Structure Determination From Single Molecule X-Ray Scattering Experiments using Photon Correlations

Benjamin von Ardenne

Göttingen 2017

Structure Determination From Single Molecule X-Ray Scattering Experiments using Photon Correlations

Benjamin von Ardenne

Dissertation
for the award of the degree
Doctor rerum naturalium (Dr.rer.nat.)
of the Georg-August University of Göttingen

within the doctoral program
International Max Planck Research School (IMPRS)
"Physics of Biological and Complex Systems"
of the
Göttingen Graduate School for Neurosciences, Biophysics,
and Molecular Biosciences (GGNB)

Submitted by
Benjamin von Ardenne
from Dresden

Göttingen, 10.08.2017

Thesis Committee:

Prof. Dr. Helmut Grubmüller
Department for Theoretical and Computational Biophysics,
Max-Planck-Institute for Biophysical Chemistry

Prof. Dr. Marcus Müller
Institute for Theoretical Physics, University of Göttingen

Prof. Dr. Holger Stark
Department of Structural Dynamics,
Max-Planck-Institute for Biophysical Chemistry

Members of the Examination Board:

First Reviewer: Prof. Dr. Helmut Grubmüller
Second Reviewer: Prof. Dr. Marcus Müller

Date of the Disputation:

18 October 2017

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Ich erkläre ausdrücklich, dass ich sämtliche in der Arbeit verwendeten fremden Quellen als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos sowohl bei wörtlich übernommenen Aussagen bzw. unverändert übernommenen Tabellen, Grafiken u. Ä. (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen bzw. von mir abgewandelten Tabellen, Grafiken u. Ä. anderer Autorinnen und Autoren (Paraphrasen) die Quelle angegeben habe.

Göttingen, den 10.08.2017

Benjamin von Ardenne

Table of Content

Abstract

Scattering experiments with femtosecond high-intensity free-electron laser pulses provide a new route to macromolecular structure determination without the need for crystallization at low material usage. In these experiments, the X-ray pulses are scattered with high repitition on a stream of identical single biomolecules and the scattered photons are recorded on a pixelized detector. The main challenges in these experiments are the unknown random orientation of the molecule in each shot and the extremely low signal to noise ratio due to the very low expected photon count per scattering image, typically well below the number of over 100 photons required by available analysis methods. The latter currently limits the scattering experiments to nano-crystals or larger virus particles, but the ultimate goal remains to retrieve the atomic structure of single biomolecules.

In light of that goal, here I present a correlation-based approach that can determine the molecular structure *de novo* from as few as three coherently scattered photons per image. I derive for the first time an analytic expression of the full three-photon correlation as a function of the molecules Fourier intensity using a spherical harmonics expansion and propose a Monte Carlo simulated annealing approach to solve the inverse problem of finding an intensity that fits the experimentally observed triple correlations. The size of the search space is reduced by using information from the analytic inversion of the two-photon correlation and the electron density is retrieved by applying an iterative phase retrieval method to the determined intensity.

Using synthetic scattering data of a small protein (46 residues) at realistic average photons counts of 10 photons per image, I demonstrate that near-atomic resolution of 3.3 Å can be achieved using $3.3 \cdot 10^9$ images, which is within experimental reach. Remarkably, the data acquisition time required to achieve the same resolution decreases to minutes if the average number of photons per image is increased to only 100 photons (equivalent to a decrease in the number of images by a factor 1000).

The noise levels in the experiment are expected to be quite high which is a challenge for all structure determination methods. To address this issue, I demonstrate that my three-photon correlation approach is robust to isotropic noise from incoherent scattering, and that the number of disordered solvent molecules attached to the macromolecular surface should be kept at a minimum.

List of Symbols

A list of mathematical symbols that are used in this thesis:

3D Structures

$\rho(\mathbf{x})$	Electron density
$A(\mathbf{k})$	Complex Fourier transform of the electron density ρ
$I(\mathbf{k})$	Intensity as the absolute square of the Fourier transform of the electron density $\rho(\mathbf{x})$
L	Spherical harmonics expansion limit
K	Number of spherical harmonics shells
k,$\mathbf{k}$	Reciprocal wave number/vector
k_{cut}	Wave number to which the intensity is eventually determined
k_{max}	Maximum wave number for which signal is measured
$\mathbf{K}$	2D vector in $k_x k_y$-plane
P_{S_+}	Positivity and support projection of the phase retrieval
P_M	Amplitudes projections (constraints) of the phase retrieval
$N(\mathbf{k}$	Intensity distribution of a noise model
γ	Signal-to-noise level

Photon Correlations

α, β	Angles between the photon correlation with α, $\beta \in [0, \pi]$
N	Number of discrete angles in the photon correlation histograms $\Delta\alpha, \Delta\beta = \pi/N$
$c(k_1, k_2, \alpha)$	Two-photon correlation function as expected for an intensity $I(\mathbf{k})$

$t(k_1, k_2, k_3, \alpha, \beta)$ Three-photon correlation function as expected for an intensity $I(\mathbf{k})$

$c_{k_1,k_2,\alpha}$ Two-photon correlation histogram counts as expected for an intensity $I(\mathbf{k})$

$t_{k_1,k_2,k_3,\alpha,\beta}$ Three-photon correlation histogram counts as expected for an intensity $I(\mathbf{k})$

$\tilde{t}_{k_1,k_2,k_3,\alpha,\beta}$ Normalized three-photon correlation histogram

$h_{k_1,k_2,\alpha}$ Measured two-photon correlation histogram counts

$h_{k_1,k_2,k_3,\alpha,\beta}$ Measured three-photon correlation histogram counts

$\tilde{h}_{k_1,k_2,k_3,\alpha,\beta}$ Normalized measured three-photon correlation histogram counts

$\mathbf{U_l}$ Arbitrary $2l+1$-dimensional unitary matrix

Spherical Harmonics

$A_{lm}(k)$ Spherical harmonics coefficients of intensity $I(\mathbf{k})$

$F_{lm}(k)$ Spherical harmonics coefficients of Fourier density $A(\mathbf{k})$

$R_{lm}(k)$ Spherical harmonics coefficients of electron density $\rho(\mathbf{r})$

$D^l_{mm\prime}(\alpha, \beta, \gamma)$ Wigner-D matrix element for the rotation of spherical harmonics

$P_l(\cos\theta)$ Legendre polynomial of order l

$sphP_l(\cos\theta)$ Spherical legendre polynomial of order l

$Y_{lm}(\theta, \varphi)$ Spherical harmonics basis function

$j_l\,(x)$ Spherical Bessel function of order l

1 Introduction

In Nature a large variety of biomolecules has emerged, differing each in structure, dynamics and function. The biological function is largely determined by the conformational dynamics which in turn is almost exclusively encoded in the molecular structure i.e., the exact positions of the residues and comprising atoms in the three-dimensional fold of the molecules. Without accurate models of these structures, e.g., predictions about the dynamics and functions of biomolecules by molecular dynamics simulations or structure-based drug design become challenging.

To this end, early on, structural biology has sought atomistic structure determination of proteins, nucleic acids, lipids, carbohydrates, and complexes thereof. The first atomic structures of larger biomolecules were determined using X-ray crystallography, e.g., Deoxyribose Nucleic Acid (DNA) by Franklin and Wilkins [1] in 1953 (Nobel Prize in Physiology in 1962), Myoglobin by Max Perutz [2] in 1957 (Nobel Prize in Chemistry in 1962 shared with Kendrew) or Lysozym by Blake [3] in 1965. Because rigid biomolcules are the easiest to grow crystals from, they were the first to be studied and therefore thought to be overall "quite rigid"[4]. Over time, the rigid picture was replaced by that of dynamic proteins that constantly move at physiological temperatures and explore the conformational space around the averaged structures that are usually measured in X-ray experiments. Determining the biomolecular structure in the presence of large conformational changes remains a big challenge in the field.

Today, many techniques are used for structure determination, most commonly X-ray crystallography, cryo electron microscopy (cryo-EM), nuclear magnetic resonance spectroscopy (NMR) and molecular modelling. In addition, many other experimental methods are used to support structure visualization, albeit lacking the capability of *de novo* atomistic structure determination of entire proteins. Among them are fluorescent imaging techniques such as fluorescence resonance energy transfer (FRET) and stimulated emission depletion (STED) microscopy, atomic force microscopy (AFM) and small-angle scattering (SAS) both with X-rays and neutrons (SAXS/SANS).

Over 20 Nobel prizes have been awarded for work related to structure determination, for example, for the structure determination of the photosynthetic reaction centre by Deisenhofer, Huber, and Michel in 1988, water and ion-channels by Agre and MacKinnon in 2003, the RNA polymerase by Kornberg in 2006, the ribosome by Ramakrishnan, Steitz, and Yonath in 2009 and the G protein-coupled receptors by Lefkowitz and Kobilka in 2012 [5].

All structure determination approaches have individual advantages and disadvantages that are complementing each other, most of which are discussed in more detail in this thesis. X-ray crystallography, for example, fails when the target protein does not form crystals or cannot be purified in sufficient quantities.

NMR, on the other hand, does not require crystallization, but instead requires a substantial quantity of biomolecules in solution, which are difficult to synthesize and potentially forms unphysiological aggregates at the high concentrations needed. In contrast to scattering experiments, NMR structure determination also becomes more challenging with larger molecules (>100kDA).

In the recent years, only cryo-EM has produced a growing rate of deposited structures mainly due to improvements of the detectors and structure determination algorithms. Although cryo-EM has proven to be a very reliable method, the structure determination of small biomolecules remains challenging because very noisy single particle images are difficult to extract from the background. Both in X-ray crystallography and cryo-EM, the time resolution is limited and the molecules are imaged at unphysiological conditions (e.g., in cryo-EM the samples are plunge-frozen down to $-269°C$).

Despite the great effort in the three fields over the past 60 years, the structures of only about 0.75% of the more than 18 million known proteins [6] have been determined to high resolution [7].

In light of the large number of unsolved structures and the individual limitations the existing methods, single particle scattering experiments with high-intensity X-ray free-electron lasers (XFELs) have been proposed by Neutze *et al.* [8] as a novel approach to solving the three-dimensional atomic structure of biomolecules without the need for crystallization at low material usage [9–15].

In these experiments, high-repetition and high-intensity X-ray pulses are scattered on a stream of single randomly-oriented biomolecules and only a few photons are scattered by the molecules and recorded on an extremely sensitive pixelized photon detector. The short femtosecond pulses outrun the severe radiation damage due to Auger decay and Coulomb explosion ("diffract and destroy" experiment) and thus allow for extremely high peak brilliance to the point where individual molecules can be imaged. Like in conventional X-ray crystallography, the phases are not measured in such an experiment but in the absence of crystals the scattering patterns are continuous and the phase problem is accessible to *ab initio* phase-retrieval methods.

Whereas previous X-ray sources, including synchrotrons, have primarily engaged in studies of static structures, X-ray FELs are by their nature suited for studying dynamic systems at the time and length scales of atomic interactions. In contrast to structure determination methods that measure a structure ensemble (NMR, X-ray crystallography, SAXS, FRET), this new method can distinguish e.g., between different native conformations, by sorting the single molecule images. Further, in systems where reactions can be induced, e.g., by light, a sequence of structures

at different reaction times may be recorded which opens the window to molecular movies as a long-standing dream [16].

In the first proof of principle single molecule scattering experiments with the available X-ray source in Stanford (LCLS), the 3D structure of single mimivirus particles was determined to a resolution of 125 nm [17, 18], using images with more than 10^7 scattered photons each. However, for a medium sized molecule and an expected XFEL fluence of $6.3{\cdot}10^7$ photons/nm^2 [19] at a 100 nm focus diameter and 5 keV beam energy, only about 10-50 coherently scattered photons per scattering image are expected [20].

Standard analysis methods cannot cope with the high statistical noise in this extreme Poisson regime, and hence so far all XFEL structure determination attempts resorted to nano-crystals [21–28]. A particular challenge is to determine the orientation of the molecule for each image to assemble all recorded images in 3D Fourier space for subsequent electron density determination.

For single molecule scattering experiments, several orientation determination methods were developed [29–36], which however require at least 100 photons per image. Alternatively, manifold reconstruction algorithms (manifold embedding) [37–40] forego the explicit assembly in Fourier space and instead use the similarity between scattering images to determine the manifold of orientations. Because these algorithms work solely on the manifold level, they are not guaranteed to generate a self-consistent 3D intensity and are prone to instability in the presence of noise. Moreover, also for these methods, successful structure determination was reported only for much more than 100 photons per image.

Photons correlations, as a summary statistic of the structure which is independent of the image orientation, are a possible solution to the very low photon counts per image because they can be either sampled by recording more photons per image or by recording more images. In fluorescence microscopy or cryo-electron microscopy, time integrated and time-correlated single-photon counting has been successfully used at extremely low signal-to-noise ratios [41]. In the context of single molecule X-ray scattering, Saldin *et al.* were the first in 2010 to demonstrate the use of two-photon correlations for the determination of the molecular shape of symmetric particles [42, 43] and the structure of particles randomly oriented around an axis [44, 45]. However, as already shown by Kam [46] in 1980, two-photon correlations do not contain enough information to retrieve the 3D structure *de novo*.

Based on early analytic work by Kam on degenerate three-photon correlations [46] – two out of three photons are recorded at the same position – structure determination of mesoscopic cylindrical particles (2012) [47] and of a highly symmetric icosahedral virus (2015) [48, 49] was demonstrated. As this approach is limited to only a small fraction of the recorded correlations, however, also this method has so far not been applied to *de novo* single molecule structure determination.

Despite the limited application of his method at the low photon counts, Kam's method demonstrated that the combined information of the two-photon and degenerate three-photon correlation fully encodes the 3D structure. Based on this assertion, I concluded that, instead, the *full* three-photon correlation should be used for the structure determination because it is sampled much better by the few photon scattering images. However, it was unclear if the additional information in the rest of the three-photon correlation is sufficient to compensate the sparsely sampled degenerate part and if unique solutions can be found. To this end, in this Doctoral thesis, I derived the analytic expression of the *full* three-photon correlation and developed an approach which uses these correlations, for the first time, for *de novo* atomistic structure determination from the sparse single molecule X-ray scattering images.

The next-generation free electron lasers are still under construction or testing and therefore experimental data of proteins is not available yet. As a preparation for the application of the method to experimental scattering data, I will validate the method using synthetic scattering images of a medium-sized Crambin molecule as a test-system using realistic estimates for the number of scattered photons. In particular, I will address the question, how the achieved resolution depends on the number of recorded images and further determine how these numbers change at different average photon counts per image.

Noise due to incoherent scattering, the photoelectric effect, background radiation, contaminants such as water molecules that adhere to the molecules' surface or detector noise will most likely be the limiting factor in single molecule structure determination. I will therefore also demonstrate the structure determination in the presence of additional non-Poissonian noise and study the dependence of the achieved resolution on the shape and strength of the noise.

As further assessment of the method, I will evaluate the impact of the phasing error on the resolution, discuss what the best model parameters are for maximizing the resolution and minimizing the computational effort and investigate at which point over-fitting occurs given the finite number of sparse scattering images.

Thesis Overview

In Chapter 2 I will begin with a brief overview of proteins which have emerged with a large variety of structures and functions and are the main subject of the presented structure determination method. In Section 2.2, I will discuss the three major established structure determination methods (X-ray crystallography, NMR and cryo-EM) with respect to their scope of application and their advantages and disadvantages in contrast to single molecule X-ray scattering. Next, I will describe the novel experimental setup of single molecule scattering in Sec. 2.3 along with the operation of a free-electron laser and the physics behind the extremely-high peak brilliance, which eventually allows for single molecule imaging. In the

short overview of coherent diffraction theory in Sec. 2.3.2, I will explain how the photon distribution of a scattering image is analytically connected to the electron density of the molecule and calculate an estimate for the number of coherently and incoherently scattered photons, both by the protein and the potential unstructured water shell using realistic beam intensities.

In Chapter 3, I will discuss already proposed single molecule X-ray scattering analysis methods and compare them with respect to their advantages and disadvantages. In particular, I will focus on the work that has been done on correlation-based methods and finish with the current state of research, clearly separating my contribution to the structure determination problem.

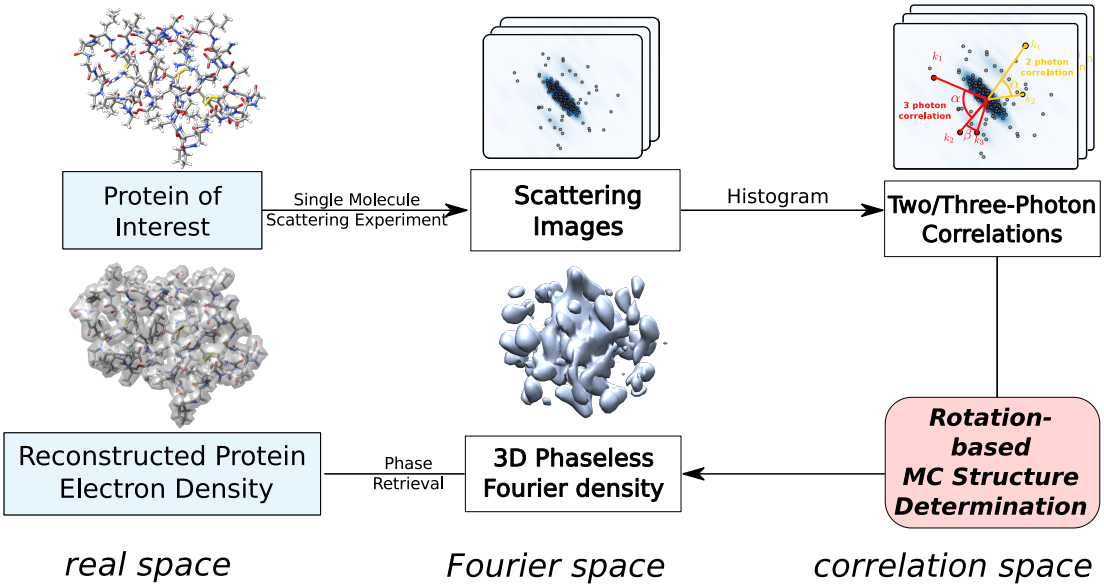


Figure 1.1: Summary of the correlation-based *de novo* structure determination method. The two- and three-photon correlations in the scattering images are histogrammed. The Monte Carlo simulated annealing approach (pink) determines the optimal intensity that fits both the two-photon and the three-photon correlation. The retrieved intensity is phased and the final electron density is obtained.

In Chapter 4, I will introduce my novel *de novo* structure determination approach which uses the *full* three-photon correlation. First, I will define the two- and three-photon correlations and derive, for the first time, the analytic expressions of the three-photon correlation in terms of spherical harmonics expansions in Sec. 4.1. The three-photon correlation is difficult to invert analytically which is why I chose a probabilistic approach and determine the most-likely structure that has generated the experimentally-measured photon correlations (Sec. 4.2). The size of the search space is reduced by isolating the search to structures that also

fit to the measured two-photon correlation as explained in Sec. 4.3. The structure optimization is done by a Monte Carlo / simulated annealing approach which I explain in Sec. 4.4, including the definition of the ergodic Monte Carlo steps and the temperature decay protocol. The computation time, which is a limiting factor, is improved drastically by using histograms of the correlations (see Sec. 4.5) and a high degree of parallelism in the computation of the three-photon correlations (see Sec. A.2.2). After determining the structure in Fourier space, the phases are retrieved using established algorithms that I modified to work with the spherical harmonics expansion, as detailed in Sec. A.1.3.

