective dimensionality of the underlying dynamics. The latter idea, as well as precisely how this 'effective dimensionality' can be defined, clearly deserves further exploration.

4.5 Acknowledgements

We thank Nicolai Kozlowski, Malte Schäffner and Andreas Volkhardt for very helpful discussions; and Andreas Volkhardt for providing the MD-trajectories for our analysis. This work was supported by the German Ministry of Education and Research, BMBF project 05K20EGA and the German Science Foundation, grant SFB 1456.

This analysis has been implemented using the Julia programming language [95].

Chapter 5

Conclusion

5.1 Single-molecule X-ray scattering

In Chapter 2 and 3, a novel Bayesian approach for structure and structural ensemble determination from single-molecule X-ray scattering images was developed. In contrast to most established methods, this approach does not require prior hit selection, classification, or orientation determination, nor does it perform these steps itself. A a first main result, I demonstrated that single electron densities can indeed be determined using this approach. My results show that fewer images are required than using established methods, as is expected because this Bayesian approach uses all available structural information. As a second main result, my findings show that not only single electron densities but entire structural ensembles can be determined from single-molecule X-ray scattering images. For this ensemble determination, far fewer scattering images are required than expected, which may put them within reach of current experiments.

The fact that my approach does not rely on hit selection, classification, or orientation determination is particularly important considering my finding that these are indeed not possible for the sparse and highly noisy images expected for small proteins. In contrast, it is generally assumed in the field that orientation determination is required, due to the fact that, with few exceptions, all commonly used approaches do require it [34]. This assumption has so far also meant that almost only those experiments have been undertaken where orientation determination was expected to be successful [46, 62, 179]. Indeed, my results have already motivated new experiments that were previously deemed too challenging and have started collaborations with multiple other research groups. Further, a proposal for beam time at the European XFEL has been submitted by a cooperating group partly based on my results.

Many complicating experimental effects such as Ewald curvature, incoherent scattering, background scattering, beam polarization, irregular detector shapes, hits and misses, and intensity fluctuations were explicitly included in both the simulated scattering images and the likelihood function used to obtain these results. I am not aware of any other established approach that includes all these effects in such a systematic and rigorous way as is possible in the Bayesian formalism.

On the single structure level, I assessed my approach using both synthetic and experimental images. Because my approach uses all available information, it should require fewer scattering

images for a given resolution than other approaches. For the tests on synthetic images, I therefore selected the same protein crambin as was used to validate the previous three-photon correlation approach [36], and, indeed, two times fewer images were required to achieve the same resolution of about 4 Å for crambin than were required using three-photon correlations. Whereas this improvement is quite helpful, it is a bit less spectacular than expected, which suggests that the three-photon correlation already contains much of the full available information. However, the improvement for single structures may in the long run turn out to be larger than it appears from this result. Indeed, the posterior probability of my reconstructed density was still substantially lower than the reference value, suggesting than an even higher resolution could have been achieved with better sampling, using the same amount of images.

For the test on experimental images, the improvement over established methods was much larger. Here my approach successfully recovered the icosahedral structure of the coliphage PR772 virus [81], achieving the same detector-limited resolution of 9 nm as reported using the established diffusion map algorithm [62] using only a tiny fraction of the available data. Indeed, only 10% of the available images were used, which were further downsampled by a factor of 10^4, such that in total only 0.001 % of the available photons were used. Considering that the information content is quadratic in the number of photons, this means that only one billionth of the available information was used. Despite the restriction to this tiny fraction of the data, I also did not need to impose any icosahedral symmetry, in contrast to Hosseinizadeh et al. [62].

This coliphage data set has been analyzed using many other approaches [36, 63, 77, 180–182], some of which also did not impose the icosahedral symmetry or also used downsampled images, but I am not aware of any that achieved this resolution using such a small fraction of the data. Even Ayyer et al., who applied the EMC algorithm in what they termed the 'low signal limit', used about ten times the number of photons required by my approach [63]. Although they report average photon counts per downsampled image that may at first sight seem similar to mine, their photon counts are somewhat misleading. Indeed, their values refer to the number of photons outside of the central speckle [63], which is substantially lower than the full number of photons per image. Specifically, they report an average of 33.9 photons outside the central speckle for their most strongly downsampled images, whereas the downsampled images used here contained on average only about 3 photons outside the central speckle. In terms of photon counts including the central speckle, I estimate that their images contained on average about 400 photons per image, as opposed to the average of 40 photons per image in this work.

On the structural ensemble level, I tested my approach using synthetic scattering images generated from molecular dynamics trajectories. Here my approach was able to determine the conformational ensembles of alanine dipeptide, and the unfolded ensemble of the the mini-protein chignolin. These ensembles were represented by discrete weighted sets, consisting of eight conformers for alanine dipeptide and six for chignolin. In contrast to the single-structure level, no obvious comparison to previous methods is possible, because, to my knowledge, ensemble determinations have so far not been successfully attempted from such sparse scattering images.

Still, much fewer scattering images were required to determine these ensembles than expected. To understand this, I analyzed the scaling behavior of the required number of images in the number

of conformers and, for comparison, in the size of one single structure. Because an ensemble of n conformers consisting of m degrees of freedom each has the same total number of degrees of freedom as a single structure of $n \times m$ degrees of freedom, one might expect that a similar number of images should be required. My analysis revealed not only that is this is not the case, but also that the scaling behavior in both scenarios is very different. In fact, I found that to determine an ensemble of n conformers, $O(n^2)$ images are required, for which I also found a theoretical argument using the Fisher information of the scattering images. In contrast, about $O(m^5)$ images seem to be required for a single structure consisting of m Gaussian beads. Here I did not find a theoretical explanation, and it is, in fact, not clear that this relationship should indeed be given a power law. Still, these results clearly demonstrated that structural ensembles require fewer images than single structures of the same complexity.

Having validated my approach both on the single structure and the ensemble level, I next sought to assess how well it performs in the presence of noise and other complicating experimental effects. Whereas the already mentioned applications to experimental data demonstrate some resilience to these effects, the noise level in these images is actually rather low, due to preprocessing steps like background subtraction or hit selection which are not possible for smaller specimen. I therefore also applied my method to synthetic noisy images, for which, as already mentioned above, I accounted for many experimental effects such as incoherent scattering, background scattering, beam polarization, irregular detector shapes, hits and misses, and intensity fluctuations. Here, I was able to reconstruct the electron density of crambin from low images with a very low hit rate (about 2 %) and a very low signal to noise ratio (also about 2 %), albeit so far only at a lower resolution of 10.6 Å. Notably, I also showed that in this scenario, neither hit selection nor orientation determination are possible for each image, such that approaches that rely on these would not be applicable.

Even though I did not explicitly test it, structural ensemble determination should also be possible in this low signal-to-noise low hit rate scenario, of course requiring more images than from noise-free images. Further, the quadratic scaling in the number of conformers already mentioned above should also hold in this scenario. Despite this, I was so far unable to determine the conformational landscape of the coliphage PR772 virus as reported by Hosseinizadeh et al. [62] using my approach. A possible reason for this is that the downsampled images may not contain enough information.

I have so far not applied my approach directly to the full not-downsampled images from this data set, because the current implementation of my method is only optimized for sparse scattering images with smaller photon counts. In fact, problems due to floating point precision have so far prevented a direct application to images with more than about 200 to 500 photons. Although these could be addressed either by switching from single to double precision or by working with logarithmic values, this would come at a prohibitive computational cost, particularly on GPUs. Still, my approach in itself is also applicable to scattering images with higher photon counts, and it should be possible to solve these problems using a separate implementation for dense scattering images.

Another possible reason may be that my forward model, while already quite realistic, still has to be calibrated and expanded. Importantly, the implementation is flexible and general enough to

allow for this calibration. For example, while so far a simple Gaussian distribution was used to account for background scattering, a more accurate distribution for background scattering can be straightforwardly included. As an alternative, in case this distribution or other parameters are not known in advance, the Bayesian framework also allows to infer them from the data along with the electron densities. The Bayesian framework should also make the inclusion of additional experimental effects fairly straightforward. However, this might come at additional computational cost or require tedious work, as is the case for example for detector noise or other complicated detector effects.

The most challenging effect yet to take into account is the scattering on the solvation shell around the molecule. While experimental techniques could be used to keep the amount of water minimal, this would be somewhat counter-productive, as the ultimate goal is to observe the protein in solution. Therefore, the water droplet included the solvation shell has to be included in the analysis. The challenge posed by the solvation shell lies in the fact that it is structure-dependent and neither fully ordered or fully disordered. While fully ordered water would simply appear in the determined electron density and fully disordered water would appear like background scattering, partially ordered water is much more tricky to handle. To include it, a model for the distribution of this partially ordered water and for its scattering distribution given the used electron density representation is needed. Here, a combination of molecular dynamics software and thorough calibration with experimental data will be instrumental.

Notably, a similar kind of partial disorder is present in the molecular dynamics trajectories used to generate the scattering images. The scattering images were generated from an effectively continuous ensemble, but the determined ensembles are discrete. This is quite similar to the effect of the solvation shell, where many configurations of the water molecules correspond to one molecular structure, suggesting that the problems posed by the solvation shell can also be solved.

After the inclusion of all necessary experimental effects and the calibration of the forward model, my approach should be able to determine structures and structural ensembles for specimen that were so far deemed inaccessible — in fact, as mentioned before, some of these new experiments are already in planning, motivated by my results.

To plan these future experiments, realistic estimates of the required number of scattering images for given resolutions will be crucial. To be able to provide estimates for the required number of images also for other scenarios than the ones explicitly considered here, and I systematically analyzed the scaling behavior with respect to the most important parameters of the forward model, the expected number of photons per image and the amount of noise. Here, I found that the information content per image is proportional to the square of the number of photons per image, and that already small amounts of noise greatly affects the required number of images. While extrapolation from this analysis suggested that atomistic resolutions may be out of reach of current experiments for small proteins like Crambin at realistic hit rates and signal-to-noise ratios, already slightly reduced resolutions of about 6 Å require much fewer images and should be achievable.

I also analyzed how the size of the specimen molecule affects the achieved resolutions. Here, I found that the number of images required for the same resolution increased slightly with the size

of the molecule, although the data are so far insufficient for a definite conclusion. In contrast, it was previously shown that the resolution achievable with an orientation determination approach is proportional to $M^{1/6}$, where M is the molecular mass. However, this refers to the resolution achievable from infinitely many images, and does therefore not contradict my finding that the achievable resolution from a given finite number of images decreases with the size of the molecule. Other previous analyses have come to the conclusion that should be independent of the size of the molecule [124, 153]. These analyses focused on larger specimen where orientation determination is possible for each image, and should therefore also be compatible with my findings, in particular considering the fact that my approach should be able to extract more information than other approaches particularly for smaller specimen, which does affect the scaling behavior.

Despite all its advantages, the Bayesian approach comes with its own challenges, the most important one being the need for sampling large solution spaces, implying excessive computational cost. In fact, a substantial computational effort was already required for the results presented in this thesis. To give specific examples, the electron density determination of crambin shown in Figure 2.2 required about one week of concurrent computation on 20 GPUs, and the scaling analysis shown in Figure 3.5 required on the order of 10,000 GPU-hours of computation. In contrast to, say, methods using photon correlations, the computational cost for each likelihood computation increases with the number of images, both in terms of the number of arithmetic operations and the memory requirements. The most important factor for the computational cost is the relative resolution, that is, the ratio between the size of the molecule and the resolution. Because both the required number of images and the required number of Monte-Carlo steps (and therefore likelihood computations) grow quickly for higher relative resolutions, the computational cost quickly becomes the main bottleneck.

Notably, this also means that the computational cost is much lower for reduced relative resolutions, which require substantially fewer images and fewer Monte-Carlo steps. To again give a specific example, for the third resolution stage in 2.2, only a about one hour of computation on a single GPU was required. In fact, for the expected first applications to experiments, already such reduced resolutions will likely be sufficient.

The fact that ‘only’ sampling is prohibitive can also be considered an advantage, because many efficient sampling methods have been developed, which have not been explored in this work. While my specialized hierarchical simulated annealing scheme has already greatly enhanced the sampling efficiency, compared to standard simulated annealing by a factor at least in the hundreds or thousands, other more sophisticated optimization or sampling algorithms may further increase the performance.

Promising here are, for example, alternating projection algorithms which have been very successful for phase reconstruction [57], and have already been successfully applied using photon correlations [77, 78]. My method could also easily be modified into an iterative updating scheme similar to the EMC algorithm [49]. Similarly, it may be possible to utilize gradient information. While simple gradient descent is ill-suited to this particular problem, due to the highly non-convex target function, other, more sophisticated methods may be better suited, like the Metropolis adjusted Langevin algorithm [183]. The Metropolis sampling may also be enhanced by utilizing the

three-photon correlation function [36] or other summary statistics for better proposals, for instance by a delayed acceptance scheme [184], or by an improved sampling technique like replica exchange (also known as parallel tempering) [185]. The latter may be particularly useful in combination with a modified version of my hierarchical sampling method, as the higher temperature replicas could also be sampled at a lower-resolution. I have already begun investigating some of these methods, like delayed acceptance and replica exchange, but so far have not reached a definite conclusion which will be the ideal choice.

As an alternative or addition to algorithmic improvements, it may be possible to include additional prior information. By including this prior information, the space of potential structures would become smaller, such that fewer scattering images would be required. This information may come from other experiments, such as the established structure determination techniques (crystallography, NMR and cryo-EM), from structural databases like the PDB [130] or from molecular dynamics software like GROMACS [127]. Particularly molecular dynamics force fields should greatly reduce the size of the solution space, but would also make the solution space much more non-convex, introducing an additional energy landscape with many barriers and local minima, thus posing a substantial additional sampling challenge. Still, including molecular dynamics as a prior may be a worthwhile approach to investigate, potentially in a simplified form better suited to sampling, for example excluding electrostatic interactions. It may also be possible to guide a molecular dynamics simulations by the likelihood function, that is, introduce it as an additional energy term in the simulation.

Considering the ever increasing efficacy of artificial intelligence, this prior information may also come from structure prediction methods like AlphaFold or RoseTTAFold. While these methods have so far focused on the single structure level, they may nevertheless be quite useful as prior information on structural ensembles. Important here is that my method and its implementation do not rely on a particular structure representation, such that is could flexibly be applied to those used by these structure prediction methods.

Finally, while Bayesian approaches are particularly known for their excessive computational cost, the established approaches are also not computationally cheap either. For instance, both the EMC algorithm as well as manifold embedding methods require exceedingly costly computations when applied to sets of many images [49, 66]. Even the approaches based on photon correlations are limited by this, despite the fact that their computational cost does not increase with the number of images [36]. Notably, the three-photon correlation approach may also be improved by the same improved sampling of optimization methods as discussed above, showing that the development of such methods is of more general use.

For all results presented in this thesis the same structure representation consisting of Gaussian beads with variable positions was used, because in this way relatively few degrees of freedom are required to represent protein structures. Further, specimen of arbitrary size can be represented using such Gaussians, which can loosely correspond to amino acids, subdomains of larger proteins, or even larger structures in a virus. In the extreme, one Gaussian bead could also represent a single atom, but the much higher number of degrees of freedom would lead to highly increased

required numbers of images such that this is likely only useful in combination with a molecular dynamics prior as mentioned above.

However, representing electron densities using a small number of degrees of freedom can also mean that the true solution is not representable or is too far away from the closest representable density. Therefore, it may be worthwhile to consider alternatives. To give a specific example, a representation consisting of radial shells and spherical harmonics in Fourier space as used for example by the three-photon correlation method [36] may be worth considering. Similarly, simply using voxels in Fourier space may provide good results. This would also allow for a simpler version of the hierarchical sampling used here, by cooling down the innermost shells before the outer shells. As a downside, however, an additional phase retrieval step would be required.

Also on the structural ensemble level alternative representations will be worth investigating. While representing structural ensembles by sets of discrete conformers has so far been adequate, it may eventually turn out to be a limiting factor. Continuous representations as have been proposed for cryo-EM [19] are much more flexible, and should, for instance, be better suited to represent pathways of conformational changes or disordered ensembles. Alternatively, it may be possible to develop a regularization scheme that includes the discrepancy between continuous and discrete ensembles in the Bayesian formalism. For instance, by including an integral over a small random perturbation of the electron density in the likelihood function, the posterior could be biased towards discrete ensembles that coincide with the centers of conformational states (local maxima) in the continuous ensemble.

From a more general viewpoint, the Bayesian approach developed here may also be very useful to derive structures from other kinds of experimental data, including, but not limited to, other scattering experiments. To give a particular example, cryo-EM is, from a mathematical perspective, very similar to single-molecule scattering, also involving random particle orientation and unknown conformations. The techniques developed here, including the hierarchical sampling method as well as the tools used to compute the likelihoods should be applicable also to cryo-EM in a very similar form. In fact, Bayesian approaches have been proposed for cryo-EM that are formally very similar to my method [92].