In Section 4.6 I further explain the choice of the optimal spherical harmonics parameters that minimize the computational effort and maximize the resolution. In the end of the chapter, I will discuss the structure determination in the presence of additional isotropic Non-Poissonian noise in Sec. 4.7. Figure 1.1 summarizes the steps involved in determining the electron density from sparse scattering images using the proposed three-photon correlations approach.

In Chapter 5, I discuss how I validated the method using synthetic scattering image of a 45-residue Crambin protein. In particular, I will explain the rejection method used to generate the images and describe how I calculate the resolution of the phased electron densities using the Fourier shell correlation (FSC).

The structure determination approach and the validation methods are implemented in the *ThreePhotons.jl* software package which I will briefly discuss in Chapter 5.3. The library contains more than 5000 lines of well-tested and highly-optimized code and has been published open-source[1].

In the last Chapter 6, I will show structure determination results of Crambin that were retrieved using up to $3.3 \cdot 10^9$ synthetic scattering images with 10 photons on average. The structure result with the maximum achieved resolution are shown in Sec. 6.1 and the dependence of the resolution on the number of scattering images for a fixed number of photons per image is discussed in Sec. 6.2. Similarly, in Section 6.3, I will assess how the resolution changes if the photons are distributed on fewer or more images using scattering images with on average 10, 25, 50 or 100 photons.

The change of the structure resolution in the presence of additional noise is shown in Sec. 6.4, where I will use a Gaussian noise model with different widths and noise-levels to mimic different sources of noise. In the end of the Chapter, I will evaluate the error due to phasing and determined the structure with different model parameters to study at which point over-fitting occurs.

[1] `https://github.com/h4rm/ThreePhotons.jl`

2 Background on Structure Determination

2.1 Proteins - The Building Blocks of Life

Proteins are large biomolecules that are often referred to as "the molecular machines of the body" due to their various shapes and functions. In their complex structure, in most cases, the position of every atom is predetermined by nature and important for the dynamics and functions, motivating numerous efforts in atomic structure determination (including single molecule X-ray scattering experiments). Here, I will give a brief overview of proteins and in particular discuss their fundamental building blocks – the amino acids –, how they are typically assembled into higher-order structures and what complex functions have emerged from these structures within organisms.

Human proteins are comprised of only 20 different amino acids (residues) as shown in Fig. 2.1, though in certain cases also selenocysteine (denoted Sec or U) and archaea-pyrrolysine (denoted Pyl or O) are incorporated. Amino acids consist of amine (-NH2) and carboxyl (-COOH) functional groups and differ only by the side chain (R group)[2]. They are classified into seven chemical groups, defined by the properties of these side chains [50]: aliphatic (alanine, glycine, isoleucine, leucine, proline, valine), aromatic (phenylalanine, tryptophan, tyrosine), ccidic (aspartic acid, glutamic acid), basic (arginine, histidine, lysine), hydroxylic (serine, threonine), sulphur-containing (cysteine, methionine) and amidic (asparagine, glutamine).

These chemical properties make them either a weak acid or a weak base, or a hydrophile if the side chain is polar or a hydrophobe if it is nonpolar. Often, these properties are the key to their interaction with their physiological environment, e.g., the formation of hydrophilic and hydrophobic surfaces allow the protein to be stably embedded into proteins [51].

In the scattering experiments, however, the chemical properties and the covalent bonds are less important because the photons are scattered on the individual carbon, oxygen, and nitrogen atoms. Nevertheless, the chemical knowledge is used

[1]Provided by Andy Brunning under Creative Commons: `http://www.compoundchem.com/2014/09/16/aminoacids/`

[2]`https://en.wikipedia.org/wiki/Amino_acid`

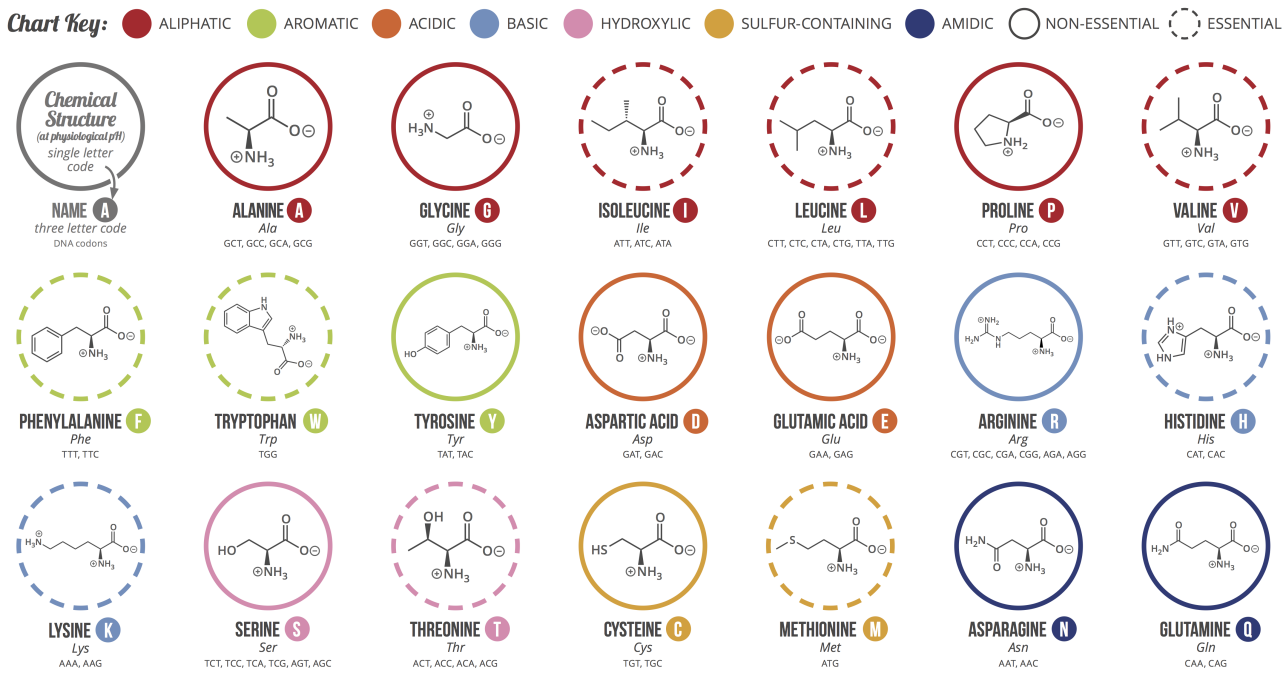


Figure 2.1: The human genetic code directly encodes 20 amino acids which are obtained from diet (essential) or synthesis (non-essential)[1]. Amino acids are classified into seven chemical groups: aliphatic, aromatic, acidic, basic, hydroxylic, sulphur-containing or amidic which make them weak acids, weak bases, hydrophilic or hydrophobic.

to reconstruct the position of hydrogen atoms which are usually not resolved in the scattering experiments due to their low scattering cross-sections. See Section 2.3.3 for a discussion of the coherent X-ray scattering cross-sections of the involved atoms.

The sequence of the proteins amino acids is encoded linearly in the DNA and assembled in-vitro by the ribosome through subsequent processes called transcription, the read-out of the genetic information from DNA to mRNA, and translation, the linear assembly of the protein by matching amino acids with the appropriate codons in the mRNA [51]. The protein sequence is extracted experimentally with little work, typically with mass spectrometry or Edman degradation using a protein sequenator [52]. The length of the polypeptide chain of a protein usually exceeds 20-30 amino acids residues, smaller chains are simply referred to as peptides.

Over 60% of eukaryotic proteins fold to one or more specific 3D conformations, while the rest remains mainly intrinsically disordered[3]. The 3D structure is comprised of a hierarchy with 4 levels[4]:

[3] https://en.wikipedia.org/wiki/Intrinsically_disordered_proteins
[4] https://en.wikipedia.org/wiki/Protein

- Primary structure: The linear sequence of the amino acids as produced by the ribosome.
- Secondary structure: Local structures such as α-helices and β-sheets, emerging from patterns of hydrogen bonds between the main-chain peptide groups.
- Tertiary structure: Three-dimensional structure of the monomeric (one component, e.g., ubiquitin) or multimeric (multiple components, e.g., aquaporin) protein.
- Quaternary structure: Aggregation of two or more proteins to a large macromolecule (e.g., ribsome with 2 subunits).

Protein folding is a complex process in which the residues move at physiological temperatures and form secondary and tertiary structures, mainly driven by hydrophobic interactions, the formation of hydrogen bonds and van der Waals forces. In nature the exact same folds are usually achieved within milliseconds to seconds. This is surprising because even for a medium-sized protein, the time to explore all possible confirmations would quickly exceeds the time of the universe according to the Levinthal's paradox [53]. As a solution, many structures have evolve whose motions are characterized by steep funnel-like energy landscapes which guide the folding motion through a small part of phase space and exclude large parts of the conformational entropy through high energy barriers. Even larger multi-domain proteins solve the conformational challenge with a "divide and conquer"-method by folding in multiple sub-steps. In some cases Chaperones aid the folding process by shield the spontaneously folding proteins from external disturbances.

If, despite the effort, proteins fail to fold into their native structure, they become inactive and in same cases even have toxic functionality. Several neurodegenerative (and other) diseases are suspected to arise from aggregates of misfolded proteins and many allergies are caused by the incorrect folding of proteins.

Predicting the fold of a protein from the primary structure alone remains one of the big challenges in the field. For small molecules, the accuracy and the times scales accessible by molecular dynamics simulations are usually sufficient to extensively sample the conformational space and to localize the folded structure as the free energy minimum. For larger molecules, homology modeling methods may derive the 3D fold of a protein from experimental structures of evolutionarily-related proteins.

De novo 3D structure determination rests on three established methods – X-ray crystallography, NMR and cryo-EM – as discussed in the following Section. From the limited set of 20 amino acids, a versatile zoo of structures has emerged as shown in Figure 2.2.

[5]Provided by Axel Griewel under Creative Commons: `https://en.wikipedia.org/wiki/Protein_structure`.

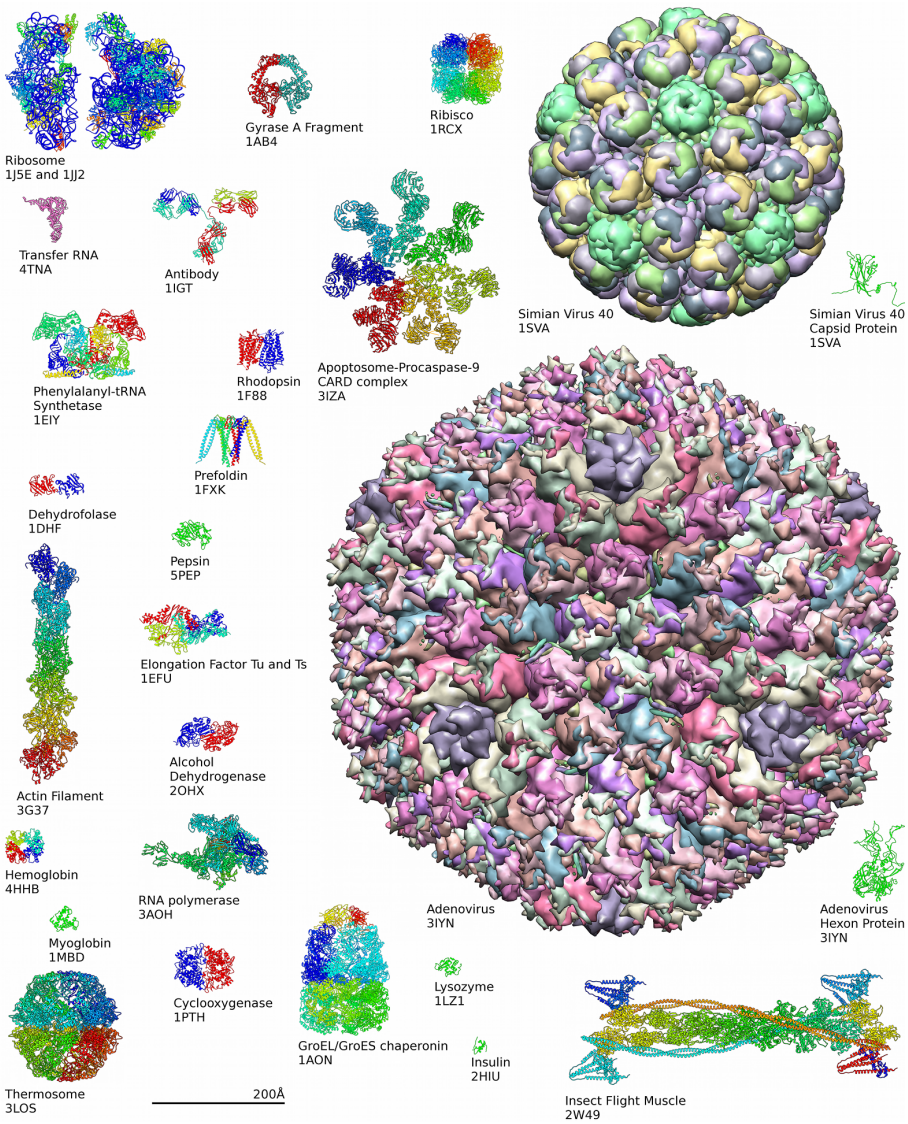


Figure 2.2: Example for the diversity of protein structures available from the PDB and EMDB[5].

Once folded, the proteins are usually not rigid. Instead, they undergo a variety of (fast) vibrations and (slower) structural rearrangements, the latter being called 'conformational transitions'. These conformational transitions are often implied by the the protein structure and are usually responsible for the biological function. The atomic motions happen on a nanometer length scale and a femtosecond timescale which makes them difficult to observe experimentally.

The motion of the folded structure in solution determines the protein's function. Among many other tasks in the body, proteins are involved in

- transporting molecules (e.g., ion-channels [54] or water-conducting channels aquaporin [55])
- responding to stimuli (e.g., SNARE proteins in the synaptical vesicles for neuronal transmissions [56])
- synthesizing other proteins (e.g., ribosome [57])
- catalyzing metabolic reactions (e.g., lactase, alcohol dehydrogenase or DNA polymerase)[6]
- composing structural elements in connective tissues (e.g., microtubules or actin filaments) [51]

So far, only molecular dynamics simulations give a complete view of the conformational dynamics, albeit depending on the accuracy of the force fields (which model/approximate various electron-related effects that influence the dynamics). All three established structure determination experiments address the challenge of imaging "molecular movies" of conformational changes or molecular reactions on a nano-scale, but no universal method exists yet that resolves the 3D trajectory *de novo*.

2.2 Established Structure Determination Experiments

Three major structure determination methods solve biomolecular structures *de novo*: X-ray crystallography, cryo electron microscopy (cryo-EM) and nuclear magnetic resonance spectroscopy (NMR). All three methods utilize different physical effects to image the atoms of the molecules.

In X-ray crystallography, high-energetic photons are coherently scattered by the electrons of the molecules (elastic photon scattering). In contrast, in cryo-EM, high-energetic electrons are coherently scattered at the positive protons of the atoms (Coulomb interaction). In NMR, the spin in the nucleus of the proteins are aligned with a strong constant magnetic field and probed with a high-frequency radio-pulse, resulting in distinctive measurable resonance signals (Nuclear Overhauser effect).

Here, I will review the three methods with respect to their mutual advantages and disadvantages, the latter motivating the development of single molecule X-ray scattering.

[6] `https://en.wikipedia.org/wiki/Enzyme`

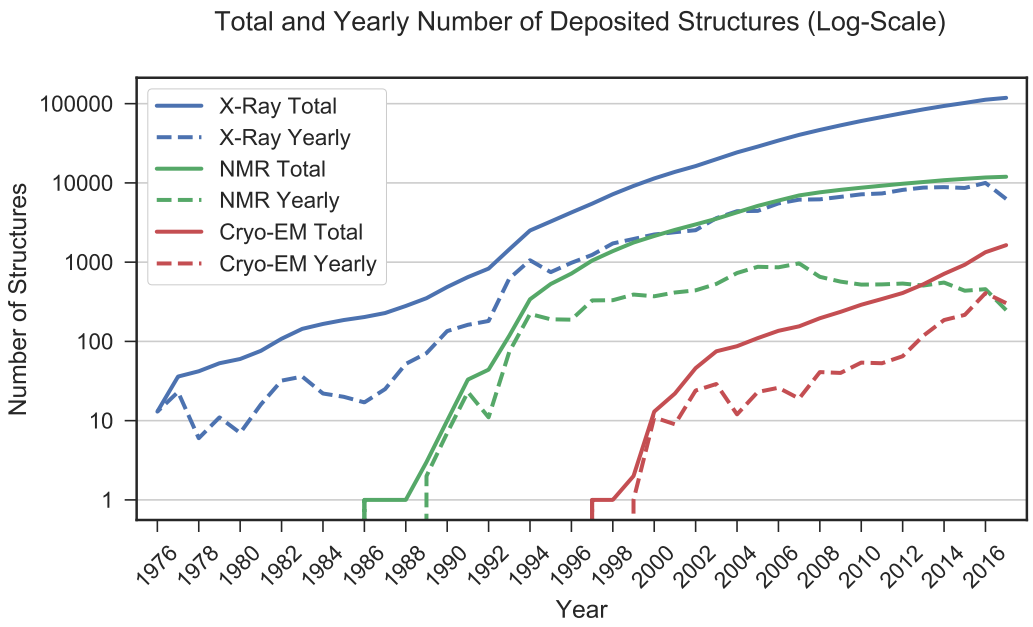


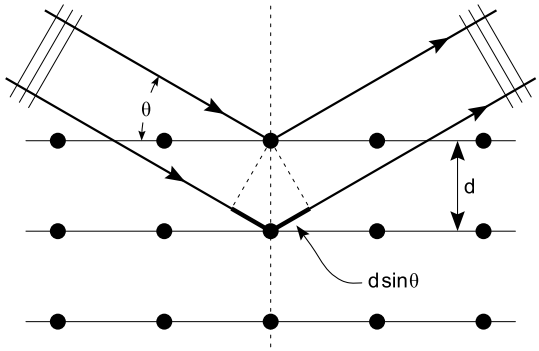
Figure 2.3: Logarithm of the total and yearly number of deposited structures in the PDB [7] from X-ray crystallography, NMR and cryo-EM between 1976 and 2017.

2.2.1 X-Ray Crystallography

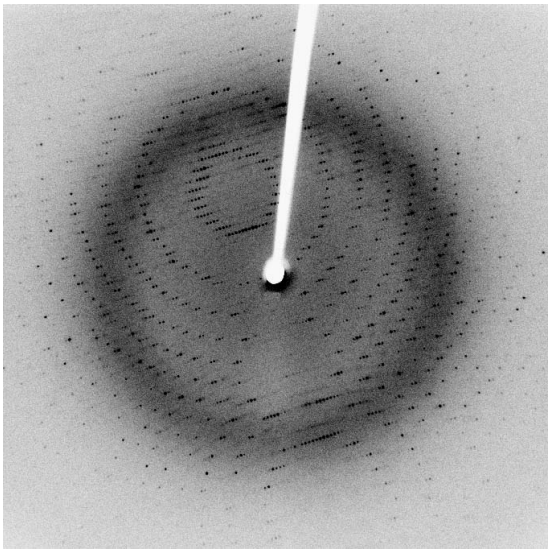
The X-ray crystallography method is the oldest structure determination method which was pioneered by William Henry and his son William Lawrence Bragg in 1912. The first of such crystallographic experiments with biological molecules were carried out in 1923 on hexamethylenetetramine [58] and later structures of cholesterol, vitamin B12 and penicillin were determined by Dorothy Crowfoot Hodgkin between 1937 and 1945. Starting in the 1950s, larger biomolecules were resolved with X-ray crystallography, e.g., DNA by Franklin and Wilkins in 1953, Myoglobin by Max Perutz in 1957 or Lysozym by Blake [3] in 1965. Since then, X-ray crystallography has proven a reliable method to produce the largest output of over 89% of the known structures according to the RCSB Protein Data Bank (PDB) [7] as shown in Fig. 2.3.