An analogous Bayesian approach may also be applied to fluctuation X-ray scattering [186]. Here, instead of on a single particle, the XFEL pulses are scattered on droplets containing hundreds or thousands of randomly oriented molecules. The resulting images are typically analyzed using photon correlations [186]. A Bayesian algorithm as here would involve a costly integral over all of the independent orientations, which may however still be possible to compute if efficient approximations can be found.

Overall, my results strongly suggest that the de novo determination of structural ensembles for proteins and other biomolecules is within reach of state-of-the-art single-molecule X-ray scattering experiments, which will be an important step towards the ultimate goal of time-resolved structures. The most important next step will be do demonstrate my approach on more experimental data, likely first for larger test specimen before the most challenging case of small proteins. Further, my method may also be of more general use, with possible applications in cryo-EM or other scattering experiments.

5.2 Time-lagged independent component analysis

In Chapter 4, I analyzed time-lagged independent component analysis (tICA), a dimension reduction algorithm commonly applied to molecular dynamics trajectories. Specifically, I sought to understand the tICA-projections of high dimensional random walks, and to compare these projections to the tICA-projections of molecular dynamics trajectories. By a combination of analytical and numerical means, I was able to determine the expected tICA-projections of high dimensional random walks. Just like the PCA-projections of random walks [111, 112], the tICA-projections turned out to resemble cosine functions; however, in contrast to PCA, for tICA the projections show a much more complex behavior and a much richer mathematical structure.

I was unfortunately not able to obtain a fully analytical expression for the expected random walk tICA-projections, only one that is restricted to a special case. It will be very interesting to see if such an expression or even a fully rigorous proof can be obtained in the future. For this it may also be worth considering to analyze tICA outside of the linear algebra picture.

The tICA-projections of short molecular dynamics trajectories for large proteins did indeed turn out to be very similar to those or random walks. Remarkably, the complex patterns found for the random walk projections near identically appear for the molecular dynamics trajectories. This suggests that not only the ensemble properties of finite protein trajectories resemble those of random walks, but also their time correlations and dynamical properties.

My results also provide a measure for the convergence of a molecular dynamics trajectory based on tICA; in fact, they show that the cosine-content [112], which was originally devised for PCA, can also be used in combination with tICA. Because the additional lag-time parameter provides a much richer structure, the cosine content may in fact be a more sensitive measure for convergence when applied in combination with tICA. It should also be possible to devise a more specialized way to measure the similarity to random walk tICA-projections across all lag-times, which may result in even higher sensitivity. Here, it should also be taken into account that the tICA-projection strongly depend on the dimensionality of the random walk.

This relationship between the dimensionality of the random walk and its tICA-projections also means that my analysis offers a way to estimate an effective dimensionality of a molecular dynamics trajectory, by determining the dimension of the random walk which maximizes the similarity of the tICA-projections. It will be very interesting to further understand this and compare this with other attempts to estimate such an effective dimensionality [187, 188]. This idea, as well as precisely how this 'effective dimensionality' can be defined, clearly deserves further exploration. It will also be worth investigating the tICA-projections of random walks on other spaces, for example on high-dimensional cubes or tori. A random walk on such a hypercube would more closely mimic the behavior of proteins, such that this may offer an even better measure of convergence or a more accurate effective dimensionality.

In summary, my analysis revealed that the tICA-projections of protein trajectories can be strikingly similar to those of high-dimensional random walks, and thereby suggest a new measure of convergence for such trajectories. Finally, my results also offer a new way to estimate an effective dimensionality of protein dynamics.

Bibliography

1. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **254,** 1598–1603. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1749933 (Dec. 1991).
2. Karplus, M. & Kuriyan, J. Molecular Dynamics and Protein Function. *Proceedings of the National Academy of Sciences* **102,** 6679–6685. DOI: 10.1073/pnas.0408930102 (May 10, 2005).
3. Levinthal, C. How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House. *Monticello, Illinois (Debrunnder JTP, Munck E.., eds.),* 22–24 (1969).
4. Zwanzig, R., Szabo, A. & Bagchi, B. Levinthal's Paradox. *Proceedings of the National Academy of Sciences* **89,** 20–22. DOI: 10.1073/pnas.89.1.20 (Jan. 1992).
5. Benjin, X. & Ling, L. Developments, Applications, and Prospects of Cryo-electron Microscopy. *Protein Science : A Publication of the Protein Society* **29,** 872–882. ISSN: 0961-8368. DOI: 10.1002/pro.3805 (Apr. 2020).
6. Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. & Valencia, A. The Structural Coverage of the Human Proteome before and after AlphaFold. *PLoS Computational Biology* **18,** e1009818. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1009818 (Jan. 24, 2022).
7. McPherson, A. & Gavira, J. A. Introduction to Protein Crystallization. *Acta Crystallographica Section F: Structural Biology Communications* **70,** 2–20. ISSN: 2053-230X. DOI: 10.1107/S2053230X13033141 (1 Jan. 1, 2014).
8. Gauto, D. F. *et al.* Integrated NMR and Cryo-EM Atomic-Resolution Structure Determination of a Half-Megadalton Enzyme Complex. *Nature Communications* **10,** 2697. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10490-9 (1 June 19, 2019).
9. Hu, Y. *et al.* NMR-Based Methods for Protein Analysis. *Analytical Chemistry* **93,** 1866–1879. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.0c03830 (Feb. 2, 2021).
10. Huang, C. & Kalodimos, C. G. Structures of Large Protein Complexes Determined by Nuclear Magnetic Resonance Spectroscopy. *Annual Review of Biophysics* **46,** 317–336. DOI: 10.1146/annurev-biophys-070816-033701 (2017).
11. Jiang, Y. & Kalodimos, C. G. NMR Studies of Large Proteins. *Journal of Molecular Biology. John Kendrew's 100th Anniversary Special Edition* **429,** 2667–2676. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2017.07.007 (Aug. 18, 2017).

12. Liu, Y., Huynh, D. T. & Yeates, T. O. A 3.8 Å Resolution Cryo-EM Structure of a Small Protein Bound to an Imaging Scaffold. *Nature Communications* **10,** 1864. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09836-0 (1 Apr. 23, 2019).

13. D'Imprima, E. & Kühlbrandt, W. Current Limitations to High-Resolution Structure Determination by Single-Particle cryoEM. *Quarterly Reviews of Biophysics* **54,** e4. ISSN: 0033-5835, 1469-8994. DOI: 10.1017/S0033583521000020 (Jan. 2021).

14. Bendory, T., Bartesaghi, A. & Singer, A. Single-Particle Cryo-Electron Microscopy: Mathematical Theory, Computational Challenges, and Opportunities. *IEEE signal processing magazine* **37,** 58–76. ISSN: 1053-5888. DOI: 10.1109/msp.2019.2957822 (Mar. 2020).

15. Bock, L. V. & Grubmüller, H. Effects of Cryo-EM Cooling on Structural Ensembles. *Nature Communications* **13,** 1709. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29332-2 (1 Mar. 31, 2022).

16. Abbamonte, P. *et al. New Science Opportunities Enabled by LCLS-II X-Ray Lasers* SLAC-R-1053 (SLAC National Accelerator Lab., Menlo Park, CA (United States), June 1, 2015). DOI: 10.2172/1630267.

17. Bock, L. V. *et al.* Energy Barriers and Driving Forces in tRNA Translocation through the Ribosome. *Nature Structural & Molecular Biology* **20,** 1390–1396. ISSN: 1545-9985. DOI: 10.1038/nsmb.2690 (12 Dec. 2013).

18. Bonomi, M., Heller, G. T., Camilloni, C. & Vendruscolo, M. Principles of Protein Structural Ensemble Determination. *Current Opinion in Structural Biology. Folding and Binding • Proteins: Bridging Theory and Experiment* **42,** 106–116. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2016.12.004 (Feb. 1, 2017).

19. Kinman, L. F., Powell, B. M., Zhong, E. D., Berger, B. & Davis, J. H. Uncovering Structural Ensembles from Single-Particle Cryo-EM Data Using cryoDRGN. *Nature Protocols,* 1–31. ISSN: 1750-2799. DOI: 10.1038/s41596-022-00763-x (Nov. 14, 2022).

20. Rosenbaum, D. *et al. Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images Using VAEs* http://arxiv.org/abs/2106.14108. preprint.

21. Jumper, J. *et al.* Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **596,** 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2 (7873 Aug. 2021).

22. Baek, M. *et al.* Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **373,** 871–876. DOI: 10.1126/science.abj8754 (Aug. 20, 2021).