In the experiment, bright X-ray light from synchrotron radiation or free electron lasers is diffracted by millions of identical biomolecules which are placed on a three-dimensional crystalline grid [59]. This results in a pointed diffraction pattern which is comprised of the so called Bragg-peaks or Bragg-reflections (see Fig. 2.4b). These peaks form when scattered light from multiple molecules in the crystal interfere

[7] https://en.wikipedia.org/wiki/X-ray_crystallography



(a) 2D projection of the reflection of coherent planar light waves on successive parallel crystal planes[7]. The path difference between two reflected waves is $2d\sin(\theta)$ and the waves interfere constructively when this difference is a multiple n of the wavelength λ. In all other directions, the inference condition is not met and the waves from many plane reflections interfere destructively.



(b) Exemplary X-ray diffraction pattern of crystallized 3Clpro, a SARS protease with signal up to 2.1 Å resolution[7]. The pattern of spots (reflections) and the relative strength of each spot (intensities) are used to determine the structure. (Image provided by Jeff Dahl licensed under Creative Commons)

such that they cancel out in almost all directions except the scattering directions where the Bragg-peaks lie.

Constructive interference in direction θ happens if the path difference $2d\sin(\theta)$ between the reflected waves is an integer multiple of the wavelength λ, as stated by *Bragg's law* (see Fig. 2.4a),

$$2d\sin\theta = n\lambda. \tag{2.1}$$

The position and distance between the reflections depend on the inter-planar distance d and the wavelength of the beam λ. The effect of the constructive or destructive interference intensifies because of the cumulative effect of reflection in successive crystallographic planes of the lattice (as described by Miller notation (hkl)).

Depending on the biomolecules, different crystal structure are used for crystallization, e.g., cubic, tetragonal or rhombohedral Bravais lattices[8] resulting in

[8] `https://en.wikipedia.org/wiki/Bravais_lattice`, `https://en.wikipedia.org/wiki/Crystal_structure`

different geometries of the reflecting crystal planes and the reflection patterns, respectively. The molecular structure itself is encoded in the relative intensities of the peaks F_{hkl} (*structure factors*) which are associated with the amplitudes of the Fourier transform of the molecules' electron density.

These structure factors F_{hkl} are the product of the Fourier transform of the lattice and the Fourier transform of the molecule's electron density, $\mathcal{F}[\text{lattice}] \times \mathcal{F}[\text{molecule}]$ (convolution theorem),

$$F_{hkl} = \sum_{j=1}^{N} f_j \mathrm{e}^{[-2\pi i \mathbf{k}\cdot\mathbf{x}]} \tag{2.2}$$

$$= \sum_{j=1}^{N} f_j \mathrm{e}^{[-2\pi i (hx_j + ky_j + lz_j)]} \tag{2.3}$$

Here, the sum is over all N atoms at positions $\mathbf{x}$ in a unit cell and f_j is the *atomic form factor* of the j'th atom, $f_j(\mathbf{k}) = \int \rho(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}} \mathrm{d}^3\mathbf{r}$. The wave vector k is expressed in the basis $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ of the lattice, $\mathbf{k} = h\mathbf{a} + k\mathbf{b} + l\mathbf{c}$ and the Miller indices (hkl) define a reciprocal lattice point which corresponds to the real-space crystal plane at which scattering occurred (depending on the lattice type).

In a perfect crystal, the scattering factors are discrete numbers at position $\mathbf{k}(hkl)$ on the detector and the intensity at $\mathbf{k}$ depends on the squared modulus $I(\mathbf{k}) \propto |F_{hkl}|^2$ of the crystallographic structure factors. Please note, that in experiments on single or disordered molecules, in contrast to crystallography, the continuous atomic form factors are measured as further discussed in the derivation of coherent diffraction theory in Sec. 2.3.2.

The phases are not measured in X-ray crystallography experiments and numerous phase-retrieval methods have been developed, among them, e.g., ab initio phasing (similar to phasing in single molecule experiments) [60, 61], molecular replacement using the phases of similar structures [62], anomalous X-ray scattering (MAD or SAD phasing) [63] or heavy atoms methods [62]. The phased electron density map is used to build an atomic model of the protein by first fitting the backbone structure and subsequently optimizing the orientations of the side-chains of the amino-acids. These methods for model building are not exclusive to X-ray crystallography and may also be applied in the context of single molecule scattering.

Due to the high number of scattering sources in the crystal, a strong signal is detected also in the high-angle scattering regions which corresponds to high-resolution information about the molecule's electron density (see Sec. 2.3.2 for a detailed discussion of coherent scattering and the resulting spatial resolution). For that reason, molecular structures with a resolution better than 1.0 Å have been determined by X-ray crystallography (e.g., the human aldose reductase at 0.66 Å resolution), sometimes even resolving small hydrogen atoms to precise position [64]. X-ray crystallography is particularly useful for large molecules such as proteins.

However, crystallography requires the growth of large crystals from purified biomolecules which is not always possible, in particular for very flexible molecules. Also sometimes, the biomolecules have to be slightly altered (e.g., cutting of loops) or embedded in a non-physiological environments (e.g., solvents different from water) to stably form large crystals. This has rendered some classes of proteins inaccessible for X-ray crystallography, e.g., disease-associated protein aggregates, disordered proteins and membrane proteins [65].

Carrying out the experiment usually requires expensive beam time at large synchrotrons or free electron lasers which is scarce and difficult to obtain, especially for smaller research teams.

In the recent years, the development of new and brither free electron lasers enabled scattering experiments on nano-crystals [21–28] which are easier to grow. In order to record sufficient signal and to avoid radiation damage, in serial nanocrystallography, various methods have been devised to successively record many images of different nano-crystals in similar setup as single molecule scattering. Within this experimental framework, recording molecular movies of the kinetics of non-equilibrium chemical reactions or light-induced conformational changes becomes feasible by looking at identically prepared nano-crystals at different subsequent points in time. X-ray crystallography is still an evolving field and will keep its significance for the foreseeable future.

2.2.2 Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance spectroscopy determines the structure of proteins in solution without the need for crystallization [66–68]. In 1938, Isidor Rabi first described and measured the nuclear magnetic resonance effect in molecular beams (Nobel Prize in Physics in 1944[9])) by extending the the Stern–Gerlach experiment. Felix Bloch and Edward Mills Purcell further improved the concept of NMR for the use on liquids and solids in 1946 (shared Nobel Prize in Physics in 1952[10]).

In the experiment, the nuclei spins S are measured in a constant external magnetic field B_0, in which the energy difference between the spin levels becomes $\Delta E = \gamma\hbar B_0$. In order to flip the spin e.g, from $-1/2$ to $1/2$ in the case of isolated hydrogen or carbon atoms, an electromagnetic field oscillating with the Larmor frequency ν_0 needs to be applied, such that the energy of the field matches the energy difference between the spin states, $\Delta E = h\nu_0 = \gamma\hbar B_0$. Throughout time, the flipped spins relax back towards their thermal equilibrium and the nuclear magnetic resonance response is measured.

In the complex electrostatic environment of a protein, however, the surrounding electrons (carrying a spin themselves and producing a magnetic field opposite the

[9] http://www.nobelprize.org/nobel_prizes/physics/laureates/1944/
[10] http://www.nobelprize.org/nobel_prizes/physics/laureates/1952/

external field) act as a "shield" and reduce the magnetic field at the individual nucleus. As a result, the resonance frequencies of the spins are shifted in a specific way, uniquely defined by the 3D structure of the chemical environment, and the frequencies encode the structural information.

Normally, instead of slowly recording spectra in the frequency domain, radiofrequency pulses are applied to the sample and time-dependent responses are Fourier transformed to retrieve an NMR spectrum [69]. In two-dimensional nuclear magnetic resonance spectroscopy (2D-NMR), a series of pulses manipulates the coherence of the spins and the decay signal is measured similar to one-dimensional FT-NMR. The shapes, frequencies and durations of these pulses distinguish different NMR experiments from one another. In 2D NMR experiments there are two frequency axes representing a chemical shift and the axis are associated with the length of the pulsing period and the time elapsed during the detection period. After Fourier transformation, the measured data is comprised of intensity value for each pair of frequency variables as can be seen in the exemplary NMR spectrum shown in Fig. 2.5.

For protein structure determination, the cross-relaxation (a mechanism related to spin–lattice relaxation) is measured by perturbing the magnetization of a spin and observing the change in magnetization of the other spins as the equilibrium is reestablished (nuclear Overhauser enhancement effect (NOE)). The strength of the NOE is inversely proportional to distance between the interacting spins with $\sim r^{-6}$, thus limiting NOE signals to interactions within 5 Å [68]. The result of a NOSEY spectrum are interatomic distances between close atoms and residues (see Fig. 2.5) from which a structural model is built using (metric matrix) distance geometry [71].

In 2002 Wüthrich was awarded the Nobel prize in Chemistry for using the nuclear Overhauser effect spectroscopy (NOESY) to determine the three-dimensional structure of biological macromolecules in solution from two-dimensional NMR spectrocopy [69]. Today, over 10426 proteins or 8% of all known proteins have been determined with NMR [7] but the yearly structure depositions are declining (see Fig. 2.3). Recently, magic-angle spinning solid-state NMR has proven to resolve the structure of biomolecules, for which X-ray crystallography or solution NMR spectroscopy fail, such as membrane proteins and disease-related protein aggregates (see Ref. [65] for more detail on MAS NMR).

The strength of all NMR methods is the ability to resolve protein structures under physiological conditions (temperature, ion-concentration, solvent). In contrast to cryo-EM, very small and flexible proteins can be resolved with NMR techniques. Also, NMR spectrocopy does not deteriorate the sample through, e.g., radiation damage, because the spin flips are a reversible process.

[11] With kind permission from Alan Brash (Vanderbilt University School of Medicine) and PNAS.

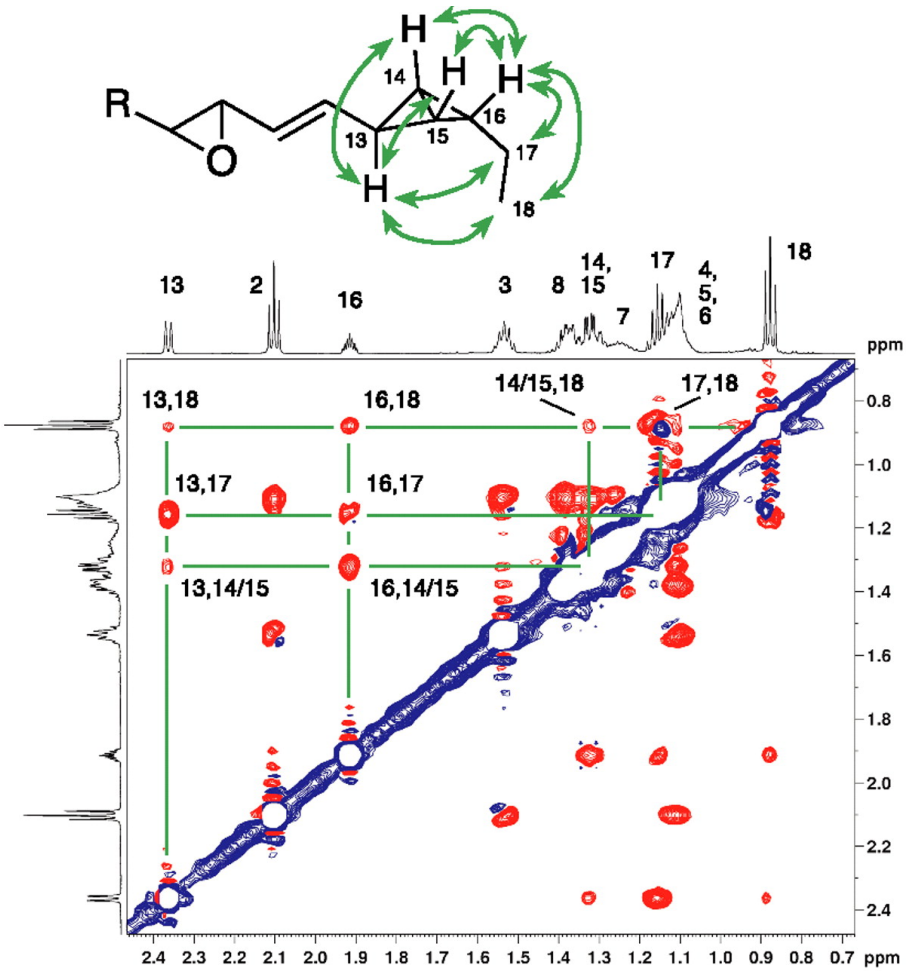


Figure 2.5: Exemplary NOESY NMR spectrum of the bicyclobutane ring of product 1 [70][11]. In contrast to X-ray crystallography, hydrogen atoms are also measured.

Kinetic changes of the molecules or reaction-induced shifts in the ensemble population can be traced by observing the change of the resonance peaks over time. Similarly, the flexibility of the ensemble can be derived from the width of the peaks.

On the other hand, NMR techniques have drawbacks. Only proteins with less than 100 kDa (approx. 1000 residues) molecular weight can be determined, although in some special cases, complexes with > 1 MDa have already been resolved with NMR [72, 73]. Proteins must be in solution and should not aggregate up to a concentration of at least 1 mM [68]. In order to measure sufficient signal, a high protein concentration is needed which is a typical bottle-neck of the method because protein synthesis in large quantities is challenging.

Interpreting NMR data can be challenging because the spectra of large proteins are complex with many overlapping signals and mapping the spectroscopy peaks to the inter-atomic distances requires good models [74]. Also, the superimposition of different conformation in the measurement makes *de novo* structure determination difficult, especially for flexible proteins. In the extreme case of a very flexible disordered protein, no resonance peaks are visible.

2.2.3 Cryo Electron Microscopy

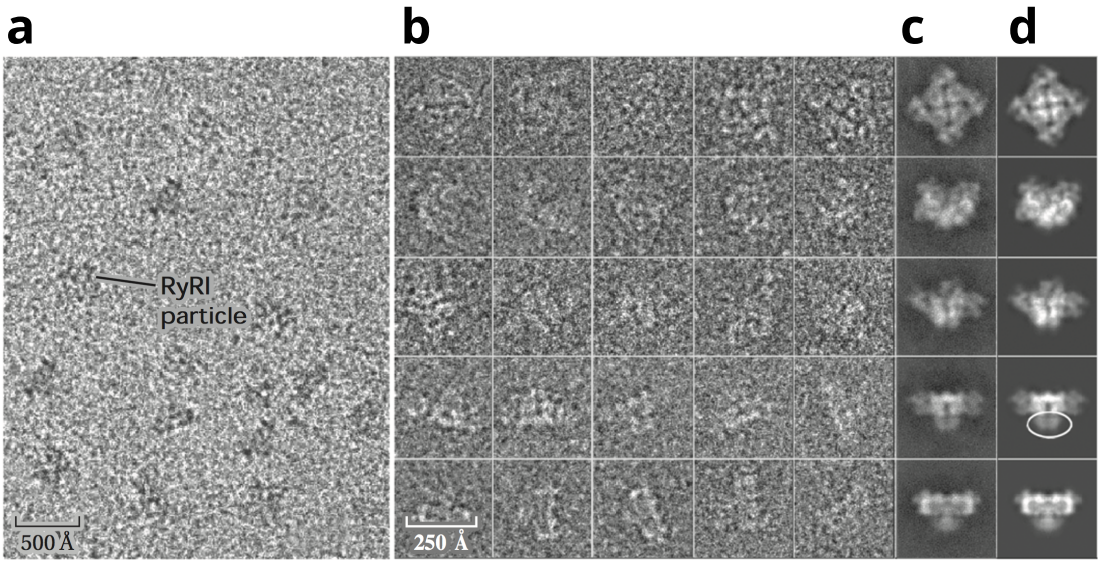


Figure 2.6: An example of single-particle structure determination using cryo-EM [75][12]. **a**: Exemplary micrograph image of ryanodine receptor 1 (RyR1) particles in the raw phase-contrast image. **b**: Particles after post-processing (phase reversals of the contrast-transfer function) and orientation classification. Each row of particles corresponds to a similar orientation. **c**: Averages of five different orientations over ~ 100 images in each class. **d**: 2D projections of the reassembled 3D density map in comparison to (c). The final electron density was determined by Ludtke *et al.* [75] with a resolution of 9.6 Å from the noisy images, an impressive demonstration of the strength of probabilistic structure determination algorithms.

Electrons have been used since the 1920s to study the microscopic structure of matter. In contrast to X-rays, the wavelength of electrons is much shorter (0.0197 Å for 300 keV electrons vs. 1.0Å for hard X-rays). Electrons are scattered

[12]With kind permission from Fred Sigworth (Yale University School of Medicine) and the APS Journal.

by the positive potential of the protons of the atoms (Coulomb interaction), while X-rays interact with the electron cloud. As the result, the interaction of electrons with matter is much stronger (10^6-10^7 times higher) than that of X-rays, although multiple scattering events become a problem.

The first proof-of-principle electron microscope was demonstrated by Ruska and Knoll in 1931, allowing four-hundred-power magnification, and eventually even exceeding the resolution of optical light microscopy two years later [76]. In 1937, Manfred von Ardenne pioneered the scanning electron microscope for which he scanned the specimens with a raster pattern to build up the full highly-resolved image, a method that he already successfully used for rastering the images in televisions or electronic cameras [77, 78].

Early on, the microscopes were used to image biological specimens but the radiation damage required the sample to be cooled down to cryogenic temperatures [79, 80], with a positive side effect that the evaporation of the solvent is also avoided.

In the experiment, a purified solution of the biomolecules is spread on a thin holey carbon film such that a thin liquid layer is formed across the holes in the carbon film. The film is then plunge-frozen in liquid ethane cooled by liquid nitrogen to preserve the native structure to the atomic level, to prevent dehydration of biological samples within the vacuum of an electron microscope and to reduce the effects of radiation damage. The film is then imaged using an electron beam yielding 2D projections of multiple randomly oriented particles.

The images have a very low signal-to-noise ratio, mainly due to the additional water (and other solvents) around the molecule and the limited electron exposure that is tolerated before radiation damage becomes too severe [81]. Determining both, the correct orientation and translation of each individual particle within a single image is required to average the respective projections and to reconstruct the 3D Coulomb potential density map, similar to the electron density map in X-ray crystallography (see Fig. 2.6). This requires, however, that the individual molecules in the images are in identical or similar conformations.