23. Bertoline, L. M. F., Lima, A. N., Krieger, J. E. & Teixeira, S. K. Before and after AlphaFold2: An Overview of Protein Structure Prediction. *Frontiers in Bioinformatics* **3,** 1120370. ISSN: 2673-7647. DOI: 10.3389/fbinf.2023.1120370 (Feb. 28, 2023).

24. Ourmazd, A., Moffat, K. & Lattman, E. E. Structural Biology Is Solved — Now What? *Nature Methods* **19,** 24–26. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01357-3 (1 Jan. 2022).

25. Lane, T. J. Protein Structure Prediction Has Reached the Single-Structure Frontier. *Nature Methods* **20,** 170–173. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01760-4 (2 Feb. 2023).

26. Chapman, H. N., Caleman, C. & Timneanu, N. Diffraction before Destruction. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369,** 20130313. DOI: 10.1098/rstb.2013.0313 (July 17, 2014).

27. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. Potential for Biomolecular Imaging with Femtosecond X-ray Pulses. *Nature* **406,** 752–757. ISSN: 1476-4687. DOI: 10.1038/35021099 (6797 Aug. 2000).

28. Schlichting, I. Serial Femtosecond Crystallography: The First Five Years. *IUCrJ* **2,** 246–255. ISSN: 2052-2525. DOI: 10.1107/S205225251402702X (2 Mar. 1, 2015).

29. Chapman, H. N. Structure Determination Using X-Ray Free-Electron Laser Pulses. *Methods in Molecular Biology (Clifton, N.J.)* **1607,** 295–324. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-7000-1_12 (2017).

30. Oda, K. *et al.* Time-Resolved Serial Femtosecond Crystallography Reveals Early Structural Changes in Channelrhodopsin. *eLife* **10** (eds Kruse, A. C., Swartz, K. J. & Hekstra, D.) e62389. ISSN: 2050-084X. DOI: 10.7554/eLife.62389 (Mar. 23, 2021).

31. Tenboer, J. *et al.* Time-Resolved Serial Crystallography Captures High-Resolution Intermediates of Photoactive Yellow Protein. *Science (New York, N.Y.)* **346,** 1242–1246. ISSN: 1095-9203. DOI: 10.1126/science.1259357 (Dec. 5, 2014).

32. Barends, T. R. M., Stauch, B., Cherezov, V. & Schlichting, I. Serial Femtosecond Crystallography. *Nature Reviews Methods Primers* **2,** 1–24. ISSN: 2662-8449. DOI: 10.1038/s43586-022-00141-7 (1 Aug. 4, 2022).

33. Zastrau, U. *et al.* The High Energy Density Scientific Instrument at the European XFEL. *Journal of Synchrotron Radiation* **28,** 1393–1416. ISSN: 1600-5775. DOI: 10.1107/S1600577521007335 (Sept. 1, 2021).

34. Sun, Z., Fan, J., Li, H. & Jiang, H. Current Status of Single Particle Imaging with X-ray Lasers. *Applied Sciences* **8,** 132. ISSN: 2076-3417. DOI: 10.3390/app8010132 (1 Jan. 2018).

35. Gaffney, K. J. & Chapman, H. N. Imaging Atomic Structure and Dynamics with Ultrafast X-ray Scattering. *Science* **316,** 1444–1448. ISSN: 0036-8075. DOI: 10.1126/science.1135923 (2007).

36. Von Ardenne, B., Mechelke, M. & Grubmüller, H. Structure Determination from Single Molecule X-ray Scattering with Three Photons per Image. *Nature Communications* **9,** 2375. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04830-4 (1 June 18, 2018).

37. Ourmazd, A. Cryo-EM, XFELs and the Structure Conundrum in Structural Biology. *Nature Methods* **16,** 941–944. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0587-4 (10 Oct. 2019).

38. Van Thor, J. J. Advances and Opportunities in Ultrafast X-ray Crystallography and Ultrafast Structural Optical Crystallography of Nuclear and Electronic Protein Dynamics. *Structural Dynamics* **6,** 050901. DOI: 10.1063/1.5110685 (Sept. 2019).

39. Chapman, H. N. *et al.* Femtosecond Diffractive Imaging with a Soft-X-ray Free-Electron Laser. *Nature Physics* **2,** 839–843. ISSN: 1745-2481. DOI: 10.1038/nphys461 (12 Dec. 2006).

40. Chapman, H. N. *et al.* Femtosecond X-ray Protein Nanocrystallography. *Nature* **470,** 73–77. ISSN: 1476-4687. DOI: 10.1038/nature09750 (7332 Feb. 2011).

41. Boutet, S. *et al.* High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography. *Science*. DOI: 10.1126/science.1217737 (July 20, 2012).

42. Fromme, P. & Spence, J. C. Femtosecond Nanocrystallography Using X-ray Lasers for Membrane Protein Structure Determination. *Current Opinion in Structural Biology. Engineering and Design / Membranes* **21,** 509–516. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2011.06.001 (Aug. 1, 2011).

43. Kirian, R. A. *et al.* Femtosecond Protein Nanocrystallography—Data Analysis Methods. *Optics Express* **18,** 5713–5723. ISSN: 1094-4087. DOI: 10.1364/OE.18.005713 (Mar. 15, 2010).

44. Roedig, P. *et al.* High-Speed Fixed-Target Serial Virus Crystallography. *Nature Methods* **14,** 805–810. ISSN: 1548-7105. DOI: 10.1038/nmeth.4335 (8 Aug. 2017).

45. Seibert, M. M. *et al.* Single Mimivirus Particles Intercepted and Imaged with an X-ray Laser. *Nature* **470,** 78–81. ISSN: 1476-4687. DOI: 10.1038/nature09748 (7332 Feb. 2011).

46. Ekeberg, T. *et al.* Three-Dimensional Reconstruction of the Giant Mimivirus Particle with an X-Ray Free-Electron Laser. *Physical Review Letters* **114,** 098102. DOI: 10.1103/PhysRevLett.114.098102 (Mar. 2, 2015).

47. Hosseinizadeh, A. *et al.* High-Resolution Structure of Viruses from Random Diffraction Snapshots. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369,** 20130326. DOI: 10.1098/rstb.2013.0326 (July 17, 2014).

48. Shneerson, V. L., Ourmazd, A. & Saldin, D. K. Crystallography without Crystals. I. The Common-Line Method for Assembling a Three-Dimensional Diffraction Volume from Single-Particle Scattering. *Acta Crystallographica Section A: Foundations of Crystallography* **64,** 303–315. ISSN: 0108-7673. DOI: 10.1107/S0108767307067621 (2 Mar. 1, 2008).

49. Loh, N.-T. D. & Elser, V. Reconstruction Algorithm for Single-Particle Diffraction Imaging Experiments. *Physical Review E* **80,** 026705. DOI: 10.1103/PhysRevE.80.026705 (Aug. 24, 2009).

50. Walczak, M. & Grubmüller, H. Bayesian Orientation Estimate and Structure Information from Sparse Single-Molecule x-Ray Diffraction Images. *Physical Review E* **90,** 022714. DOI: 10.1103/PhysRevE.90.022714 (Aug. 20, 2014).

51. Kassemeyer, S. *et al.* Optimal Mapping of X-Ray Laser Diffraction Patterns into Three Dimensions Using Routing Algorithms. *Physical Review E* **88,** 042710. DOI: 10.1103/PhysRevE.88.042710 (Oct. 28, 2013).

52. Elser, V. Three-Dimensional Structure from Intensity Correlations. *New Journal of Physics* **13,** 123014. ISSN: 1367-2630. DOI: 10.1088/1367-2630/13/12/123014 (Dec. 2011).

53. Tegze, M. & Bortel, G. Atomic Structure of a Single Large Biomolecule from Diffraction Patterns of Random Orientations. *Journal of Structural Biology* **179,** 41–45. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2012.04.014 (July 1, 2012).

54. Flamant, J., Le Bihan, N., Martin, A. V. & Manton, J. H. Expansion-Maximization-Compression Algorithm with Spherical Harmonics for Single Particle Imaging with x-Ray Lasers. *Physical Review E* **93,** 053302. DOI: 10.1103/PhysRevE.93.053302 (May 11, 2016).

55. Ayyer, K., Lan, T.-Y., Elser, V. & Loh, N. D. Dragonfly: An Implementation of the Expand–Maximize–Compress Algorithm for Single-Particle Imaging. *Journal of Applied Crystallography* **49,** 1320–1335. ISSN: 1600-5767. DOI: 10.1107/S1600576716008165 (4 Aug. 1, 2016).