To this day, only 1650 or $\sim 1\%$ of the protein structures deposited in the Protein Data Bank originate from cryo-EM experiments. However, with a steep increase in the yearly deposited high-resolution structures since 2013 (see Fig. 2.3), it is expected that cryo-EM will produce a larger fraction of determined structures in the coming years. This increase is mainly attributed to the development of new generation of electron detectors and the improvement of the highly-parallel probabilistic structure determination algorithms [82] that handle the extremely low signal-to-noise ratios (see Fig. 2.6a for the high noise in the images of RyRl).

Cryo-EM has demonstrated enormous potential in determining large biomolecular structures such as the ribosome up to atomic resolution of <3.0 Å [79, 83, 84]. The method approaches X-ray crystallography in terms of resolution in some cases and can be used to determine atomic structures of macromolecules for which

crystallization has so far been unsuccessful or which are difficult to crystallize in specific functional states.

Cryo-EM experiments record real space images which contain both the amplitudes and the phases and therefore eliminate the phasing problem. The resolution of structure increases with the number of particle images (in analogy to the size of a 3D crystal) because the accuracy of image alignment is increased with every image. Therefore the method requires a high computational effort and many data, both of which is addressed by a high degree of automization. Especially in contrast to X-ray crystallography, where both the growth of the large crystals and the limited beam time at large syncrotrons and free electron lasers are major limiting factors, cryo-EM experiments on biomolcules can be carried out with less effort at any research site with (comparably) inexpensive electron microscopes.

Since cryo-EM is also a single molecule method and the whole ensemble of structures in equilibrium is imaged, information on structural heterogeneity and kinetics is accessible [84, 85] at the post-processing stage. Here, the images are not just sorted into orientational classes but also into different conformational states which are then linked to the states of the in-vivo dynamics (equilibrium – thermodynamics). Each structure usually corresponds to the free energy minimum in the respective part of phase-space [86]. If the molecule is flexible and many conformations are present in the ensemble, however, the determination of orientation and conformational classes at the same time becomes challenging, posing the limit on the time resolution for the dynamics.

Despite the many advantages of cryo-EM, the problem remains that the plunge-freezing may not be fast enough to avoid conformational changes due to the cooling and as a result, unphysiological conformations are imaged. Also, molecular movies of induced reactions, as proposed by numerous serial nanocrystallographic experiments, will most-likely not be possible with the frozen specimens in cryo-EM experiments.

The energy of the electrons used for imaging (80-300 kV) is high enough that covalent bonds are broken and the radiation damage destroys the samples much faster than e.g., in X-ray crystallography, decreasing the signal-to-noise ratio over time. Although structure determination methods handle the extreme signal-to-noise ratios very well, the extraction of the single particle images from the background remains challenging for small biomolecules.

2.3 Single Molecule X-Ray Scattering

Despite the great effort in structure determination, the structures of only about 0.75% ($\sim$ 132,000) of the more than 18 million known proteins [6] have been determined to high resolution [7]. Over the past years the yearly number of new structure depositions have been stagnating at 10,000 structures from X-ray crys-

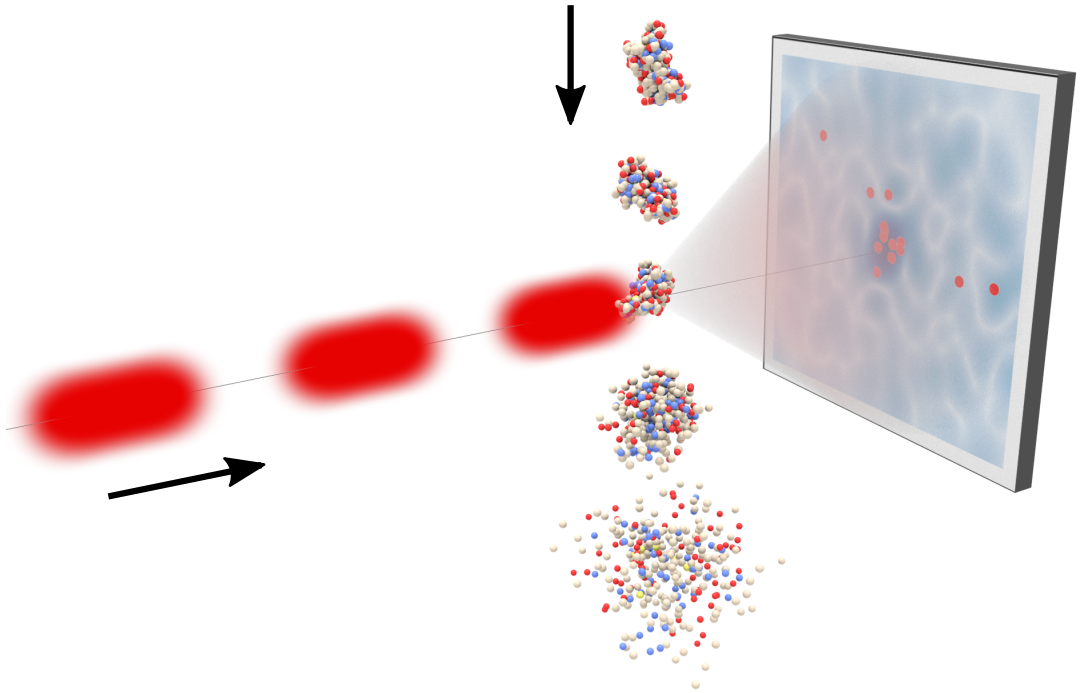


Figure 2.7: Sketch of a single molecule scattering experiment. A stream of randomly-oriented particles is injected into the high-intensity short-pulsed FEL beam, hit sequentially by femtosecond X-ray pulses, and the few coherently scattered photons (red dots) are recorded on the pixel detector. The spatial distribution of the photons follows the Fourier intensity of the molecule which is depicted here in light blue in the background of the photon pattern. After illumination, ionization effects charge the molecules and the resulting Coulomb forces quickly disintegrate the molecule. Note that the size of the FEL beam, the size of the detector and the distance between particle and detector are not shown to scale for visualization purposes.

tallography and even regressing below 1000 structures from NMR, as shown in Fig. 2.3. The growth of existing structure determination methods, except cryo-EM, seem to have reached a fundamental limit and new approaches are required to extend the knowledge of biomolecular structures.

To this end, X-ray scattering experiments with single biomolecules have been proposed by Neutze *et al.* as a novel *de novo* structure determination approach for proteins without the need for crystallization [8–10, 12, 14]. Single molecule X-ray imaging becomes possible due to newly-developed free electron laser that produce very high-intensity femtosecond-short X-ray pulses with a focus size of

down to 100 nm (see Sec. 2.3.1). Here, despite the much higher cross-section of electron scattering, photons are used instead of electrons, because due to the diverting Coulomb forces between the electrons, the electron beams, in contrast to the X-ray beam, cannot be focused sharp enough to achieve the fluence required for single-shot molecule imaging.

As sketched in Fig. 2.7, in the experiment, a stream of (typically) hydrated and randomly oriented proteins enters the pulsed X-ray beam at a rate of one molecule per pulse. Despite the high photon flux of the incident beam, only a few photons are scattered by the molecules and recorded on the pixelized detector (see Sec. 2.3.2 for the theory of image formation in coherent scattering experiments).

Sample delivery is non-trivial due to the nanoscopic size of the biomolecules and several solutions have been proposed, e.g., using electrospraying techniques [87], gas focused liquid jets [88], oil/water droplet immersion jet [89] or embedding the molecules into polymers to save material (lipidic cubic phase injector) [90]. In each sample delivery method, it is important that the single molecules stay in their physiological environment in order to observe the natural conformations.

In the scattering process, ionization (Auger decay) charges the atoms in the molecule and leads to Coulomb explosion, coining the method as a "diffract and destroy" experiment. However, the short pulses, usually less than 100 fs long, outrun the severe radiation damage because the molecular motion in response to the changed electronic configuration is estimated to take longer than 100 fs [8, 91] and the incident photons are scattered by the unperturbed structure before the molecule degenerates.

Like in conventional X-ray crystallography, only the intensities and not the phases are measured. However, due to the absence of crystals, the measured signal is the continuous Fourier transformation of the molecule, rendering the phase problem accessible to established *ab initio* phase-retrieval methods.

Whereas previous X-ray sources, including synchrotron sources, have primarily engaged in studies of static structures, X-ray FELs are by their nature suited for studying dynamic systems at the time and length scales of atomic interactions. In contrast to methods that measure a structure ensemble (NMR, SAXS, FRET), this method gives access to single molecule images and, with a seed model, the images could be e.g., sorted probabilistically to distinguish between different native conformations. Further, similar to nano-crystallography, in systems where reactions can be easily induced, e.g., by light, a sequence of structures at different reaction times may be recorded which opens the window to molecular movies as a long-standing dream [16]. Even without sorting, the variance of the native conformations can be assessed via the variance of the determiend electron density in which flexible regions would be smeared out more than rigid protein motifs.

2.3.1 Free Electron Laser

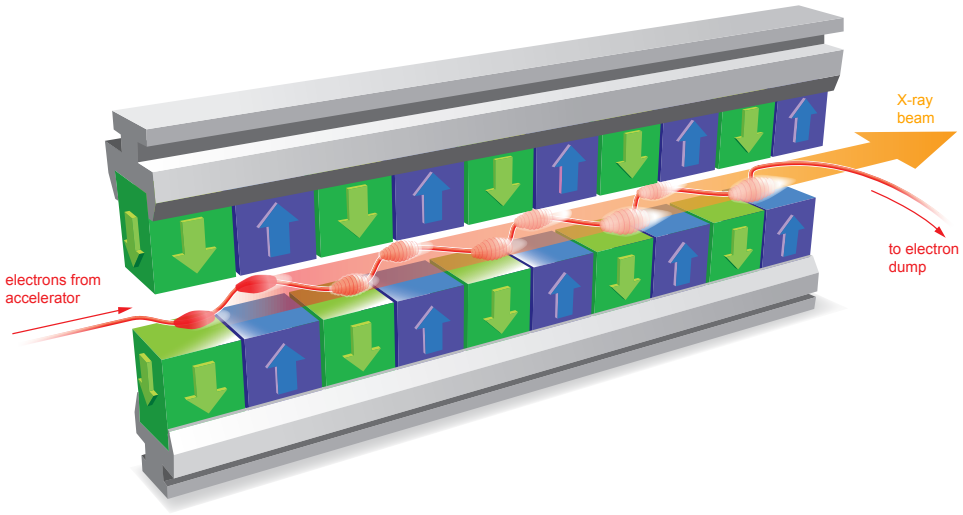


Figure 2.8: Sketch of the undulator of a free electron laser[13]. The electrons beam (red) enters the undulator, which consists of alternating magnets (green and blue), and is forced on a sinusoidal "wiggling" motion transversely to the beam axis. Due to this motion, X-ray photons are emitted in the direction of the beam that interact with the electrons and further increase the formation of bunches, an effect called self-amplified spontaneous emission (SASE). As a result, a very high-intensity short-pulsed X-ray beam is emitted.

Single molecule scattering experiments only have become possible because of the development of very powerful free electron lasers (FELs) which were initially invented in 1971 by John Madey in 1971 at Stanford University [92]. Today, they are the brightest available X-ray sources and have many applications beyond single molecule scattering on biomolecules.

The laser consist of two major parts [93, 94]. In the first part of the apparatus, an electron gun, produces a beam of electrons which is accelerated linearly to relativistic speed.

In the second part, the beam traverses a periodically alternating magnetic field created by the undulator magnets (see Fig. 2.8). Due to the Lorentz force of the magnetic field, the electrons are brought onto a sinusoidal path about the axis of the beam which results in the release of monochromatic incoherent photons. The electric field of the photons then interacts with the electron current which leads

[13]With kind permission from the European XFEL GmbH. Source: `http://www.xfel.eu/overview/how_does_it_work/`.

to an energy loss or gain of the electrons, depending on the phase of the photons. Eventually microbunches of electrons are formed which themselves emit more coherent photons. This process is called *self-amplified spontaneous emission* (SASE) and eventually it results in an exponential increase of emitted radiation power which leads to high beam intensities and laser-like properties. The wavelength λ_r of the radiated X-rays,

$$\lambda_r \propto \frac{\lambda_u}{2\gamma^2}, \tag{2.4}$$

is determined by the distance between the undulator magnets λ_u (the spatial period of the magnetic field), the relativistic Lorentz factor γ and a proportionality constant which depends on the undulator geometry[14]. In contrast to conventional lasers, the X-ray pulse is produced by a single pass of radiation through the undulators because no mirrors are available that can reflect the X-rays as resonant cavities. The pulse length, defined as the full-duration at half-maximum, varies between a few and about 100 fs [94, 95]

The first high intense free electron laser was demonstrated with the Lineac Coherent Light Source (LCLS) at SLAC, Stanford with a record short wavelength of 1.5 Å in 2009 [96]. The European XFEL at DESY, which came into operation in 2017 and has a total length of 3.4 kilometers, is currently the most powerful X-ray free electron laser available [97]. It reaches up to 27.000 pulses per second, electron energies of 17.5 GeV, a minimum wavelength of $\lambda = 0.5$ Å and a peak brilliance of $5 \cdot 10^{33}$ (photons / s / mm2 / mrad2 / 0,1% bandwidth) which is "a billion times higher than that of the best conventional X-ray sources".

The XFEL soon will be complemented by the equally powerful LCLS-2 at Stanford [16, 98] which starts operation in the early 2020s. Other hard X-ray FELs are available at slightly lower brightness, e.g., SACLA at RIKEN Harima Institute in Japan and SwissFEL at the Paul Scherrer Institute in Zürich, Switzerland and two soft X-ray FELs, FLASH and Fermi, are also in operation at DESY in Hamburg, Germany, and in Sincrotrone Trieste, Italy.

Beyond single molecule scattering experiments, the next-generation FELs will presumably used in many serial nano-crystallography experiments for *ab initio* biomolecular structure determination and imaging of molecular movies. Beyond these atomic structure determination, FELs will also be used for imaging chemical and structural processes over a wide range of length and time scales of other organic and inorganic specimens. This includes imaging and modifying matter in extreme environments, imaging nanoscale materials, heterogeneity and fluctuations and observing emergent phenomena in quantum materials through a range of scattering (e.g., time-resolved and high-resolution resonant inelastic X-ray scattering) or spectroscopy (time-resolved photoemission or nonlinear X-ray spectroscopy) ex-

[14] `https://en.wikipedia.org/wiki/Free-electron_laser`

periments. I recommend the LCLS-II proposal [16] for more detailed information on possible applications of FELs.

2.3.2 Coherent X-ray Scattering on Biomolecules

The images in single molecule X-ray scattering experiments are formed by the scattering of the photons on the electron density of the atoms within the molecules.

Often the terms diffraction and scattering are used interchangeably in the context, but while "diffraction" describes the directional change of the light wave and its interference, "scattering" is used in the context of single photons and refers to the momentum change, possibly in all directions. In the experiment, the X-ray pulse is comprised of over 10^{12} photons and the mathematical derivation of coherent scattering can be described in both pictures (scattering and diffraction).

Here, I use the wave formulation to derive the coherent scattering amplitudes but in the following chapters about photon correlations I will use the photon formulation, because due to the low signal, only single photons are recorded in the experiment. I will derive the elastic X-ray scattering in the far field (*Fraunhofer diffraction*) in the absence of other resonates or multiple scattering events and under the assumption that the molecules size is much smaller then the distance between the molecule and the detector. I further assume that, first, there is no reflection/interference at the particles interface and second, there is not phase change within the particle (coherence). The derivation presented here mainly follows Ref. [99–102].

In elastic scattering processes, the wavelength λ of the incoming and outgoing light waves is the same, i.e. the energy of the wave is conserved, and only the direction of the light wave changes. Here, I will describe the incoming and scattered waves with the wave vectors $\mathbf{k}_i$ and $\mathbf{k}_s$, respectively, each with length $|k| = 2\pi/\lambda$. Further, the oscillating field of the incoming light is described as a complex planar wave,

$$E_i(\mathbf{x}) = A \cdot \exp(i\mathbf{k}_i \cdot \mathbf{x}). \tag{2.5}$$

If the incoming planar wave is scattered at two different points $\mathbf{x}_0$ and $\mathbf{x}$ in the same direction $\mathbf{k}_s$ as depicted in Fig. 2.9a, the diffracted waves arrive at a point far away from the object with a path difference $\delta l = \delta l_1 + \delta l_2$ or the corresponding phase difference $\Delta\varphi = 2\pi\delta l/\lambda = -\mathbf{k}\cdot\mathbf{x}$. Here, the path and phase difference depend on the scattering vector $\mathbf{k}$ which is defined as the difference between the incoming and the scattered wave vectors, $\mathbf{k} = \mathbf{k}_i - \mathbf{k}_s$ (see Fig. 2.9b). The expression for the phase shift only holds if we assume that the (almost) parallel scattered

[15] Inspired by `http://www.rodenburg.org/theory/Ewaldsphere21.html`, `https://www.slideshare.net/lsinstruments/light-scattering-fundamentals`, `https://en.wikipedia.org/wiki/Coherent_diffraction_imaging`, `https://en.wikipedia.org/wiki/Diffraction#General_aperture` and Ref. [102]

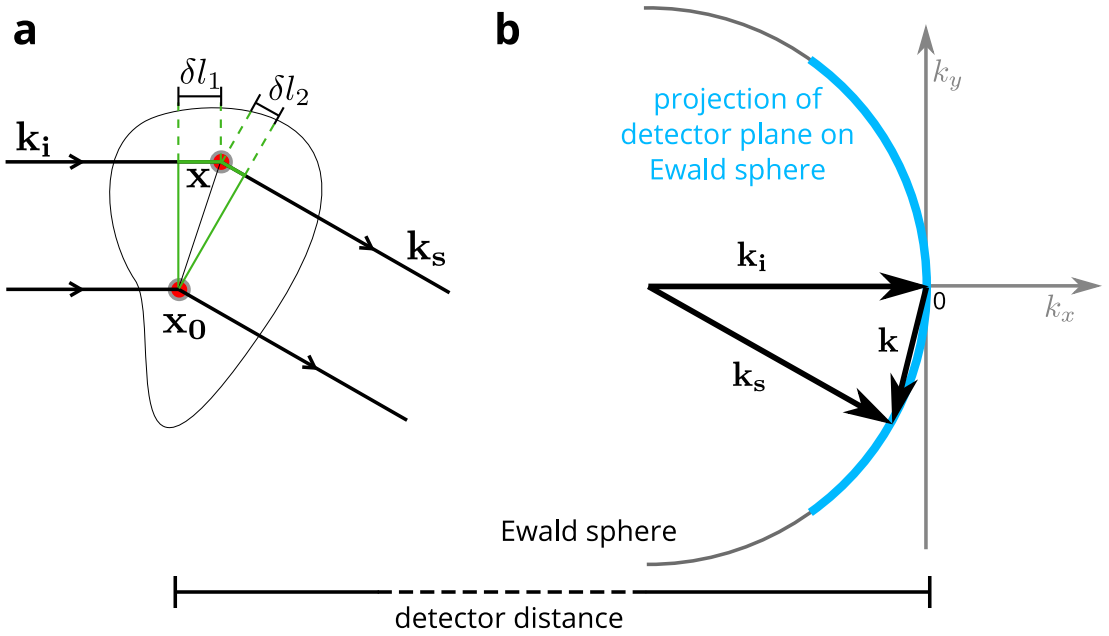


Figure 2.9: Sketch of coherent scattering on multiple atoms within a biomolecule[15]. **a**: The incident wave, denoted by the wave vector $\mathbf{k}_i$, is scattered at multiple atoms (red) in the direction θ, denoted by the wave vector $\mathbf{k}_s$. The path difference between two waves that scatter at atoms $\mathbf{x}_0$ and $\mathbf{x}$ amounts to $\delta l_1 + \delta l_2 = \delta l$ which equals a phase shifts $\Delta\varphi = 2\pi\delta l/\lambda = -\mathbf{k}\cdot\mathbf{x}$. **b**: The scattering vector $\mathbf{k} = \mathbf{k}_i - \mathbf{k}_s$ for different scattering angles lie on the Ewald sphere with radius $k = 2\pi/\lambda$. Each scattering image follows the intensity in the projection of the detector plane onto the Ewald sphere (blue).

waves from the two close points in the object plane intersect, after propagating the long detector distance, at point $\mathbf{k}$. The phase difference is the essential part in understanding interference of multiple scattering sources.

Instead of just two scattering sources, the electron density $\rho(\mathbf{x})$ of a protein is a continuous scattering volume and on the detector, the scattered waves of all the infinitely small scattering volumes interfere. In order to express the total sum of these waves, each infinitely small portion of the electron density is considered as the source of a new wavelet (spherical wave) according to the Huygens-Fresnel principle. The amplitudes of these waves depends on the local electron density and the contributions of the wavelets at any given point $\mathbf{k}$ on the detector is given by

$$\mathrm{d}E(\mathbf{k}) = \underbrace{\rho(\mathbf{x})\mathrm{d}\mathbf{x}}_{\text{source strength}} \underbrace{\frac{\exp(-i\Delta\varphi(\mathbf{k}))}{r(\mathbf{k})}}_{\text{Huygens spherical wave}} . \tag{2.6}$$

Here, d$\mathbf{x}$ is the very small volume element from which the wavelet originates. Both the radius r and the phase shift $\Delta\varphi$ depend on the position $\mathbf{k}$ but only the former can be approximated as the detector distance, $r \approx d$, because it is much larger than the object size, $d \gg |r_{\text{object}}|$.

On any point on the detector $\mathbf{k}$, the diffracted waves of all atoms interfere and the respective contributions of $\mathrm{d}E(\mathbf{k})$ (in particular the phase differences) are added up. In the limit of infinitely small volumes, this sum converges to an integral which is identical to the Fourier transform of the electron density $\rho(\mathbf{x})$ (omitting the $1/d$ factor),

$$A(\mathbf{k}) = \int \rho(\mathbf{x}) \exp(-i\mathbf{k}\cdot\mathbf{x})\mathrm{d}\mathbf{x} \tag{2.7}$$

Since the detector only records the absolute square of the scattering amplitudes $A(\mathbf{k})$, the measured signal (*intensity*) reads

$$I(\mathbf{k}) = J_0 r_e^2 P(\mathbf{x}_i, \mathbf{k}_s, \mathbf{u}) \left|A(\mathbf{k})\right|^2 \Delta\Omega, \tag{2.8}$$

with the fluence of the incident X-ray beam J_0 (photons / area), the Thomson scattering cross-section r_e^2 and the solid angles $\Delta\Omega$ of the detector pixel (considered to be small) [99].

The polarization $P(\mathbf{x}_i, \mathbf{k}_s, \mathbf{u})$ ($P(\mathbf{x}_i, \mathbf{k}_s, \mathbf{u}) = 1 - (\mathbf{k}_s \cdot \mathbf{u})^2$ for linearely-polarized light and $P(\mathbf{x}_i, \mathbf{k}_s, \mathbf{u}) = 1/2\left(1 + (\mathbf{k}_i \cdot \mathbf{k}_s)^2\right)$ for unpolarized light) is usually neglected because it can be extracted from the signal in the data analysis step. The expression for the scattering amplitudes can also be elegantly derived from first principles in quantum mechanics, see Ref. [103, 104].

Due to the energy conservation in elastic scattering experiments, the points on the planar detector are actually projections from a 2D sphere in the three-dimensional reciprocal space for which the condition $\mathbf{k} = \mathbf{k}_i - \mathbf{k}_s$ is met (see Fig. 2.9b). This Ewald sphere has a radius $k = 2\pi/\lambda$ and intersects with the origin of reciprocal space. In this thesis, however, I assume that the radius of the Ewald sphere is large due to the small wavelength of the X-ray $\lambda \sim 1.0$ Å, and for small scattering angles the slices in Fourier space are approximately planar (Note, that in Fig. 2.9b the radius of the Ewald sphere is depicted small for visual reasons).

According to the *Abbe diffraction limit*, the minimum distance that can be resolved in a scattering experiment is proportional to the wavelength (omitting numerical aperture),

$$\begin{aligned} d_{\min} &\leq \lambda \\ &= \frac{2\pi}{k_{\max}}. \end{aligned} \tag{2.9}$$

The resolution therefore is limited by the maximum wave number $k_{\max}$ for which scattered photons are recorded in the experiment.

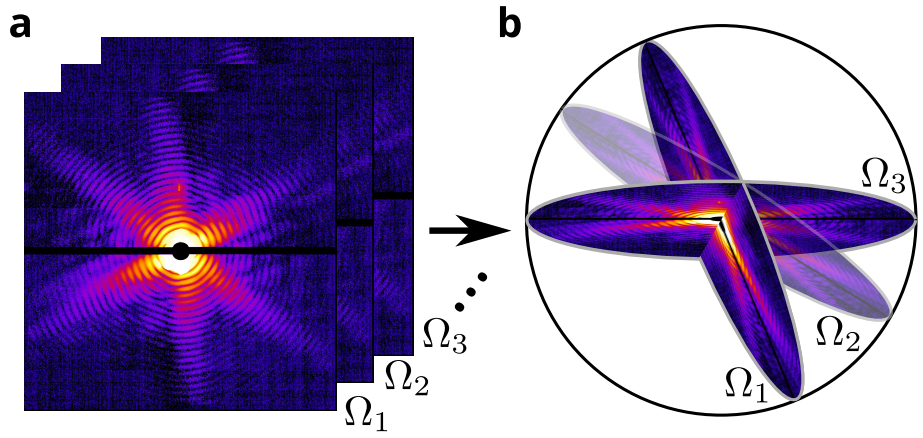


Figure 2.10: **a**: In each scattering snapshot, the molecule is imaged at a fixed orientation. **b**: If sufficiently many images with strong signals are recorded and the orientations $\{\Omega_i\}$ at the time of illumination are known, then the 3D intensity can be assembled from the images by projecting them into Fourier space.

The molecular orientation determines the orientation of the Ewald sphere (approx. planar here) in Fourier space because the rotation $\mathcal{R}$ and the Fourier transform commute, $\mathcal{R}\left[\mathcal{F}\left(\rho(\mathbf{x})\right]\right) = \mathcal{F}\left[\mathcal{R}\left[\rho(\mathbf{x})\right]\right]$. However, the molecule's orientation at the time of the illumination is unknown and the photon counts at the pixels therefore cannot be trivially mapped to the originating positions $\mathbf{k}$ in Fourier space to reassemble the Fourier density (as shown for the optimal case with known orientations in Fig. 2.10). In contrast to X-ray crystallography, $I(\mathbf{k})$ is continuous for single molecule scattering, rendering the phase problem accessible to established methods [105–108] (see Appendix A.1.3).

2.3.3 Estimation of the Number of Scattered Photons and Data-Acquisition Times

In the experiment, the photons are not just scattered coherently (Rayleigh) but also incoherently (Compton scattering) [19] by bound- and free-electrons. In fact, scattering images are expected to contain a substantial amount of noise from various sources including from inelastic scattering, the photo-electric effect, detector noise and the water-shell or other bulk material[8]. Theoretically, a signal-to-noise ratio of 1 to 10 is expected from scattering cross-sections of the photo-electric effect but early experiments on the mimivirus show a much lower noise-level to an extent that the noise is not addressed in the reconstruction. The lack of photo-electric photons can in part be attributed to the very short beam pulses [91] that outrun

the radiation damage. Therefore, I have only considered inelastic (incoherent) scattering processes so far.

Photon Energy [keV]	**C [barns/atom]**		**H [barns/atom]**		**O [barns/atom]**	
	COH	INCOH	COH	INCOH	COH	INCOH
1.0	21.57	0.2519	30.03	0.256	39.89	0.2262
2.0	16.69	0.7702	24.33	0.8165	33.55	0.756
5.0	7.219	1.984	11.1	2.225	16.4	2.323
8.0	4.205	2.497	6.244	2.858	9.145	3.139
10.0	3.248	2.697	4.722	3.093	6.805	3.439

Table 2.1: Coherent and incoherent atomic scattering cross-sections σ for multiple photon energies [109] in [barns/atoms] = [10^{-28}m^2/atoms].

Due to the limited fluence the experiment, the scattering images are a sparse sampling of the planar cuts with noise according to a Poisson distribution (also called "shot noise"). The total number of scattered photons can be approximated with a simple model of the biomolecules consisting of only J carbon, nitrogen and oxygen atoms as

$$N_{total} = \frac{N_{\text{photons}}}{(d/2)^2\pi} \sum_{j=1}^{J} \sigma_j N_j. \tag{2.10}$$

This assumes, as mentioned earlier, that each atom scatters independently of the other atoms such that the individual contributions can be added up. The scattering cross-sections σ for coherent X-ray scattering have been measured to high-precision and are denoted in Table 2.1. At the European XFEL, the X-ray beam will be comprised of $N_{\text{photons}} = 5.0 \cdot 10^{11}$ photons at 5 keV beam energy (according to the simulation framework for the European XFEL [19]) focused into a spot with diameter $d = 100$ Å.

In this thesis, I will validate my structure determination method with the 46 residue Crambin protein which totals to 211 carbon, 57 nitrogen and 67 oxygen atoms. At the given beam parameters of the XFEL mentioned above, I estimated an average of 20 coherently and 5 incoherently scattered photons per Crambin shot (25% noise-level) for this relatively small biomolecule. A similar number of photons (16) was obtained using the SimEx simulation framework for imaging single particles at the European XFEL using realistic beam profiles [19, 20]. In Section 4.7, I will discuss the photon distribution of incoherently scattered photons which follows the Klein-Nishina differential cross-section.

The number of photons n at each pixel position $\mathbf{k}$ is Poisson-distributed,

$$P(n \text{ photons}, \mathbf{k}) = e^{-\lambda} \frac{\lambda(\mathbf{k})^n}{n!}, \tag{2.11}$$

with the average number of photons $\lambda(\mathbf{k}) \sim I(\mathbf{k})$, given by the intensity, which also sets the variance of the distribution.

3 Existing Single Molecule X-Ray Scattering Analysis Methods

The structure determination from sparse single molecule scattering images faces two major challenges, first, the unknown orientation of the molecule at the time of illumination and, second, the low number of scattered photons along with the additional background noise, resulting in low signal-to-noise levels. Over the past years, several structure determination methods have been proposed and demonstrated which mainly fall into two major classes.

The first class of methods predicts the orientation of the molecules at the time of illumination for each scattering image either explicitly or implicitly e.g., through statistical similarities between images or by using a coarse seed model. Images that belong to the same orientation are averaged and these averages are assembled into the 3D intensity similar to cryo-EM (see Fig. 2.10). However, almost all of the orientation classification methods are limited to scattering datasets with more than 100 average photons per image.

The second class of methods forgoes the classification of orientations by using photon correlations as an averaged summary statistics of the entire image dataset that is independent of the individual orientations. So far, these methods are either limited by low photon counts, require a seed model or are only applicable to molecules which are highly symmetric or only randomly rotated about a single axis such as membrane molecules.

Here, I give a brief overview on the current state of research within the two approaches. For a detailed discussion of photon correlations, the main subject of this thesis, please read up in the following Chapter 4.

3.1 Methods Classifying Pattern Orientations

Common Line Approach

A common approach is to estimate the orientation of the molecule at the time of illumination explicitly. Images corresponding to the same orientation and Fourier slice, respectively, are averaged and the differently-oriented continuous images are projected back into the 3D Fourier space to reassemble the diffraction volume.

Huldt *et al.* [10], Shneerson *et al.* [29] and Bortelt *et al.* [110] compared two patterns using only their *common line* of intersection (see Fig. 2.10) which allows to orient the patterns with respect to each other. This technique has successfully been used in electron microscope imaging but it requires more than 10 photons per detector pixel and image which is about three orders of magnitude larger than the expected scattering count in the experiment. Similar to cryo-EM, the discretization of the orientations requires up to 10^6 orientational classes to unambiguously determine a 3 Å resolution structure of a 100 Å molecule [29]. These initial averaging approaches have since been superseded by other methods that overcome the issue of low signal-to-noise by treating the data globally.

Manifold Embedding Methods

Manifold reconstruction algorithms forego the explicit reconstruction of the intensity and use the similarity between scattering patterns as a proxy for the similarity between orientations to reconstruct the manifold of orientations. Since these algorithms work solely on the manifold level they are not guaranteed to generate a self-consistent 3D intensity.

The first method was proposed by Fung *et al.* [37] and is based on a Generative Topographic Mapping (GTM), "a Bayesian nonlinear factor-analytical approach originally developed for data projection, visualization and neural network applications". The method relies on the fact that the Fourier intensity is smooth and continuous. The correlations between scattering images are used together with a maximum likelihood measure to map the images into three-dimensional space of possible orientations which itself is embedded in the N-dimensional space of pixels.

The manifold generation is done for a large number of diffraction patterns and the closest patterns are classified into orientational classes which are eventually used for averaging. In contrast to the common-line method, the Bayesian approach determines the orientations of the patterns with a global optimum.

However, statistical uncertainty is introduced when classifying the scattering images by orientation, and the method depends on the angular discretization of orientation space. The method works with a low scattering signal but still requires more than 100 average photons per image. As an example, Fung [37] claim to have reconstructed the 500 kD chignolin with a 1.8 Å resolution at $4 \cdot 10^{-2}$ photons per pixel.

More recently, diffusion map techniques have been developed to compute low-dimensional manifolds from XFEL diffraction data [39, 40]. These techniques allow to sort the images into more than three dimensions, incooperating other experimental variables such as sample heterogeneity or changing beam conditions. The algorithm determines the dimension of the manifold space autonomously which removes the human bias but makes the interpretation of the manifold challenging. One way to explicitly relate the data-space to rotations is via mapping geodesic

distances as demonstrated by Kassemeyer *et al.* in the imaging of 150 nm nanorice particle [33].

Recently, manifold-embedding was used to determine the structure of single icosahedral Rice Dwarf virus particles (70 nm diamter) using hard X-ray at LCLS to 6.7 Å resolution [111].

Expansion-Maximization-Compression (EMC) Algorithms

Loh and Elser have introduced the *Expansion-Maximization-Compression* (EMC) algorithm which maximizes the likelihood of an intensity model to fit a set of diffraction images using Bayes' theorem [30]. Bayes methods can handle low signal-to-noise levels and allow to incorporate information about the noise statistics.

However, also in this method, the orientations for individual diffraction images are determined, but in contrast to previous methods, the algorithm calculates for each image a probability (likelihood) for every possible orientation, accounting for statistical uncertainty in the classification of the orientation. Further, an intensity model is used which is iteratively updated in each step by averaging the aligned images in 3D reciprocal space. The method was used to determine the structure of a GroEL molecule at 20 Å resolution from 106 synthetic diffraction images.

Another EMC-like algorithm was proposed by Tegze and Bortel [112] which addressed the high computational effort by reducing the pattern comparisons and by assigning the orientations of the images through the best fit to the iteratively refined intensity instead of a probabilistic way.

Walczak and Grubmüller [31] developed a solution for the classification uncertainty that utilizes a seed model of the structure in real space to speed up convergence and to discriminate between different conformations. The seed model of the structure can be a low resolution version of the protein, obtained from NMR or X-ray crystallography, or a homology model. A variant of the method involves a Monte Carlo search in which the tertiary structure of polypeptides is determined by using the primary and secondary structure as a starting point and modifying bond angles. Very few photons per picture in the range of $N_{photons} \approx 10 - 100$ are sufficient to successfully reconstruct viable structure information, but the necessity of a seed model makes it impractical for the study of large protein structures because the search space grows combinatorially.

EMC algorithms are computationally expensive and the effort grows with the recorded scattering pattern because the overall probability depends on all images. Large datsets of $> 10^4$ images that are expected in the experiment will be challenging.

To this end, Flament *et al.* [32] improved the EMC method by using shell-wise spherical harmonics expansions, allowing the method to scale in terms of angular and radial resolution. Their method uses the hierarchical properties of the spherical harmonics expansion to resolve, first, the intensity with few patterns and

lower resolution and proceeding to higher resolutions and more patterns, increasing the convergence speed.

Ekeberg *et al.* have demonstrated in a proof of principle experiment that it is possible to determine a low resolution 3D structure of a Mimivirus based on XFEL single molecule scattering data using the EMC algorithm [18].

Donatelli *et al.* proposed a "multitiered iterative phasing algorithm to reconstruct structural information from single-particle diffraction data by simultaneously determining the states, orientations, intensities, phases, and underlying structure in a single iterative procedure" [36]. The method leverages real-space constraints on the structure to improve convergence from very few images (<100) with $\sim$ 8000 photons each, yielding a resolution of 5.5 Å.

3.2 Methods Using Photon Correlation

Photon correlations are a measure for how often certain angles between photon doublets, triplets, etc... and distances of the photons to the detector center are observed in the scattering images. Only photons within a single image are correlated but the correlation can be refined by collecting more images.

If the photons of individual doublets or triplets should were coherently scattered on a single particle, then these correlations encode structural information. The sum of all photon correlations throughout all scattering images present an orientation-independent summary statistic of the scattering data which can be obtained without sorting of the images into orientational classes beforehand.

In other fields such as fluorescence microscopy, time integrated and time-correlated single-photon counting [41] is successful when working with low signal-to-noise ratios.

Two-Photon Correlations

Saldin *et al.* [42] first proposed to use of photon correlations in the context of single molecule scattering experiments in 2009 with an approach in which two-photon correlations are used to reconstruct the shape function of the protein. However, the shape is a binary function which describes the nonzero regions of the electron density and cannot be used to retrieve the atomic positions of the molecule.

In a follow up paper, Saldin *et al.* [43–45] showas that the structure can be solved from two-photon correlations when the molecules only rotate about a fixed axis in the experiment. This approach is a feasible for membrane bound proteins (e.g. potassium or aquaporin channel proteins) *in situ* embedded in the membrane but it excludes a large group of solvated proteins. Two-photon correlations are also applicable for the determiantion of the full electron density if the studied particle

has a high degree of symmetry, such as an icosahedral virus as demonstrated by Saldin in another paper in 2011 [48].

Two-Photon Correlations in Fluctuation X-ray scattering (CXS)

Two-photon correlations have also been successfully used in fluctuation X-ray scattering (CXS) which is a completely different type of experiment similar to solution scattering. However, the methods and results are also interesting for single molecule scattering.

In conventional solution scattering, the orientational averaging that occurs during the X-ray illumination results in signal which carries only 1-dimensional (radial) intensity information and all angular information is averaged out. In CXS experiments, however, the X-ray pulses from synchronous or free electron lasers are much shorter than the orientational diffusion times of the molecules such that they appear to be fixed in space.