56. Elser, V., Rankenburg, I. & Thibault, P. Searching with Iterated Maps. *Proceedings of the National Academy of Sciences* **104,** 418–423. DOI: 10.1073/pnas.0606359104 (Jan. 9, 2007).

57. Luke, D. R. Relaxed Averaged Alternating Reflections for Diffraction Imaging. *Inverse Problems* **21,** 37–50. ISSN: 0266-5611. DOI: 10.1088/0266-5611/21/1/004 (Nov. 2004).

58. Fienup, J. R. Phase Retrieval Algorithms: A Comparison. *Applied Optics* **21,** 2758–2769. ISSN: 2155-3165. DOI: 10.1364/AO.21.002758 (Aug. 1, 1982).

59. Miao, J., Kirz, J. & Sayre, D. The Oversampling Phasing Method. *Acta Crystallographica Section D: Biological Crystallography* **56,** 1312–1315. ISSN: 0907-4449. DOI: 10.1107/S0907444900008970 (10 Oct. 1, 2000).

60. Huldt, G., Szőke, A. & Hajdu, J. Diffraction Imaging of Single Particles and Biomolecules. *Journal of Structural Biology. Analytical Methods and Software Tools for Macromolecular Microscopy* **144,** 219–227. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2003.09.025 (Oct. 1, 2003).

61. Bortel, G., Faigel, G. & Tegze, M. Classification and Averaging of Random Orientation Single Macromolecular Diffraction Patterns at Atomic Resolution. *Journal of Structural Biology* **166,** 226–233. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2009.01.005 (May 1, 2009).

62. Hosseinizadeh, A. *et al.* Conformational Landscape of a Virus by Single-Particle X-ray Scattering. *Nature Methods* **14,** 877–881. ISSN: 1548-7105. DOI: 10.1038/nmeth.4395 (Sept. 2017).

63. Ayyer, K. *et al.* Low-Signal Limit of X-ray Single Particle Diffractive Imaging. *Optics Express* **27,** 37816–37833. ISSN: 1094-4087. DOI: 10.1364/OE.27.037816 (Dec. 23, 2019).

64. Tegze, M. & Bortel, G. Incorporating Particle Symmetry into Orientation Determination in Single-Particle Imaging. *Acta Crystallographica Section A: Foundations and Advances* **74,** 512–517. ISSN: 2053-2733. DOI: 10.1107/S2053273318008999 (5 Sept. 1, 2018).

65. Donatelli, J. J., Sethian, J. A. & Zwart, P. H. Reconstruction from Limited Single-Particle Diffraction Data via Simultaneous Determination of State, Orientation, Intensity, and Phase. *Proceedings of the National Academy of Sciences* **114,** 7222–7227. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1708217114 (July 11, 2017).

66. Fung, R., Shneerson, V., Saldin, D. K. & Ourmazd, A. Structure from Fleeting Illumination of Faint Spinning Objects in Flight. *Nature Physics* **5,** 64–67. ISSN: 1745-2481. DOI: 10.1038/nphys1129 (1 Jan. 2009).

67. Schwander, P., Giannakis, D., Yoon, C. H. & Ourmazd, A. The Symmetries of Image Formation by Scattering. II. Applications. *Optics Express* **20,** 12827–12849. ISSN: 1094-4087. DOI: 10.1364/OE.20.012827 (June 4, 2012).

68. Giannakis, D., Schwander, P. & Ourmazd, A. The Symmetries of Image Formation by Scattering. I. Theoretical Framework. *Optics Express* **20,** 12799–12826. ISSN: 1094-4087. DOI: 10.1364/OE.20.012799 (June 4, 2012).

69. Winter, M., Saalmann, U. & Rost, J. M. Enhancing Scattering Images for Orientation Recovery with Diffusion Map. *Optics Express* **24,** 3672–3683. ISSN: 1094-4087. DOI: 10.1364/OE.24.003672 (Feb. 22, 2016).

70. Moths, B. & Ourmazd, A. Bayesian Algorithms for Recovering Structure from Single-Particle Diffraction Snapshots of Unknown Orientation: A Comparison. *Acta Crystallographica Section A: Foundations of Crystallography* **67,** 481–486. ISSN: 0108-7673. DOI: 10.1107/S0108767311019611 (5 Sept. 1, 2011).

71. Saldin, D. K., Shneerson, V. L., Fung, R. & Ourmazd, A. Structure of Isolated Biomolecules Obtained from Ultrashort X-Ray Pulses: Exploiting the Symmetry of Random Orientations. *Journal of Physics: Condensed Matter* **21,** 134014. ISSN: 0953-8984. DOI: 10.1088/0953-8984/21/13/134014 (Mar. 2009).

72. Saldin, D. K., Poon, H.-C., Schwander, P., Uddin, M. & Schmidt, M. Reconstructing an Icosahedral Virus from Single-Particle Diffraction Experiments. *Optics Express* **19,** 17318–17335. ISSN: 1094-4087. DOI: 10.1364/OE.19.017318 (Aug. 29, 2011).

73. Saldin, D. K. *et al.* Beyond Small-Angle x-Ray Scattering: Exploiting Angular Correlations. *Physical Review B* **81,** 174105. DOI: 10.1103/PhysRevB.81.174105 (May 7, 2010).

74. Saldin, D. K. *et al.* New Light on Disordered Ensembles: Ab Initio Structure Determination of One Particle from Scattering Fluctuations of Many Copies. *Physical Review Letters* **106,** 115501. DOI: 10.1103/PhysRevLett.106.115501 (Mar. 14, 2011).

75. Saldin, D. K. *et al.* Structure of a Single Particle from Scattering by Many Particles Randomly Oriented about an Axis: Toward Structure Solution without Crystallization? *New Journal of Physics* **12,** 035014. ISSN: 1367-2630. DOI: 10.1088/1367-2630/12/3/035014 (Mar. 2010).

76. Starodub, D. *et al.* Single-Particle Structure Determination by Correlations of Snapshot X-ray Diffraction Patterns. *Nature Communications* **3,** 1276. ISSN: 2041-1723. DOI: 10.1038/ncomms2288 (1 Dec. 11, 2012).

77. Kurta, R. P. *et al.* Correlations in Scattered X-Ray Laser Pulses Reveal Nanoscale Structural Features of Viruses. *Physical Review Letters* **119,** 158102. DOI: 10.1103/PhysRevLett.119.158102 (Oct. 12, 2017).

78. Donatelli, J. J., Zwart, P. H. & Sethian, J. A. Iterative Phasing for Fluctuation X-ray Scattering. *Proceedings of the National Academy of Sciences* **112,** 10286–10291. DOI: 10.1073/pnas.1513738112 (Aug. 18, 2015).

79. Kam, Z. The Reconstruction of Structure from Electron Micrographs of Randomly Oriented Particles. *Journal of Theoretical Biology* **82,** 15–39. ISSN: 0022-5193. DOI: 10.1016/0022-5193(80)90088-0 (Jan. 7, 1980).

80. Zhuang, Y. *et al.* Unsupervised Learning Approaches to Characterizing Heterogeneous Samples Using X-ray Single-Particle Imaging. *IUCrJ* **9,** 204–214. ISSN: 2052-2525. DOI: 10.1107/S2052252521012707 (2 Mar. 1, 2022).

81. Reddy, H. K. N. *et al.* Coherent Soft X-ray Diffraction Imaging of Coliphage PR772 at the Linac Coherent Light Source. *Scientific Data* **4,** 170079. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.79 (1 June 27, 2017).

82. Dashti, A. *et al.* Trajectories of the Ribosome as a Brownian Nanomachine. *Proceedings of the National Academy of Sciences* **111,** 17492–17497. DOI: 10.1073/pnas.1419276111 (Dec. 9, 2014).

83. Barty, A. *et al.* Cheetah: Software for High-Throughput Reduction and Analysis of Serial Femtosecond X-ray Diffraction Data. *Journal of Applied Crystallography* **47,** 1118–1131. ISSN: 1600-5767. DOI: 10.1107/S1600576714007626 (3 June 1, 2014).

84. Andreasson, J. *et al.* Automated Identification and Classification of Single Particle Serial Femtosecond X-ray Diffraction Data. *Optics Express* **22,** 2497–2510. ISSN: 1094-4087. DOI: 10.1364/OE.22.002497 (Feb. 10, 2014).

85. Yoon, C. H. *et al.* Unsupervised Classification of Single-Particle X-ray Diffraction Snapshots by Spectral Clustering. *Optics Express* **19,** 16542–16549. ISSN: 1094-4087. DOI: 10.1364/OE.19.016542 (Aug. 15, 2011).