In each image multiple particles with different orientations are recorded and as a result speckle patterns emerge from which angular correlations are calculated [113]. Kam [46] pointed out that the correlated signal, when averaged over many images with multiple particles, converges to the two-photon correlation of a single biomolecule, and that this occurs even if an extremely large number of biomolecules are illuminated simultaneously in each shot. The information content of the resulting correlated scattering data set grows in proportion to the cube of the particle size, unlike isotropic SAXS data, where the information content grows linearly with particle size [99].

Mendez *et al.* [114] have determined two populations of nanoparticle domains from the two-photon correlations in solution, and in a more general approach, Kirian *et al.* demonstrated the wide use of correlated X-ray scattering (CXS) in the determination of 2D projections of particles [99].

Liu *et al.* demonstrated *ab initio* model reconstruction from fluctuation X-ray scattering profiles using two-photon correlation ("average angular auto-correlation") [115], however, at a relatively low resolution of 48 Å. The method represents the electron density on a grid in real space and in successive Monte Carlo steps, the density is locally perturbed by random dilations or an erosions. The expected two-photon correlations from the proposed densities are then compared with the correlations from the experimental fluctuation data and the difference is minimized.

Donatelli *et al.* has proposed to phase directly from the two-photon correlation in fluctuation X-ray experiments [113] to overcome the information deficiency in the two-photon correlations, but the method relies on prior information on the structure.

Degenerate Three-Photon Correlation

The presented two-photon correlation methods have demonstrated that, remarkably, two photons per image already carry important structural information, but as shown by numerous publications [43, 44, 46, 116], they are not sufficient to retrieve the 3D structure unambiguously (see Appendix A.1.2). This analytic observation was a motivation for my work to look into higher order correlations, in particular three-photon correlations.

The first analytic work on three-photon correlations has been done by Kam in 1980 in the context of electron microscopy. He demonstrated that in principle three photons per picture are sufficient to retrieve the Fourier density [46] by compensating the information deficiency in the two-photon correlation with additional information from the degenerate part of three-photon correlation, i.e., triplets in which two photons are recorded at the same detector position. However, Kam's method cannot universally be applied to few photon single molecule X-ray scattering data because the degenerate three photon events, which are only a small subset of the entire three-photon correlation, are not sufficiently sampled at the low photon counts (in contrast to the dense electron micrograph images).

Based on Kam's approach, Starodub *et al.* determined the structure of comparably large cylindrical particles (polystyrene spheres with a 91 nm diameter) at 200 Å resolution [47] and Poon *at al.* resolved the missing signs of the spherical harmonics coefficients of a highly symmetric icosahedral virus at 220 Å resolution [49].

In conclusion, all existing correlation-based approaches are either limited to symmetrical molecules or have only demonstrated low-resolution structure determination. No available approach so far has made use of the *full* three-photon correlation information available in the images.

4 The Three-Photon Structure Determination Method

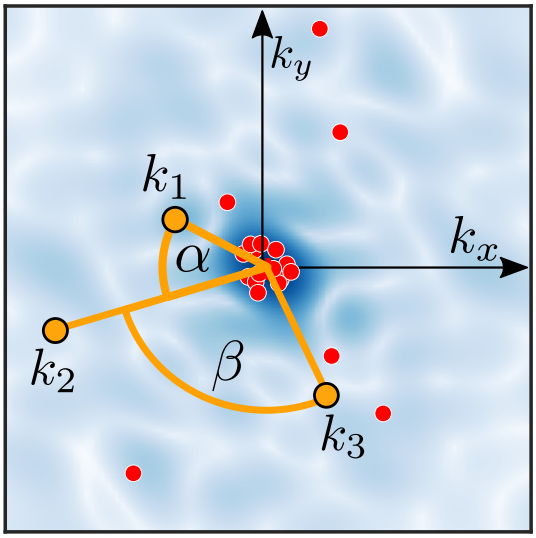


Figure 4.1: Schematic depiction of the three-photon correlation using an exemplary synthetic single molecule scattering images of Crambin with ca. 20 coherently scattered photons. In the detector plane $k_x k_y$ the recorded photons are grouped into triplets, each of which is characterized by distances k_1, k_2, k_3 to the detector center (orange lines) and the angles α and β between the respective photons (orange circular arcs).

In this Chapter I will present the three-photon correlation method that I have developed for the *de novo* structure determination from few photon single molecule X-ray scattering images. The main idea is to determine the intensity $I(\mathbf{k})$ from the *full* three-photon correlation $t(k_1, k_2, k_3, \alpha, \beta)$ which is accumulated from all photon triplets in the recorded scattering images as illustrated in Fig. 4.1.

A single triplet is characterized by the angles α and β between the photons and the distances of the photons to the detector center. Each triplet is comprised of three correlated doublets $(k_1, k_2, \alpha,)$, (k_2, k_3, β) and $(k_1, k_3, \alpha + \beta)$. The angles are chosen as the minimum difference between the pairs, $\alpha, \beta \in [0, \pi]$.

In the following Section I will derive an analytic expression of the full three-photon correlation $t(k_1, k_2, k_3, \alpha, \beta)$ as a function of the 3D intensity $I(\mathbf{k})$ using

shell-wise spherical harmonics (SH) expansions [117] for $I(\mathbf{k}) = \sum_{lm} A_{lm}(|\mathbf{k}|) Y_{lm}(\theta, \varphi)$. For a brief introduction to spherical harmonics expansions, see Appendix A.1.1.

The correlations, expressed in spherical harmonics terms, are faster to calculate than e.g., the numerical integration, and they allow for adapting the number $K(L^2 + 3L + 2)/2$ of spherical harmonics basis functions to the target resolution via the largest considered wave number k_{cut}, the number K of used shells between $0...k_{\text{cut}}$, and the expansion order L. The hierarchical properties of spherical harmonics basis functions further allow to determine the structure first with low angular resolution and then to successively refine it to higher resolutions and higher expansion limits, respectively (see Sec. 4.4 for implementation details).

4.1 Derivation of the Three-Photon Correlation expressed in Spherical Harmonics

The following derivation follows Kam [46], but further generalizes it to the full three-photon correlation. I assume that, due to the small wavelength of the X-ray photons, the Ewald sphere has a large radius and its intersections with the Fourier space density (from which the scattering images are sampled) are approximately planar.

The triple correlation $t(k_1, k_2, k_3, \alpha, \beta)$ is the orientational average $\langle\rangle_\omega$ of the product between three intensities $I(\mathbf{k})$ that lie in the same plane in Fourier space (see Fig. 4.2),

$$t(k_1, k_2, k_3, \alpha, \beta)_{I(\mathbf{k})} = \langle I_\omega(\mathbf{k}_1(k_1, 0)) \cdot I_\omega(\mathbf{k}_2(k_2, \alpha)) \cdot I^*_\omega(\mathbf{k}_3(k_3, \beta)) \rangle_\omega . \tag{4.1}$$

Here, without loss of generality, the $k_x k_y$-plane is chosen as the detector plane and the vectors $\mathbf{k_1} = (k_1, 0, 0)$, $\mathbf{k_2} = k_2(\cos\alpha, \sin\alpha, 0)$ and $\mathbf{k_3} = k_3(\cos\beta, \sin\beta, 0)$ on this plane are chosen as one arbitrary realization of the triplet $(k_1, k_2, k_3, \alpha, \beta)$, characterized by the angles α and β between the vectors and distances to the detector k_1, k_2 and k_3, respectively (see Fig. 4.1).

For the orientational average $\langle\rangle_\omega$ it is assumed that in the experiment the orientation of the molecule is unknown and uniformly sampled. Note that the orientational average can either be expressed as an average over all rotations of $I_\omega(\mathbf{k})$ for fixed $\mathbf{k}_{1,2,3}$ (our approach) or as an average over all rotations of the vectors $\mathbf{k}_{1,2,3,\omega}$ for a fixed $I(\mathbf{k})$.

Next, $I(\mathbf{k})$ is decomposed into spherical shells with radius k and each shell is expanded using a spherical harmonics basis [117],

$$I(\mathbf{k}) = \sum_{lm} A_{lm}(k) Y_{lm}(\theta, \varphi) . \tag{4.2}$$

The coefficients $A_{lm}(k)$ describe the intensity function on the respective shells and are non-zero only for even $l \in \{0, 2, 4, ..., L\}$ because of the symmetry of

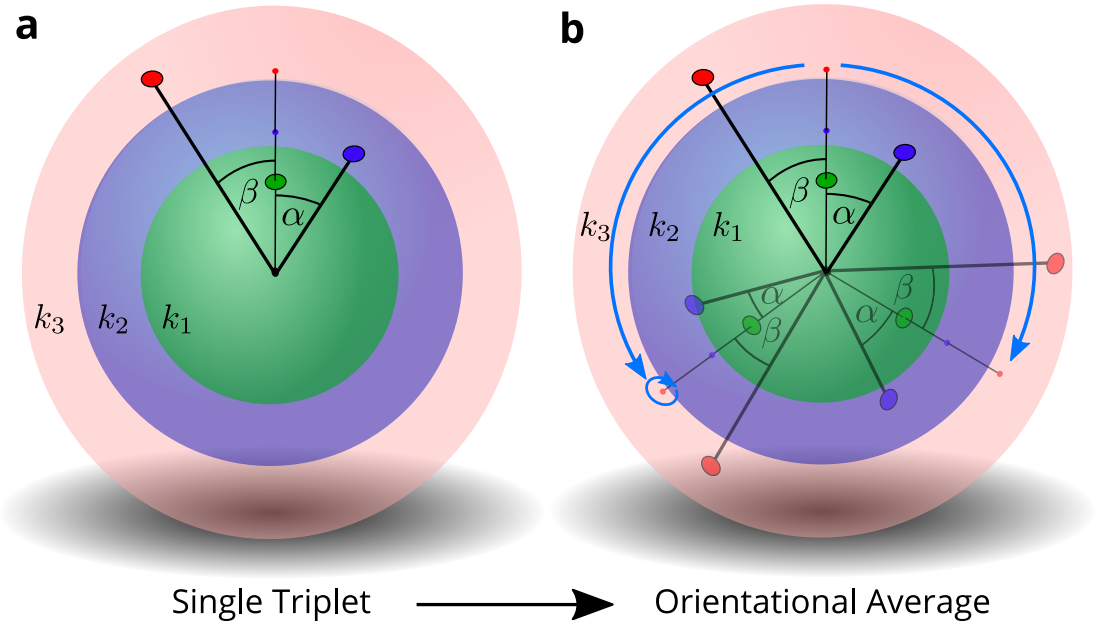


Figure 4.2: Sketch of the orientational average of the product of three intensities as a visualizing of the integration of the three-photon correlation. **a:** Three intensities in the 3D Fourier space on different shells k_1 (green), k_2 (blue) and k_3 (red) with angles α (green,blue) and β (green, red) forming one realization of the triplet $(k_1, k_2, k_3, \alpha, \beta)$. **b:** The orientational average rotates the triplets with all possible Euler rotations ω. Depicted with opacity are two possible rotations of the triplet $(k_1, k_2, k_3, \alpha, \beta)$. The right triplet is rotated about the z-axis (out of page) in clockwise direction and the left triplet is rotated counterclockwise and additionally rotated about the axis given by the reciprocal vector $\mathbf{k}_1$. Each of the three triplets correspond to different points in the 3D intensity and therefore contribute different intensity triple products (proportional to the probability of observing the triplet) to the integral of the three-photon correlation.

$I(\mathbf{k}) = I^*(-\mathbf{k})$ (Friedel's law). In this description, a 3D Euler rotation ω of $I(\mathbf{k})$ is expressed by transforming the spherical harmonics coefficients according to $A^{\mathrm{rot}}_{lm}(k) = \sum_{mm\prime} D^l_{mm\prime} A^{\mathrm{unrot}}_{lm\prime}(k)$, using the rotation operators $D^l_{m'm}$ which are composed of elements of the Wigner D-matrix as defined, e.g., in Ref. [117], yielding the rotated intensity,

$$I_\omega(\mathbf{k}) = \sum_{lmm'} A_{lm}(k)\, Y_{lm'}(\theta, \varphi)\, D^l_{m'm}(\omega)\,. \tag{4.3}$$

Inserting the spherical harmonics expansion of the rotated intensity $I_\omega(\mathbf{k})$, evaluated at positions $\mathbf{k}_1$, $\mathbf{k}_2$ and $\mathbf{k}_3$ in the $k_x k_y$ plane ($\theta = \pi/2$), into the expression for the three-photon correlation, Eq. (4.1), yields

$$\begin{aligned} t(k_1,k_2,k_3,\alpha,\beta)_{\{A_{lm}(k)\}} &= \sum_{l_1 l_2 l_3} \sum_{m_1 m_2 m_3} \sum_{m_1' m_2' m_3'} A_{l_1 m_1}(k_1) A_{l_2 m_2}(k_2) A^*_{l_3 m_3}(k_3) \\ & Y_{l_1 m_1'}\left(\frac{\pi}{2},0\right) \cdot Y_{l_2 m_2'}\left(\frac{\pi}{2},\alpha\right) \cdot Y^*_{l_3 m_3'}\left(\frac{\pi}{2},\beta\right) \\ & \left\langle D^{l1}_{m_1 m_1'} \cdot D^{l2}_{m_2 m_2'} \cdot D^{l3}_{m_3 m_3'} \right\rangle_\omega , \end{aligned} \tag{4.4}$$

such that the orientational average only involves the elements of the Wigner D-matrix $D^l_{mm'}$. Using the Wigner-3j symbols $\begin{pmatrix} l_1 & l_2 & L \\ m_1 & m_2 & -M \end{pmatrix}$ [118], the product of two rotation elements $D^l_{mm'}$ reads

$$\begin{aligned} D^{l_1}_{m_1 m_1'} D^{l_2}_{m_2 m_2'} &= \sum_{L=|l_1-l_2|}^{l_1+l_2} \sum_{MM'} (2L+1)(-1)^{M-M'} \\ & \begin{pmatrix} l_1 & l_2 & L \\ m_1 & m_2 & -M \end{pmatrix} \\ & \begin{pmatrix} l_1 & l_2 & L \\ m_1' & m_2' & -M' \end{pmatrix} D^L_{MM'}. \end{aligned} \tag{4.5}$$

With the orthogonality theorem for orientational averages of the product of two Wigner D operators,

$$\left\langle D^L_{MM'} D^{l3*}_{m_3 m_3'} \right\rangle_\omega = \frac{1}{2L+1} \delta_{l_3 L} \delta_{m_3 M} \delta_{m_3' M'}, \tag{4.6}$$

the three-photon correlation finally reads

$$\begin{aligned} t(k_1,k_2,k_3,\alpha,\beta)_{\{A_{lm}(k)\}} &= \sum_{l_1 l_2 l_3} \sum_{m_1 m_2 m_3} A_{l_1 m_1}(k_1) A_{l_2 m_2}(k_2) A^*_{l_3 m_3}(k_3) \\ & \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & -m_3 \end{pmatrix} \\ & \sum_{m_1' m_2' m_3'} (-1)^{m_3 - m_3'} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1' & m_2' & -m_3' \end{pmatrix} \\ & Y_{l_1 m_1'}\left(\frac{\pi}{2},0\right) Y_{l_2 m_2'}\left(\frac{\pi}{2},\alpha\right) Y^*_{l_3 m_3'}\left(\frac{\pi}{2},\beta\right). \end{aligned} \tag{4.7}$$

This expression only involves sums of products of three spherical harmonics coefficients $A_{lm}(k)$ with known Wigner-3j symbols and spherical harmonics basis functions $Y_{lm}(\theta,\varphi)$. The numerical calculation of the three photon correlation is the computationally limiting step in the structure determination approach. I have devised a fast implementation using GPUs (CUDA) which I will discuss in Appendix A.2.2.

4.2 Structure Determination using Three-Photons

I was unable to invert the analytic expression of the three-photon correlation in Eq. (4.7), and the number of unknowns (e.g., 4940 for $K = 26$, $L = 18$) is too large for a straightforward numeric solution. Instead, I chose a probabilistic approach and asked which intensity $I(\mathbf{k})$ is most likely to have generated the complete set of measured scattering images and triplets, respectively. To this end, I considered the Bayesian probability p (with uniform prior) that a given intensity $I(\mathbf{k})$, expressed in spherical harmonics by $\{A_{lm}(k)\}$, generated the set of triplets, $\{k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i\}_{i=1...T}$,

$$p\left(\left\{k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i\right\}_{i=1...T} \middle| \{A_{lm}(k)\}\right) = \prod_{i=1}^{T} \tilde{t}(k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i)_{\{A_{lm}(k)\}}. \quad (4.8)$$

Due to the statistical independence of the triplets, this probability p is a product over the probabilities $\tilde{t}(k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i)$ of observing the individual triplets i which is given by the normalized three-photon correlation $\tilde{t}\,(k_1, k_2, k_3, \alpha, \beta)$. Here, $\tilde{t}\,(k_1, k_2, k_3, \alpha, \beta)$ was calculated using Eq. (4.7) for varying intensity coefficients $\{A_{lm}(k)\}$ and the coefficients that maximized $p\,(\{k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i\})$ were determined using a Monte Carlo scheme as described in Sec.4.4.

In contrast to the direct inversion, the probabilistic approach has the benefit of fully accounting for the Poissonian shot noise implied by the limited number of photon triplets that are extracted from the given scattering images. I note that this approach also circumvents the limitation faced by Kam [46], where only triples are considered, in which two photons are recorded at the same detector position. Because all other triples had to be discarded, Kam's approach is limited to very high beam intensities, and cannot be applied in the present extreme Poisson regime.

4.3 Reduction of the Search Space using Two-Photon Correlations

The high-dimensional search space was reduced by utilizing the structural information contained within the two-photon correlation. In analogy to the three-photon correlation, the two photon-correlation is expressed as a sum over products of spherical harmonics coefficients $A_{lm}(k)$ weighted with Legendre polynomials P_l [42, 46],

$$c_{k_1,k_2,\alpha} = \sum_l P_l\left(\cos\left(\alpha\right)\right) \sum_m A_{lm}\left(k_1\right)\left(\omega\right) A_{lm}^*\left(k_2\right). \quad (4.9)$$

The inversion yields coefficient vectors $\mathbf{A}_l^0(k) = (A_{l-m}^0, ..., A_{lm}^0)$ for all $l \leq L \leq K_{\max}/2$ and $-l < m < l$, as first demonstrated by Kam [46] (See Appendix A.1.2

for detailed inversion of the two-photon correlation). However, all rotations in the $2l+1$-dimensional coefficient eigenspaces of $\mathbf{A}_l^0(k)$ by $\mathbf{U}_l$ are also solutions,

$$\mathbf{A}_l(k) = \mathbf{U}_l \mathbf{A}_l^0(k). \tag{4.10}$$

The result implies that the inversion only gives a degenerate solution for the coefficients and the intensity cannot be determined solely from two photons. Note that the maximum L, corresponding to the angular resolution of the intensity model, scales with the number of shells $K_{\max}$ (or the inverse of the shell spacing Δk respectively) used for the two-photon inversion.

Here, I used Eq. (4.10) to search for the optimal rotations $\mathbf{U}_l$ instead of optimal coefficients $A_{lm}^{\text{all}}(k)$, which reduced the size of the search space from $(\frac{1}{2}L^2 + \frac{3}{2}L + 1) \cdot K$ to $\frac{1}{3}(L^3 + \frac{15}{4}L^2 + \frac{7}{2}L)$ unknowns (e.g., reducing the number of unknowns from 4940 coefficients to 2370 rotation angles for $L = 18$ and $K = 26$). A rotation in dimension D has $D(D-1)/2$ free angles and for $D = 2l+1$ the sum over $2l^2+l$ free angles for $l \in \{2, 4, ..., L\}$ yields $\frac{1}{3}(L^3 + \frac{15}{4}L^2 + \frac{7}{2}L)$. Note that the number of rotation angles does not depend on the number of shells K anymore, and the difference in the number of unknowns further increases with the number of shells K.

4.4 Monte Carlo Simulated Annealing

The probability p from Eq. (4.8) was maximized by a Monte Carlo / simulated annealing approach on the 'energy' function

$$\begin{aligned} E\left(\left\{k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i\right\} | \{A_{lm}(k)\}\right) &= -\log p\left(\left\{k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i\right\} \middle| \{A_{lm}(k)\}\right) \\ &= -\sum_i \log \tilde{t}(k_1^i, k_2^i, k_3^i, \alpha^i, \beta^i)_{\{A_{lm}(k)\}}, \end{aligned} \tag{4.11}$$

in the space of all rotations $\mathbf{U}_l$ given by the inversion of the two-photon correlation discussed in the previous Section. Each Monte Carlo run was initialized with a random set of rotations $\{\mathbf{U}_l\}$ and the set of unaligned coefficients $\{\mathbf{A}_l^0\}$. In each Monte Carlo step j, all rotations $\mathbf{U}_l^j$ were varied by small random rotations $\mathbf{\Delta}_l(\beta_l)$ such that the updated rotations for each l ($l \leq L$) read $\mathbf{U}_l^{j+1} = \mathbf{\Delta}_l(\beta_l) \cdot \mathbf{U}_l^j$ using stepsizes β_l. In order to escape local minima, a simulated annealing was performed using an exponentially decaying temperature protocol, $T(j) = T_{\text{init}} \exp(j/\tau)$. Steps with an increased energy were also accepted according to the Boltzmann factor $\exp(-\Delta E/T)$. I further used adaptive stepsizes such that all $\beta(l)$ were increased or decreased by a factor μ when accepting or rejecting the proposed steps, respectively. Convergence was improved by using a hierarchical approach in which the

intensity was first determined with low angular resolution and further increased to high resolution. To this end, the variations of low-resolution features were "frozen out" faster than the variations of high-resolution features.

The random rotations $\left\{\mathbf{U}_l \in R^{2l+1\times 2l+1}\right\}$ were generated using QR decompositions of matrices whose entries were drawn from a normal distribution as described by Mezzadri [119]. The rotational variations $\mathbf{\Delta}_l(\beta)$ were calculated via the basis transformation

$$\mathbf{\Delta}_l(\beta) = \mathbf{R}_l \mathbf{S}_l(\beta)\, \mathbf{R}_l^{-1} \tag{4.12}$$

with

$$\mathbf{S}_l(\beta) = \begin{pmatrix} \cos(\beta) & -\sin(\beta) & 0 & ... & 0 \\ \sin(\beta) & \cos(\beta) & 0 & ... & 0 \\ 0 & 0 & I_{2l+1-2} & & \\ ... & ... & & & \\ 0 & 0 & & & \end{pmatrix} \tag{4.13}$$

and random rotation matrices $\mathbf{R}_l$ [120]. Here, sub-matrix $\mathbf{I}_{2l-1}$ in $\mathbf{S}_l$ is a $2l-1$-dimensional unity matrix.

By using the small rotational variations $\mathbf{\Delta}_l(\beta)$, the SO(n) is sampled ergodically. Approximately $[1/(2-2\cos(\beta))]n \cdot \log(n)$ steps are necessary to achieve sufficient sampling aaccording to Ref. [120]. For the largest search space of $L = 18$ with a rotation dimension of $n = 37$ ($n = 2L+1$) and a minimum stepsize of $\beta = 0.025$ rad, 213.777 steps were required to sample rotations in $SO(37)$ sufficiently dense. To ensure that the search space is exhaustively explored, I aimed at an optimization length of over 200.000 Monte Carlo steps. To this end, a time constant for the temperature decrease of $\tau = 50000$ steps was chosen. The initial temperature T_{init} was calculated as 10% of the standard deviation of the energy within 50 random steps away from the starting structure using the initial stepsizes. Further, I used a factor $\mu = 1.01$ for the adaptive stepsizes. The hierarchical approach was implemented by distributing the initial stepsizes according to $\beta(l) = (l-1)\pi$ such that spherical harmonics coefficients with larger expansion orders l are always varied with a larger stepsize $\beta(l)$ than coefficients with lower orders.

4.5 Efficient Computation of the Energy using Histograms

Calculating the probability from Eq. (4.8) (and energy in the Monte Carlo scheme) is computationally expensive due to the typically large number of triples T. I therefore approximated this product by grouping triplets with similar α, β angles and distances k into bins and calculated the function $t(k_1, k_2, k_3, \alpha, \beta)$ for each bin only once, denoted $t_{k_1,k_2,k_3,\alpha,\beta}$, thus markedly reducing the number of function evaluations to the number of bins. To improve the statistics for each bin, the

intrinsic symmetry of the triple correlation function was also used. In particular, all triplets were mapped into the sub-region of the triple correlation that satisfies $k_1 \geq k_2 \geq k_3$. Special care was taken to correct for the fact that triplets with $k_1 = k_2 \neq k_3$ or $k_1 \neq k_2 = k_3$ or $k_1 = k_3 \neq k_2$ occur 3 times more often than $k_1 = k_2 = k_3$ and triplets with $k_1 \neq k_2 \neq k_3$ occur 6 times more often. To compensate for different binsizes, each bin was normalized by $k_1 k_2 k_3$.

In our study, the two-photon and three-photon correlations were histogrammed using sets of scattering images ranging from $1.3 \cdot 10^6$ to $3.3 \cdot 10^9$ images with an average of 10 photons per shot. I further used $K_{\max} = 38$ shells and $N = 32$ ($\Delta\alpha, \Delta\beta = 5.6°$) as bin sizes in correlation space. In Section 4.6, the choice for number of shells $K_{\max}$ and its impact on the resolution is discussed. In this work, the α and β discretization was varied e.g., to $N = 48$ but without an increase in the resolution of the retrieved structures, indicating that $N = 32$ is sufficiently large.

The above histogramming, required us to calculate the probability p differently. In the triplet histogram $\{n_{k_1,k_2,k_3,\alpha,\beta}\}$, the intensity is integrated over different shell volumes with width Δk each. Depending on the fluctuation of the intensity within these volumes, this leads to different integration errors for different (k_1, k_2, k_3)-combinations. However, this error decreases with smaller shell distances Δk.

To avoid this error, I compared the intensities only by the expected (α, β)-distribution of the triplets, omitting the expected relative number of triplets per (k_1, k_2, k_3)-combination. Hence, the probability p from Eq. (4.8) was calculated as

$$p\left(\{n(k_1,k_2,k_3,\alpha,\beta)\} \,|\, \{A_{lm}(k)\}\right) = \prod_{k_1,k_2,k_3} \prod_{\alpha,\beta} \left(\tilde{t}_{k_1,k_2,k_3,\alpha,\beta}\right)^{\tilde{n}_{k_1,k_2,k_3,\alpha,\beta}}, \tag{4.14}$$

normalizing the probabilities

$$\tilde{t}_{k_1,k_2,k_3,\alpha,\beta} = \frac{t_{k_1,k_2,k_3,\alpha,\beta}}{\sum_{\alpha,\beta} t_{k_1,k_2,k_3,\alpha,\beta}} \tag{4.15}$$

and histogram counts

$$\tilde{n}_{k_1,k_2,k_3,\alpha,\beta} = \frac{n_{k_1,k_2,k_3,\alpha,\beta}}{\sum_{\alpha,\beta} n_{k_1,k_2,k_3,\alpha,\beta}}, \tag{4.16}$$

for each (k_1, k_2, k_3)-combination individually. Note that the radial shape of the intensity is already encoded in the two-photon correlation.

4.6 Choice of Optimal Spherical Harmonics Parameters

Three parameters of the spherical harmonics expansion and the histogramming control the resolution of the determined structure. First, for a maximum wave

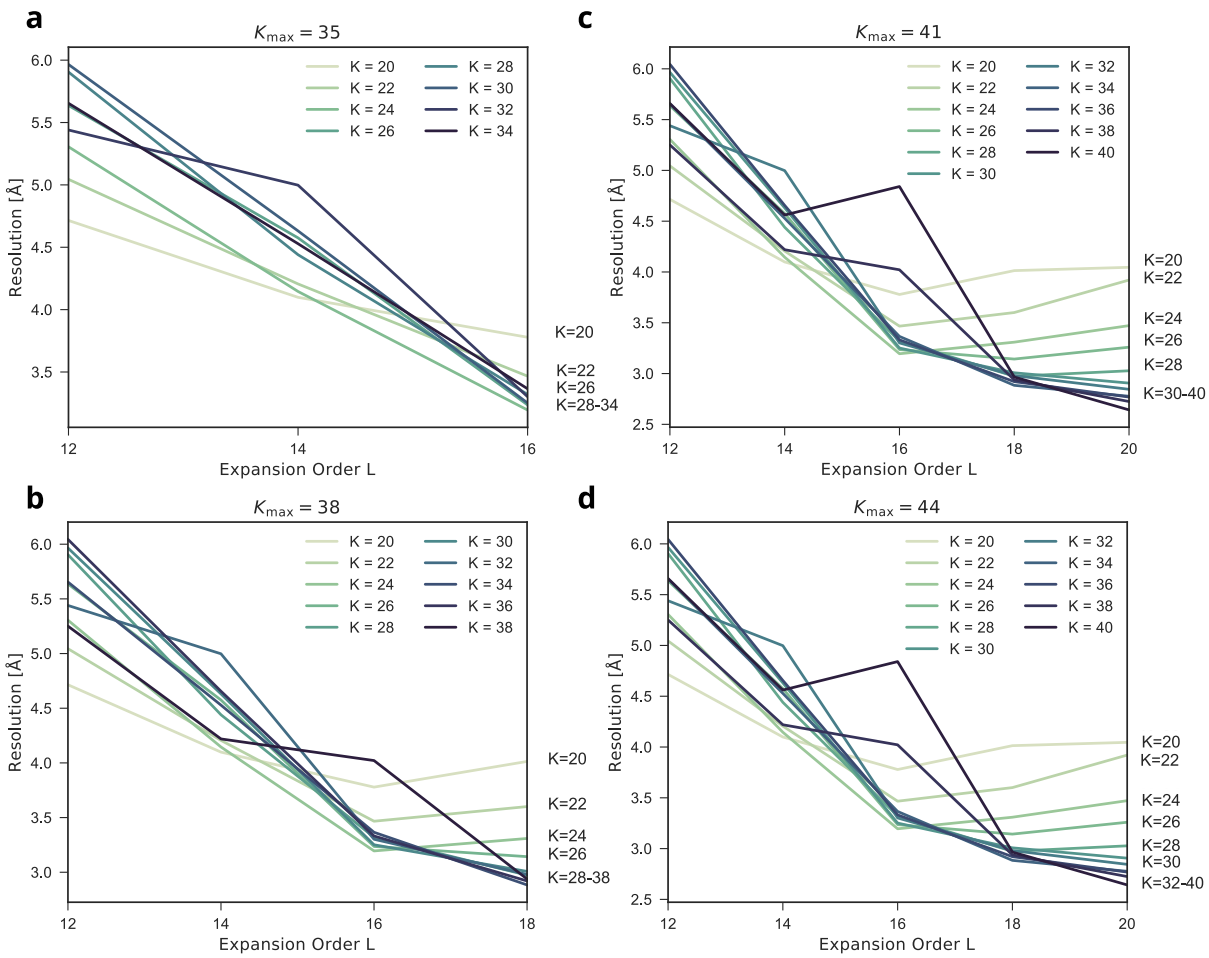


Figure 4.3: Comparison of the effect on resolution of $K_{\max}$, K and L for different parameter combinations. By increasing $K_{\max}$ (35(a), 38(b), 41(c), 43(d)), higher order terms in the spherical harmonics expansion and larger K result in increased resolution.

number $k_{\max}$ up to which sufficient signal is detected, the number of shells $K_{\max}$ that is used in the inversion of the two-photon correlation can be chosen freely. The choice of $K_{\max}$ determines both the shell spacing Δk and the maximum expansion order $L_{\max} = K_{\max}/2$ to which the intensity model, used in the Monte Carlo search, is initially determined. The second parameter is the number of shells $K \leq K_{\max}$ of the intensity model in the structure determination, which determines the maximum wave number $k_{\text{cut}} = K \cdot \Delta k$ and sets an upper bound for the resolution. The third parameter is the expansion order $L \leq K_{\max}/2$ of the intensity model, which controls the angular resolution of the intensity model. The angular resolution of the intensity does not directly correspond to the resolution of the real-space electron density which is why the impact of L on the resolution is indirect. However, for

each wave number k_{cut}, there is a minimum L that is required to describe the intensity "sufficiently" accurately.

Here, I aimed at the optimal set of parameters (K_{opt}, L_{opt}, $K_{\mathrm{max,opt}}$) by which a specific resolution is achieved with minimal computational effort. As discussed in Sec. A.2.2, the computational effort is determined by the time to calculate the full three-photon correlation matrix $\mathbf{T}$ and approximated here as $K \cdot (K+1) \cdot (K+2)/6 \cdot L^4$. For our parameter optimization, I further assumed that an infinite number of photons is recorded up to the maximum wave number k_{max}.

As an example, I aimed at a resolution of 3 Å. To determine the suitable parameters, I calculated the corresponding real space resolution of intensity models with varying expansion parameters K, L and K_{max} (see Sec. 5.2 for the calculation of the resolution). Note that in this check, no structure determination was involved.

Figure 4.3 shows the achieved resolution as a function of L for various number of shells K for four different K_{max} (35, 38, 41, 44). Although the maximum possible L and K increase with K_{max} the k_{cut} ($k_{\mathrm{cut}} = K \cdot \Delta k$) of the model does not increase the same way due to the decrease of Δk. In all the cases, the optimal expansion order $L_{\mathrm{opt}} = K_{\mathrm{max}}/2$ was equivalent to the maximum possible expansion order and K_{max} and K were the limiting parameters.

From all parameter combinations yielding a resolution close to 3 Å, $K_{\mathrm{max}} = 38$, $K = 26$ and $L = 18$ minimized the computational effort, with the matrix multiplication of $\mathbf{A} \in \mathbb{R}^{163153 \times 17576}$ with $\mathbf{F} \in \mathbb{R}^{1024 \times 163153}$ for each Monte Carlo step was the limiting factor.

4.7 Structure Determination in the Presence of Additional Non-Poissonian Noise

To asses how additional noise (beyond the Poisson noise due to low photon counts) affects the achievable resolution, I have carried out synthetic scattering experiments including Gaussian distributed photons,

$$G(\mathbf{k}, \sigma) = (2\pi\sigma^2)^{-1/2} \exp\left(-|\mathbf{k}|^2/2\sigma^2\right) \tag{4.17}$$

(see Fig. 4.4), as a simple noise model. From the generated scattering images, intensities $S(\mathbf{k})$ were determined as described in Sec.4.2.

Assuming that the noise is independent of the molecular structure, the obtained intensities $S(\mathbf{k}) = I(\mathbf{k}) + \gamma N(\mathbf{k})$ are a linear superposition of the molecules' intensity $I(\mathbf{k})$ and the intensity of the unknown noise $N(\mathbf{k})$. Accordingly, the noise was subtracted from $S(\mathbf{k})$ in 3D Fourier space using our noise model $N(\mathbf{k}) = G(\mathbf{k}, \sigma)$ and the estimated signal-to-noise ratio γ. Since the spherical harmonics expansion of a Gaussian distribution is described by a single coefficient $G_{l=0,m=0}(k) = G\,(k, \sigma)$

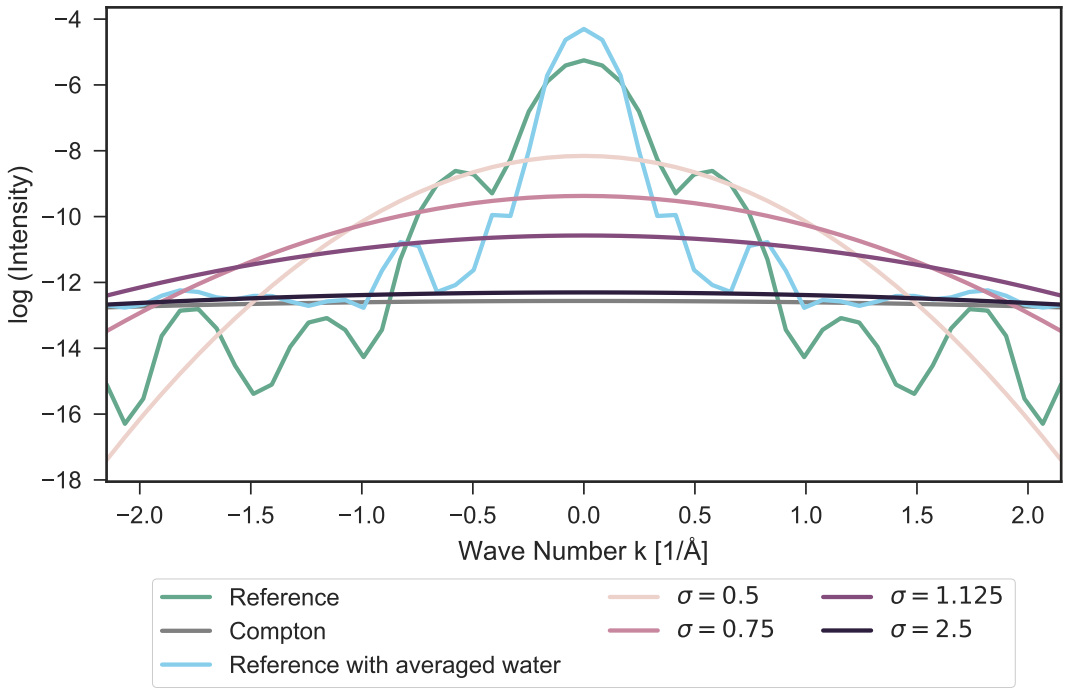


Figure 4.4: Comparison of linear cuts through the normalized intensities of noise distributed according to Gaussian functions with widths $\sigma = [0.5, 0.75, 1.125, 2.5]\,\text{Å}^{-1}$ (purple shades and black), noise from Compton scattering (grey) and noise from the a disordered water shell of 5 Å thickness (aqua). A cut through the Crambin intensity without noise (green) is given for reference. Note that, due to the normalization in 3D, the noise intensities are shown at a signal-to-noise ratio $\gamma = 100\%$; at different signal-to-noise ratios, the noise intensities are shifted vertically with respect to the Crambin intensity.

on each shell k, the noise subtraction simplified to $A^{noise-free}_{l=0,m=0}(k) = A^{noisy}_{l=0,m=0}(k) - \gamma G\,(k, \sigma)$. The noise-free intensity $I(\mathbf{k})$ was then processed as described in Sec. 5.2.