86. Grant, T., Rohou, A. & Grigorieff, N. cisTEM, User-Friendly Software for Single-Particle Image Processing. *eLife* **7** (ed Egelman, E. H.) e35383. ISSN: 2050-084X. DOI: 10.7554/eLife.35383 (Mar. 7, 2018).

87. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: Algorithms for Rapid Unsupervised Cryo-EM Structure Determination. *Nature Methods* **14,** 290–296. ISSN: 1548-7105. DOI: 10.1038/nmeth.4169 (3 Mar. 2017).

88. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-Based Classification of Cryo-EM Images Using FREALIGN. *Journal of Structural Biology* **183,** 377–388. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2013.07.005 (Sept. 1, 2013).

89. Scheres, S. H. W. in *Methods in Enzymology* (ed Crowther, R. A.) 125–157 (Academic Press, Jan. 1, 2016). DOI: 10.1016/bs.mie.2016.04.012.

90. Tang, G. *et al.* EMAN2: An Extensible Image Processing Suite for Electron Microscopy. *Journal of Structural Biology. Software Tools for Macromolecular Microscopy* **157,** 38–46. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2006.05.009 (Jan. 1, 2007).

91. Zivanov, J. *et al.* New Tools for Automated High-Resolution Cryo-EM Structure Determination in RELION-3. *eLife* **7** (eds Egelman, E. H. & Kuriyan, J.) e42166. ISSN: 2050-084X. DOI: 10.7554/eLife.42166 (Nov. 9, 2018).

92. Cossio, P. & Hummer, G. Bayesian Analysis of Individual Electron Microscopy Images: Towards Structures of Dynamic and Heterogeneous Biomolecular Assemblies. *Journal of Structural Biology* **184,** 427–437. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2013.10.006 (Dec. 1, 2013).

93. Bendory, T., Boumal, N., Leeb, W., Levin, E. & Singer, A. *Toward Single Particle Reconstruction without Particle Picking: Breaking the Detection Limit* http://arxiv.org/abs/1810.00226. preprint.

94. Guaita, M., Watters, S. C. & Loerch, S. Recent Advances and Current Trends in Cryo-Electron Microscopy. *Current Opinion in Structural Biology* **77,** 102484. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2022.102484 (Dec. 1, 2022).

95. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Review* **59,** 65–98. ISSN: 0036-1445. DOI: 10.1137/141000671 (Jan. 2017).

96. Jelsch, C. *et al.* Accurate Protein Crystallography at Ultra-High Resolution: Valence Electron Distribution in Crambin. *Proceedings of the National Academy of Sciences* **97,** 3171–3176. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.97.7.3171 (Mar. 28, 2000).

97. Honda, S. *et al.* Crystal Structure of a Ten-Amino Acid Protein. *Journal of the American Chemical Society* **130,** 15327–15331. ISSN: 0002-7863. DOI: 10.1021/ja8030533 (Nov. 19, 2008).

98. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Protein Folding Kinetics and Thermodynamics from Atomistic Simulation. *Proceedings of the National Academy of Sciences* **109,** 17845–17850. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1201811109 (Oct. 2012).

99. Karplus, M. & McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nature Structural Biology* **9,** 646–652. ISSN: 1545-9985. DOI: 10.1038/nsb0902-646 (9 Sept. 2002).

100. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of Folded Proteins. *Nature* **267,** 585–590. ISSN: 1476-4687. DOI: 10.1038/267585a0 (5612 June 1977).

101. Groot, B. L. de & Grubmüller, H. Water Permeation Across Biological Membranes: Mechanism and Dynamics of Aquaporin-1 and GlpF. *Science* **294,** 2353–2357. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1066115 (Dec. 14, 2001).

102. Zink, M. & Grubmüller, H. Mechanical Properties of the Icosahedral Shell of Southern Bean Mosaic Virus: A Molecular Dynamics Study. *Biophysical Journal* **96,** 1350–1363. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2008.11.028 (Feb. 18, 2009).

103. Perilla, J. R. & Schulten, K. Physical Properties of the HIV-1 Capsid from All-Atom Molecular Dynamics Simulations. *Nature Communications* **8,** 15959. ISSN: 2041-1723. DOI: 10.1038/ncomms15959 (1 July 19, 2017).

104. Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins: Structure, Function, and Bioinformatics* **17,** 412–425. ISSN: 1097-0134. DOI: 10.1002/prot.340170408 (1993).

105. Molgedey, L. & Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Physical Review Letters* **72,** 3634–3637. ISSN: 0031-9007. DOI: 10.1103/physrevlett.72.3634 (1994).

106. Naritomi, Y. & Fuchigami, S. Slow Dynamics of a Protein Backbone in Molecular Dynamics Simulation Revealed by Time-Structure Based Independent Component Analysis. *The Journal of Chemical Physics* **139,** 215102. ISSN: 0021-9606. DOI: 10.1063/1.4834695 (2013).

107. Pérez-Hernández, G., Paul, F., Giorgino, T., Fabritiis, G. D. & Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *The Journal of Chemical Physics* **139,** 015102. ISSN: 0021-9606. DOI: 10.1063/1.4811489 (2013).

108. Schwantes, C. R. & Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *Journal of Chemical Theory and Computation* **9,** 2000–2009. ISSN: 1549-9618. DOI: 10.1021/ct300878a (2013).

109. De Groot, B. L., Daura, X., Mark, A. E. & Grubmüller, H. Essential Dynamics of Reversible Peptide Folding: Memory-Free Conformational Dynamics Governed by Internal Hydrogen bonds11Edited by R. Huber. *Journal of Molecular Biology* **309,** 299–313. ISSN: 0022-2836. DOI: 10.1006/jmbi.2001.4655 (May 25, 2001).

110. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* **140,** 2386–2396. ISSN: 0002-7863. DOI: 10.1021/jacs.7b12191 (Feb. 2018).

111. Hess, B. Similarities between Principal Components of Protein Dynamics and Random Diffusion. *Physical Review E* **62,** 8438–8448. ISSN: 1539-3755. DOI: 10.1103/physreve.62.8438 (2000).

112. Hess, B. Convergence of Sampling in Protein Simulations. *Physical Review E* **65,** 031910. ISSN: 1539-3755. DOI: 10.1103/physreve.65.031910 (2002).

113. Olsson, S. & Noé, F. Mechanistic Models of Chemical Exchange Induced Relaxation in Protein NMR. *Journal of the American Chemical Society* **139,** 200–210. ISSN: 0002-7863. DOI: 10.1021/jacs.6b09460 (2016).

114. Xiao, J. & Salsbury, F. R. Na + -Binding Modes Involved in Thrombin's Allosteric Response as Revealed by Molecular Dynamics Simulations, Correlation Networks and Markov Modeling. *Physical Chemistry Chemical Physics* **21,** 4320–4330. ISSN: 1463-9076. DOI: 10.1039/c8cp07293k (2019).

115. Schultze, S. & Grubmüller, H. *De Novo Structural Ensemble Determination from Single-Molecule X-ray Scattering: A Bayesian Approach* http://arxiv.org/abs/2302.09136. preprint.

116. Hajdu, J. Single-Molecule X-ray Diffraction. *Current Opinion in Structural Biology* **10,** 569–573. ISSN: 0959-440X. DOI: 10.1016/S0959-440X(00)00133-0 (Oct. 1, 2000).

117. Miao, J., Ishikawa, T., Robinson, I. K. & Murnane, M. M. Beyond Crystallography: Diffractive Imaging Using Coherent x-Ray Light Sources. *Science* **348,** 530–535. ISSN: 0036-8075. DOI: 10.1126/science.aaa1394 (2015).

118. Yoon, C. H. *et al.* A Comprehensive Simulation Framework for Imaging Single Particles and Biomolecules at the European X-ray Free-Electron Laser. *Scientific Reports* **6,** 24791. ISSN: 2045-2322. DOI: 10.1038/srep24791 (1 Apr. 25, 2016).

119. Hantke, M. F., Ekeberg, T. & Maia, F. R. N. C. Condor: A Simulation Tool for Flash X-ray Imaging. *Journal of Applied Crystallography* **49,** 1356–1362. ISSN: 1600-5767. DOI: 10.1107/S1600576716009213 (4 Aug. 1, 2016).

120. Van Heel, M. & Schatz, M. Fourier Shell Correlation Threshold Criteria. *Journal of Structural Biology* **151,** 250–262. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2005.05.009 (Sept. 1, 2005).

121. Cuturi, M. *Sinkhorn Distances: Lightspeed Computation of Optimal Transport* in *Advances in Neural Information Processing Systems* **26** (Curran Associates, Inc., 2013).

122. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation* **11,** 5525–5542. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00743 (2015).

123. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *Journal of Molecular Biology* **7,** 95–99. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(63)80023-6 (July 1963).

124. Poudyal, I., Schmidt, M. & Schwander, P. Single-Particle Imaging by x-Ray Free-Electron Lasers—How Many Snapshots Are Needed? *Structural Dynamics* **7,** 024102. ISSN: 2329-7778. DOI: 10.1063/1.5144516 (Mar. 20, 2020).

125. *SPHERE_LEBEDEV_RULE - Quadrature Rules for the Sphere* https://people.sc.fsu.edu/~jburkardt/datasets/sphere_lebedev_rule/sphere_lebedev_rule.html.

126. Gräf, M. & Potts, D. Sampling Sets and Quadrature Formulae on the Rotation Group. *Numerical Functional Analysis and Optimization* **30,** 665–688. ISSN: 0163-0563. DOI: 10.1080/01630560903163508 (2009).

127. Abraham, M. J. *et al.* GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **1–2,** 19–25. ISSN: 2352-7110. DOI: 10.1016/j.softx.2015.06.001 (Sept. 1, 2015).

128. Huang, J. *et al.* CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nature Methods* **14,** 71–73. ISSN: 1548-7105. DOI: 10.1038/nmeth.4067 (1 Jan. 2017).

129. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *The Journal of Physical Chemistry Letters* **5,** 3863–3871. ISSN: 1948-7185. DOI: 10.1021/jz501780a (Nov. 6, 2014).

130. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235 (Jan. 1, 2000).

131. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *Journal of Computational Chemistry* **20,** 786–798. ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(199906)20:8<786::AID-JCC5>3.0.CO;2-B (1999).

132. Bussi, G., Donadio, D. & Parrinello, M. Canonical Sampling through Velocity Rescaling. *The Journal of Chemical Physics* **126,** 014101. ISSN: 0021-9606. DOI: 10.1063/1.2408420 (Jan. 3, 2007).

133. Parrinello, M. & Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *Journal of Applied Physics* **52,** 7182–7190. ISSN: 0021-8979. DOI: 10.1063/1.328693 (Dec. 1, 1981).

134. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *Journal of Computational Chemistry* **18,** 1463–1472. ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H (1997).

135. Miyamoto, S. & Kollman, P. A. Settle: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *Journal of Computational Chemistry* **13,** 952–962. ISSN: 1096-987X. DOI: 10.1002/jcc.540130805 (1992).

136. Darden, T., York, D. & Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *The Journal of Chemical Physics* **98,** 10089–10092. ISSN: 0021-9606. DOI: 10.1063/1.464397 (June 15, 1993).

137. *Mdshare* mdshare. https://markovmodel.github.io/mdshare/.

138. Behboodian, J. Information Matrix for a Mixture of Two Normal Distributions. *Journal of Statistical Computation and Simulation* **1,** 295–314. ISSN: 0094-9655. DOI: 10.1080/00949657208810024 (Oct. 1, 1972).

139. Gelman, A., Gilks, W. R. & Roberts, G. O. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability* **7,** 110–120. ISSN: 1050-5164, 2168-8737. DOI: 10.1214/aoap/1034625254 (Feb. 1997).

140. Hubbell, J. H., Gimm, H. A. & O/verbo/, I. Pair, Triplet, and Total Atomic Cross Sections (and Mass Attenuation Coefficients) for 1 MeV-100 GeV Photons in Elements Z =1 to 100. *Journal of Physical and Chemical Reference Data* **9,** 1023–1148. ISSN: 0047-2689, 1529-7845. DOI: 10.1063/1.555629 (Oct. 1, 1980).

141. Allahgholi, A. *et al.* The Adaptive Gain Integrating Pixel Detector at the European XFEL. *Journal of Synchrotron Radiation* **26,** 74–82. ISSN: 1600-5775. DOI: 10.1107/S1600577518016077 (Jan. 1, 2019).

142. Sobolev, E. *et al.* Megahertz Single-Particle Imaging at the European XFEL. *Communications Physics* **3,** 1–11. ISSN: 2399-3650. DOI: 10.1038/s42005-020-0362-y (1 May 29, 2020).

143. Kirian, R. A. *et al.* Simple Convergent-Nozzle Aerosol Injector for Single-Particle Diffractive Imaging with X-ray Free-Electron Lasers. *Structural Dynamics (Melville, N.Y.)* **2,** 041717. ISSN: 2329-7778. DOI: 10.1063/1.4922648 (July 2015).

144. DePonte, D. P. *et al.* Gas Dynamic Virtual Nozzle for Generation of Microscopic Droplet Streams. *Journal of Physics D: Applied Physics* **41,** 195505. ISSN: 0022-3727. DOI: 10.1088/0022-3727/41/19/195505 (Sept. 2008).

145. Hoffman, D. J. *et al.* Microfluidic Liquid Sheets as Large-Area Targets for High Repetition XFELs. *Frontiers in Molecular Biosciences* **9.** ISSN: 2296-889X (2022).

146. Giacovazzo, C. *et al. Fundamentals of Crystallography* 866 pp. ISBN: 978-0-19-957365-3 (OUP Oxford, Feb. 10, 2011).

147. Yun, K. *et al.* Coherence and Pulse Duration Characterization of the PAL-XFEL in the Hard X-ray Regime. *Scientific Reports* **9,** 3300. ISSN: 2045-2322. DOI: 10.1038/s41598-019-39765-3 (1 Mar. 1, 2019).

148. Saldin, E. L., Schneidmiller, E. A. & Yurkov, M. V. Statistical Properties of Radiation from VUV and X-ray Free Electron Laser. *Optics Communications* **148,** 383–403. ISSN: 0030-4018. DOI: 10.1016/S0030-4018(97)00670-6 (Mar. 15, 1998).

149. Ayvazyan, V. *et al.* First Operation of a Free-Electron Laser Generating GW Power Radiation at 32 Nm Wavelength. *The European Physical Journal D - Atomic, Molecular, Optical and Plasma Physics* **37,** 297–303. ISSN: 1434-6079. DOI: 10.1140/epjd/e2005-00308-1 (Feb. 1, 2006).

150. Bielecki, J., Maia, F. R. N. C. & Mancuso, A. P. Perspectives on Single Particle Imaging with x Rays at the Advent of High Repetition Rate X-Ray Free Electron Laser Sources. *Structural Dynamics* **7,** 040901. ISSN: 2329-7778. DOI: 10.1063/4.0000024 (Aug. 6, 2020).

151. Wang, I. *et al.* Solution Structure of a K+-Channel Blocker from the Scorpion Tityus Cambridgei. *Protein Science* **11,** 390–400. ISSN: 1469-896X. DOI: 10.1110/ps.33402 (2002).

152. Hoffman, R. C., Klevit, R. E. & Horvath, S. J. Structures of DNA-binding Mutant Zinc Finger Domains: Implications for DNA Binding. *Protein Science* **2,** 951–965. ISSN: 1469-896X. DOI: 10.1002/pro.5560020609 (1993).

153. Nakano, M., Miyashita, O. & Tama, F. Molecular Size Dependence on Achievable Resolution from XFEL Single-Particle 3D Reconstruction. *Structural Dynamics (Melville, N.Y.)* **10,** 024101. ISSN: 2329-7778. DOI: 10.1063/4.0000175 (Mar. 2023).

154. Gorobtsov, O. Yu., Lorenz, U., Kabachnik, N. M. & Vartanyants, I. A. Theoretical Study of Electronic Damage in Single-Particle Imaging Experiments at x-Ray Free-Electron Lasers for Pulse Durations from 0.1 to 10 Fs. *Physical Review E* **91,** 062712. DOI: 10.1103/PhysRevE.91.062712 (June 15, 2015).

155. Fortmann-Grote, C. *et al.* Start-to-End Simulation of Single-Particle Imaging Using Ultra-Short Pulses at the European X-ray Free-Electron Laser. *IUCrJ* **4,** 560–568. ISSN: 2052-2525. DOI: 10.1107/S2052252517009496 (Sept. 1, 2017).

156. Hau-Riege, S. P., London, R. A. & Szoke, A. Dynamics of Biological Molecules Irradiated by Short X-Ray Pulses. *Physical Review E* **69,** 051906. DOI: 10.1103/PhysRevE.69.051906 (May 18, 2004).

157. Schultze, S. & Grubmüller, H. Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. *Journal of Chemical Theory and Computation* **17,** 5766–5776. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.1c00273 (Sept. 14, 2021).

158. Henzler-Wildman, K. & Kern, D. Dynamic Personalities of Proteins. *Nature* **450,** 964–972. ISSN: 1476-4687. DOI: 10.1038/nature06522 (7172 Dec. 2007).

159. Lewandowski, J. R., Halse, M. E., Blackledge, M. & Emsley, L. Direct Observation of Hierarchical Protein Dynamics. *Science* **348,** 578–581. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaa6111 (May 1, 2015).

160. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Structure, Function, and Bioinformatics* **21,** 167–195. ISSN: 1097-0134. DOI: 10.1002/prot.340210302 (1995).

161. Perilla, J. R. *et al.* Molecular Dynamics Simulations of Large Macromolecular Complexes. *Current Opinion in Structural Biology. Theory and Simulation/Macromolecular Machines and Assemblies* **31,** 64–74. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2015.03.007 (Apr. 1, 2015).

162. Grubmüller, H. Predicting Slow Structural Transitions in Macromolecular Systems: Conformational Flooding. *Physical Review E* **52,** 2893–2906. DOI: 10.1103/PhysRevE.52.2893 (Sept. 1, 1995).

163. Sugita, Y. & Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chemical Physics Letters* **314,** 141–151. ISSN: 0009-2614. DOI: 10.1016/S0009-2614(99)01123-9 (Nov. 26, 1999).

164. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *WIREs Computational Molecular Science* **1,** 826–843. ISSN: 1759-0884. DOI: 10.1002/wcms.31 (2011).

165. Faradjian, A. K. & Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *The Journal of Chemical Physics* **120,** 10880–10889. ISSN: 0021-9606. DOI: 10.1063/1.1738640 (May 24, 2004).

166. Wu, H. *et al.* Variational Koopman Models: Slow Collective Variables and Molecular Kinetics from Short off-Equilibrium Simulations. *The Journal of Chemical Physics* **146,** 154104. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4979344 (Apr. 21, 2017).

167. Antognini, J. M. & Sohl-Dickstein, J. *PCA of High Dimensional Random Walks with Comparison to Neural Network Training* in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* Montréal, Canada (Curran Associates Inc., Red Hook, NY, USA, 2018), 10328–10337.

168. Hyvärinen, A., Karhunen, J. & Oja, E. *Independent Component Analysis* ISBN: 978-0-471-40540-5. DOI: 10.1002/0471221317 (June 2001).

169. Tong, L., Liu, R., Soon, V. C. & Huang, Y. Indeterminacy and Identifiability of Blind Identification. *IEEE Transactions on Circuits and Systems* **38,** 499–509 (1991).

170. Pronk, S. *et al.* GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics (Oxford, England)* **29,** 845–854. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt055 (Apr. 1, 2013).

171. Lindorff-Larsen, K. *et al.* Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins: Structure, Function, and Bioinformatics* **78,** 1950–1958. ISSN: 1097-0134. DOI: 10.1002/prot.22711 (2010).

172. Horn, H. W. *et al.* Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *The Journal of Chemical Physics* **120,** 9665–9678. ISSN: 0021-9606. DOI: 10.1063/1.1683075 (May 22, 2004).

173. Nakatsu, T., Kato, H. & Oda, J. Crystal Structure of Asparagine Synthetase Reveals a Close Evolutionary Relationship to Class II Aminoacyl-tRNA Synthetase. *Nature Structural Biology* **5,** 15–19. ISSN: 1545-9985. DOI: 10.1038/nsb0198-15 (1 Jan. 1998).

174. Jager, M. *et al.* Structure-Function-Folding Relationship in a WW Domain. *Proceedings of the National Academy of Sciences* **103,** 10648–10653. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0600511103 (July 11, 2006).

175. Gray, R. M. Toeplitz and Circulant Matrices: A Review. *Foundations and Trends® in Communications and Information Theory* **2,** 155–239. ISSN: 1567-2190, 1567-2328. DOI: 10.1561/0100000006 (Jan. 30, 2006).

176. Davis, P. J. *Circulant Matrices* 276 pp. ISBN: 978-0-471-05771-0 (Wiley, 1979).

177. Davis, C. & Kahan, W. M. The Rotation of Eigenvectors by a Perturbation. III. *SIAM Journal on Numerical Analysis* **7,** 1–46. ISSN: 0036-1429. DOI: 10.1137/0707001 (Mar. 1970).

178. Sengupta, U., Carballo-Pacheco, M. & Strodel, B. Automated Markov State Models for Molecular Dynamics Simulations of Aggregation and Self-Assembly. *The Journal of Chemical Physics* **150,** 115101. ISSN: 0021-9606. DOI: 10.1063/1.5083915 (Mar. 15, 2019).

179. Lundholm, I. V. *et al.* Considerations for Three-Dimensional Image Reconstruction from Experimental Data in Coherent Diffractive Imaging. *IUCrJ* **5,** 531–541. ISSN: 2052-2525. DOI: 10.1107/S2052252518010047 (Sept. 1, 2018).

180. Rose, M. *et al.* Single-Particle Imaging without Symmetry Constraints at an X-ray Free-Electron Laser. *IUCrJ* **5,** 727–736. ISSN: 2052-2525. DOI: 10.1107/S205225251801120X (Nov. 1, 2018).

181. Ignatenko, A. *et al.* Classification of Diffraction Patterns in Single Particle Imaging Experiments Performed at X-Ray Free-Electron Lasers Using a Convolutional Neural Network. *Machine Learning: Science and Technology* **2,** 025014. ISSN: 2632-2153. DOI: 10.1088/2632-2153/abd916 (Feb. 2021).

182. Assalauova, D. *et al.* An Advanced Workflow for Single-Particle Imaging with the Limited Data at an X-ray Free-Electron Laser. *IUCrJ* **7,** 1102–1113. ISSN: 2052-2525. DOI: 10.1107/S2052252520012798 (Nov. 1, 2020).

183. Roberts, G. O. & Rosenthal, J. S. Optimal Scaling of Discrete Approximations to Langevin Diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **60,** 255–268. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00123 (Jan. 1, 1998).

184. Lykkegaard, M. B., Dodwell, T. J., Fox, C., Mingas, G. & Scheichl, R. *Multilevel Delayed Acceptance MCMC* http://arxiv.org/abs/2202.03876. preprint.

185. Swendsen, R. H. & Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters* **57,** 2607–2609. DOI: 10.1103/PhysRevLett.57.2607 (Nov. 24, 1986).

186. Pande, K. *et al.* Ab Initio Structure Determination from Experimental Fluctuation X-ray Scattering Data. *Proceedings of the National Academy of Sciences* **115,** 11772–11777. DOI: 10.1073/pnas.1812064115 (Nov. 13, 2018).

187. Volkhardt, A. & Grubmüller, H. Estimating Ruggedness of Free-Energy Landscapes of Small Globular Proteins from Principal Component Analysis of Molecular Dynamics Trajectories. *Physical Review E* **105,** 044404. DOI: 10.1103/PhysRevE.105.044404 (Apr. 15, 2022).

188. Phillips, J. L., Colvin, M. E. & Newsam, S. *Dimensionality Estimation of Protein Dynamics Using Polymer Models* in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Association for Computing Machinery, New York, NY, USA, Aug. 15, 2018), 675–680. ISBN: 978-1-4503-5794-4. DOI: 10.1145/3233547.3233713.

Acknowledgments

I extend my deepest gratitude to my supervisor Prof. Helmut Grubmüller the opportunity to undertake this project. It was his suggestion to employ a Bayesian method, and I have immensely benefited from our consistently constructive and enlightening discussions.

I further thank the members of my thesis committee, Prof. Simone Techert and Prof. Thorsten Hohage, for their helpful questions and feedback during our meetings.

Special thanks are due to Nicolai Kozlowski, Malte Schäffner, Andreas Volkhardt, and Carsten Kutzner for many enlightening discussions. I am also grateful to Andreas Volkhardt and Nicolai Kozlowski for granting access to their MD-trajectories.

My thanks go out to Petra Kellers and Florian Leidner for proofreading sections of this manuscript.

For their unwavering support and assistance, I thank Eveline Heinemann, Stefanie Teichmann, Ansgar Esztermann, Martin Fechner, Frauke Bergmann, and Petra Kellers.

Lastly, I convey my heartfelt gratitude also to all other members of the Department of Computational and Theoretical Biophysics at the Max-Planck-Institute for Multidisciplinary Sciences.