I assessed the effect of noise for different Gaussian widths $\sigma \in [0.5, 0.75, 1.125, 2.5]\ \text{Å}^{-1}$ and signal-to-noise ratios $\gamma \in [10\%, ..., 50\%]$. Figure 4.4 compares the Crambin intensity (green) with the different Gaussian distributions (puples shades, black) at signal-to-noise ratio of $\gamma = 100\%$.

The Figure also shows the noise expected from Compton scattering (grey), which was estimated using the Klein-Nishina differential cross-section [121, 122]

$$\mathrm{d}\sigma = \frac{1}{2}\frac{\alpha^2}{m^2}\left(\frac{E'}{E}\right)^2\left[\frac{E'}{E} + \frac{E}{E'} - \sin^2\theta\right]\mathrm{d}\Omega, \tag{4.18}$$

with the scattering angle θ, the energy of the incoming photons E, the energy of the scattered photon $E' = E/(1 + \frac{E}{m}(1 - \cos\theta))$, the fine structure constant $\alpha = 1/137.04$ and the electron resting mass $m_e = 511$ keV$/c^2$. As can be seen, the noise from Compton scattering (grey) is described well by a Gaussian distributions with width $\sigma = 2.5\,\text{Å}^{-1}$ (black), and thus was used to approximate incoherent scattering.

Finally, I also estimated the noise from the disordered fraction of the water shell by averaging the intensities of 100 Crambin structures with different 5 Å-thick water shells. The resulting intensity (aqua) is similar to the reference intensity with fewer signal in the intermediate regions ($0.2\ \text{Å}^{-1} < k < 1.0\ \text{Å}^{-1}$) and more signal in the center and the high-resolution regions ($k > 1.0\ \text{Å}^{-1}$). Since the noise of the water shell depends on the structure of the biomolecule, potentially combined with ordered water molecules, it is unlikely to be well described by our simple Gaussian model. Therefore, simple noise subtraction will be challenging, and more advanced iterative techniques will be required.

5 Methods for Validating the Approach

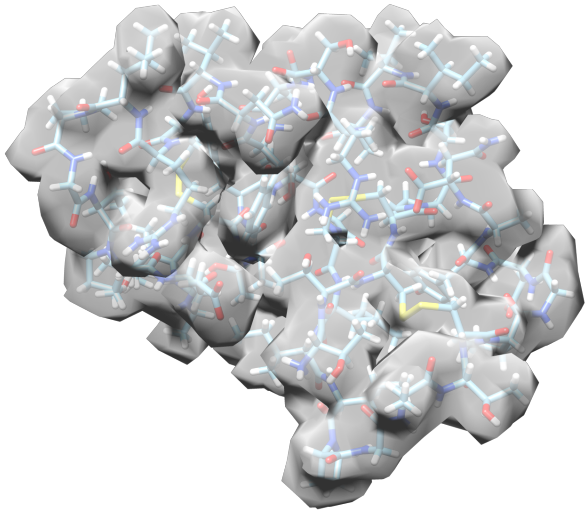


Figure 5.1: Electron density of Crambin with 2 Å resolution.

Contrary to intuition, in single molecule structure determination, smaller molecules are more challenging than larger ones [31] because they scatter fewer photons. I therefore challenged the approach with synthetic scattering images of the rather small, 46 residue comprising, Crambin protein (PDB descriptor: 3U7T), which is known to 0.45Å resolution from X-ray crystallography [123, 124] (Fig. 5.1). The choice for Crambin is based on its size and the resulting low expected signal and not on its function as a small seed storage protein in Cabbage.

5.1 Synthetic Data Generation

For the synthetic scattering experiments, I approximated the 3D electron density $\rho(\mathbf{x})$ by a sum of Gaussian functions centered at the atomic positions $\mathbf{x}_i$,

$$\rho(\mathbf{x}) = \sum_{i=1}^{N_{atoms}} N_i \exp^{-(\mathbf{x}-\mathbf{x}_i)^2/\left(2\sigma_i^2\right)} . \tag{5.1}$$

The heights and variances of the Gaussian spheres depend on the type of atom i. The variances σ_i correspond to the size of the atoms with respect to their scattering cross-section and the height is determined by N_i, the number of electrons which are the potential targets for scattering.

The absolute square of the electron densities' Fourier transformation $I(\mathbf{k}) = |\mathcal{F}[\rho(\mathbf{x})]|^2$ was used to generate synthetic scattering images. In each synthetic scattering experiment, the molecule, and thus also $I(\mathbf{k})$, was randomly oriented. In each shot, on average P photons per image were generated according to the distribution given by the randomly oriented planar slice of the intensity $I_\omega(\mathbf{K})$.

To generate the distributions numerically, first, a random set of N_{pos} positions $\{\mathbf{K}_i\}$ in the k_xk_y-plane was generated according to a 2D Gaussian distribution $G(\mathbf{K})$ with width $\sigma = 1.05\,\text{Å}^{-1}$. Given a random 3D rotation $\mathbf{U}$ (see Sec. 4.4 for uniform sampling of SO(3)), *rejection sampling* was used to accept or reject each position according to $\xi < I_\omega(\mathbf{U}\cdot\mathbf{K}_i)/(M\cdot G(\mathbf{K}_i))$ using uniformly-distributed random numbers $\xi \in [0,1]$ each. Here, the constant M was chosen as $I_{\max}\cdot \max(G(\mathbf{K}))$ such that the ratio $I_\omega(\mathbf{U}\cdot\mathbf{K}_i)/(M\cdot G(\mathbf{K}_i))$ is below 1 for all $\mathbf{K}$.

In accordance with our most conservative estimate discussed in Sec. 2.3.3, the number of positions N_{pos} was chosen such that on average 10 scattered photons were generated. For assessing the dependency of the resolution on the number of scattered photons, additional image sets with 25, 50 or 100 scattered photons were also generated (See Sec. 6.3).

For technical reasons, I used a spherical harmonics expansion of the intensity with a high expansion order $L = 35$ as a sufficiently accurate approximation for $I(\mathbf{k})$ to generate the images. The accuracy of the intensity model was cross-checked with the intensity calculated on a cubic grid (150 grid size) using the Fast Fourier Transform (FFT), resulting in a 0.9999 correlation, thus establishing sufficient accuracy. Altogether, up to $3.3\cdot 10^9$ images were generated using a high degree of parallelism.

5.2 Calculating Resolutions

Figure 5.2 summarizes the calculation of the electron densities as carried out in this work. All intensities were obtained up to an arbitrary Euler rotation (θ,ϕ,ψ) and were therefore rotationally fit to the known reference intensity for subsequent

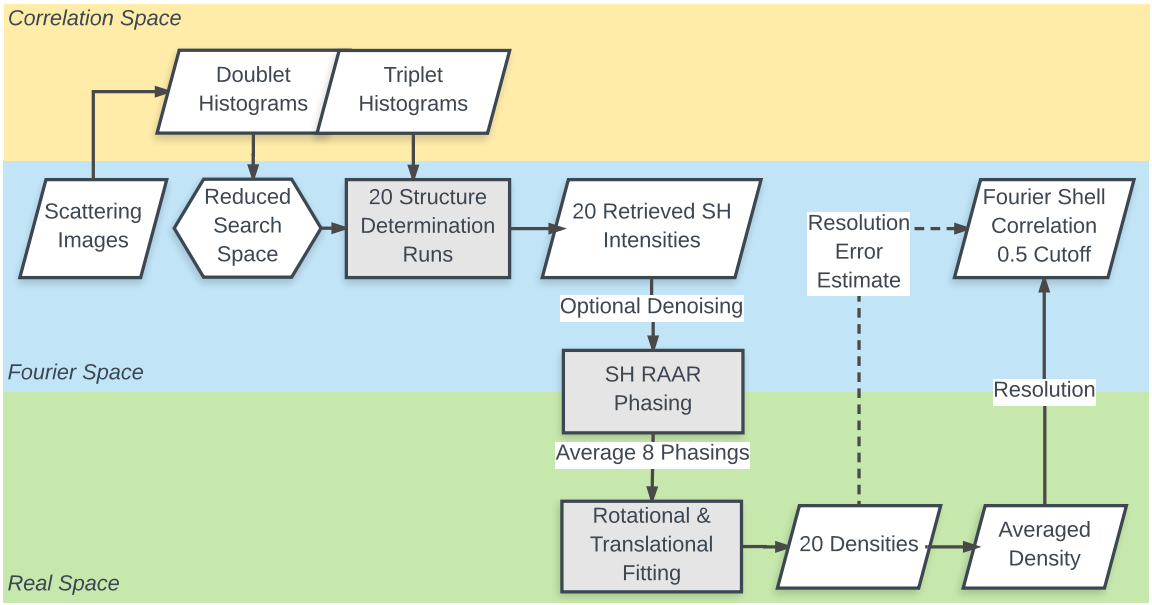


Figure 5.2: Flowchart outlining how electron densities are calculated from the scattering images.

comparison. The phases of the aligned intensities were calculated using the *relaxed averaged alternating reflections* (RAAR) method by Luke [107] as described in Appendix A.1.3. In order to reduce interpolation errors, which are introduced when switching between cubic and spherical harmonics representation, I have adapted the original RAAR implementation to work with spherical harmonics expansions. The resolution of the phased electron densities was characterized by the Fourier shell correlation (FSC),

$$\mathrm{FSC}(k) \quad = \quad \frac{\sum\limits_{k_i \in k} F_1(k_i) \cdot F_2(k_i)^*}{\sqrt[2]{\sum\limits_{k_i \in k} |F_1(k_i)|^2 \cdot \sum\limits_{k_i \in k} |F_2(k_i)|^2}}. \tag{5.2}$$

I have adopted the "gold standard" for the definition of the resolution from cryo-EM [125], where the resolution is defined as the scattering angle k_{res} at which $\mathrm{FSC}(k) = 0.5$, yielding a radial resolution $\Delta r = 2\pi/k_{\mathrm{res}}$.

Starting from an individual set of doublet and triplet histograms (Fig.5.2, top left), 20 independent intensity determination runs were carried out to asses and improve convergence of the Monte Carlo simulated annealing runs. To reduce the phasing error, the phase retrieval of one intensity was carried out 8 times and the resulting 8 electron densities were averaged. The final electron density, for

which the resolution is given, is the average of those 20 individual densities and the resolution error was estimated from the standard deviation of the resolution of the 20 individual electron densities. I chose to average in real space instead of Fourier space before phasing because I found that this sequence yielded more accurate electron densities.

5.3 Structure Determination Software Package ThreePhotons.jl

I have implemented the structure determination approach and the validation methods in the *ThreePhotons.jl* software package which is written in Julia[1] and released on `https://github.com/h4rm/ThreePhotons.jl` under the MIT License[2].

Julia was chosen as a language, among other reasons, because it was developed and optimized for scientific/numerical calculations and therefore handles mathematical expressions and vector/matrix operations very well. It is a very fast due to the just-in-time (JIT) compiler and the optional static-typing. It has elegant and extensible conversions and promotions for numeric and handles complex numbers natively. The C interface via FFI functionality was easy to use and no wrappers or special APIs were required to use existing C libraries such as the s2kit used for the spherical harmonics transformations. In many cases, I used the quick and efficient parallelization mechanisms, e.g., the generation of many scattering image, for the rotational fitting of the intensities or multiple parallel phasings. Performance critical parts of the code were found and optimized by the extensive profiling capabilities of the language. Julia further offered easy installation on all locations including institute cluster, Hydra and GWDG cluster and is free and open source (MIT licensed). I have used the seamless integration with Python libraries such as matplotlib for all graphs generated in this work.

The *ThreePhotons.jl* software package consists of over 6000 lines of code covering the following functions:

- Handling cubic and spherical harmonics represented volumes (loading, saving, set, get, basic operations, transformations)
- Loading of PDB structures, creating electron densities from atomic positions
- Calculating Fourier amplitudes and intensities in both cubic and spherical harmonics representation
- Calculating synthetic scattering images including additional Gaussian noise
- Histogramming the two- and three-photon correlation in the scattering images

[1] `http://julialang.org/`
[2] `https://opensource.org/licenses/MIT`

- Calculating two- and three-photon correlations from intensities represented as spherical harmonics expansions
- Inverting two-photon correlations
- Determining intensities from histogrammed two- and three-photon correlations using the MC simulated annealing approach
- Calculating electron densities from the retrieved intensites using the RAAR phase retrieval method
- Calculating resolutions via Fourier shell correlations

Table 5.1 give a detailed view into the structure of the package. In Appendix A.2.3 I will give source code examples that show how to generate 3D structure of PDB files, generate synthetic scattering images, histogram them and start a structure determination run using the library.

The software package makes extensive use of external libraries. The main dependencies are on the S2Kit library[3] for calculating fast spherical harmonics transformations, FFTW 3.3[4] for Fast Fourier Transform in S2Kit, Gnu Science Library (GSL)[5] for calculating spherical harmonics basis functions Y_{lm}, Legendre Polynomials P_l and Wigner-3j symbols, CUDArt.jl[6] and CUBLAS.jl[7] for fast calculation of three-photon correlations using CUDA v8.0[8], and, Distributions.jl[9] for generating synthetic scattering images.

[3] `http://www.cs.dartmouth.edu/~geelong/sphere/`
[4] `http://www.fftw.org/`
[5] `https://www.gnu.org/software/gsl/`
[6] `https://github.com/JuliaGPU/CUDArt.jl`
[7] `https://github.com/JuliaGPU/CUBLAS.jl`
[8] `https://developer.nvidia.com/cuda-toolkit`
[9] `https://github.com/JuliaStats/Distributions.jl`

File/Directory	Content
cubic.jl	Handling volumes described by cubes.
spherical_harmonics.jl	Handling volumes described by a spherical harmonics expansion.
structure.jl	Approximation of electron densities loaded from PDB files using Gaussian sphere.
data_processing.jl	Calculation of the resolution via Fourier shell correlations of many phasings.
utilities.jl	Random N-dimensional rotations, conversion between spherical and Cartesian coordinates, Gaussian distributions, data serialization.
datagen.jl	Generation of synthetic scattering images including noise. Also handles the histogramming of the images.
correlations.jl	Fast calculation of two- and three-photon correlations. Also implements inversion of two-photon correlation.
cuda.jl, cuda_kernel.cu	CUDA implementation of the three-photon correlation.
determination.jl	Monte Carlo simulated annealing search scheme.
phases.jl	RAAR phasing for spherical harmonics volumes.
tests	Unit tests for all the modules described above.
jobs	Scripts for handling the execution of the code on various clusters, in particular the OWL institute cluster.
sh	\s2kit_interface.c implements the C interface to the s2kit used by spherical_harmonics.jl to handle all SH transformations. See sh\Makefile for compilation.

Table 5.1: List of source files/directories and a description of the functionality implemented within these files.

6 Results and Discussion

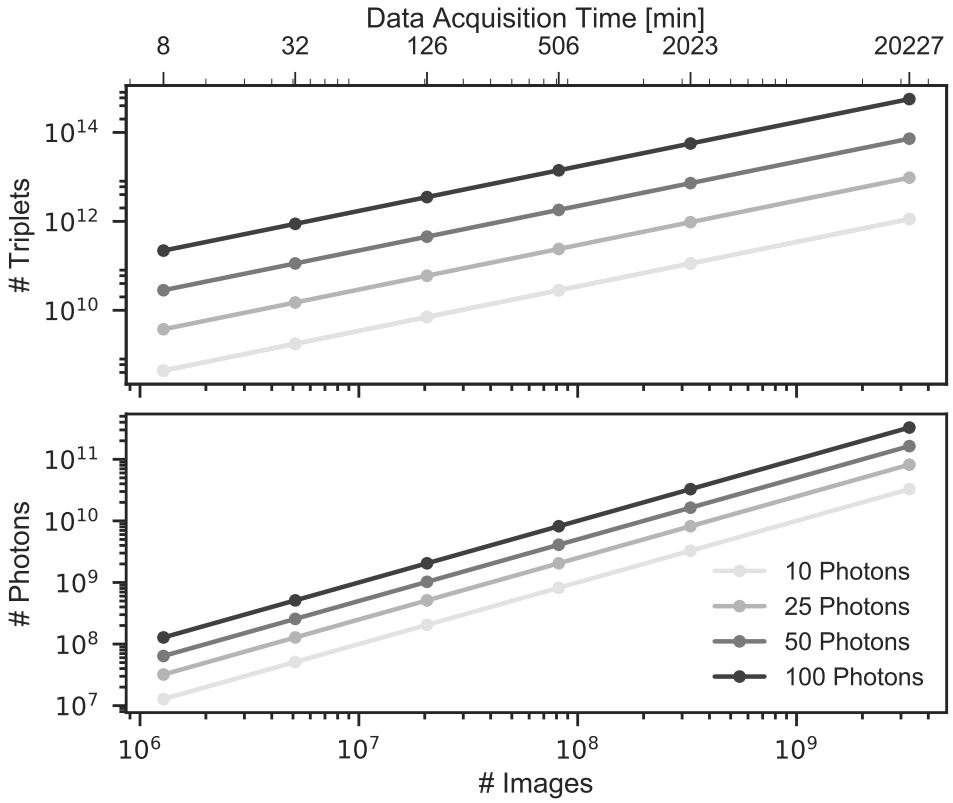


Figure 6.1: Expected number of photons and triplets as a function of image numbers for 10, 20, 50, and 100 photons per image. Also shown (top) is the estimated data acquisition time given a 27 kHz XFEL repetition rate and 10% hit rate.

To stay below the estimate of approximately 20 photons per image (see Sec. 2.3.3 for estimate), I generated up to $3.3 \cdot 10^9$ synthetic scattering images with only 10 photons on average, totalling up to $3.3 \cdot 10^{10}$ recorded photons. With an expected XFEL repetition rate of up to 27 kHz [126], and assuming a hit-rate of 10%, I expect this data to be collected within a few days (see Fig. 6.1). However, as will be discussed in the Sec. 6.3, the data acquisition time substantially decreases to e.g., approx. 30 minutes when on average 100 photons per image are recorded,

reducing the total number of required photons by a factor 100 to $3.3 \cdot 10^8$ (and reducing the number of images by a factor 1000 to $3.3 \cdot 10^6$).

6.1 Near-Atomic Structure Determination of Crambin

Starting from the histograms obtained from $3.3 \cdot 10^9$ synthetic scattering images, I performed 20 independent structure determination runs. For all runs I used an expansion order $L = 18$, $K = 26$ shells and a cutoff $k_{\mathrm{cut}} = 2.15\,\text{Å}^{-1}$, thus setting the maximum achievable resolution to 2.9 Å. Figure 6.2 compares the average intensity obtained from these 20 runs (green) with the reference intensity derived from the known X-ray structure (blue). Overall, the shape of the intensity is recovered very well and only minor deviations in the outer shells, where fewer photons are recorded, are present.

To assess the achievable resolution of the determined Fourier intensities, I calculated 20 real space electron density maps using an iterative phase retrieval algorithm [107]. Figure 6.3 compares the average of the 20 retrieved densities (d, green shaded structure) with the the reference electron density (e, blue shaded structure) which has been calculated from the Fourier density (including phases) with same cutoff k_{cut} as (d). The cross-correlation between the two densities is 0.9. For the averaged electron density, the Fourier shell correlation (FSC) was calculated as a function of the wave number k [125] and the resolution was calculated according to $\Delta r = 2\pi/k_{\mathrm{res}}$ with k_{res} being the wave number at which $\mathrm{FSC}(\mathrm{k_{res}}) = 0.5$. Here, a near-atomic resolution of 3.3 Å was achieved.

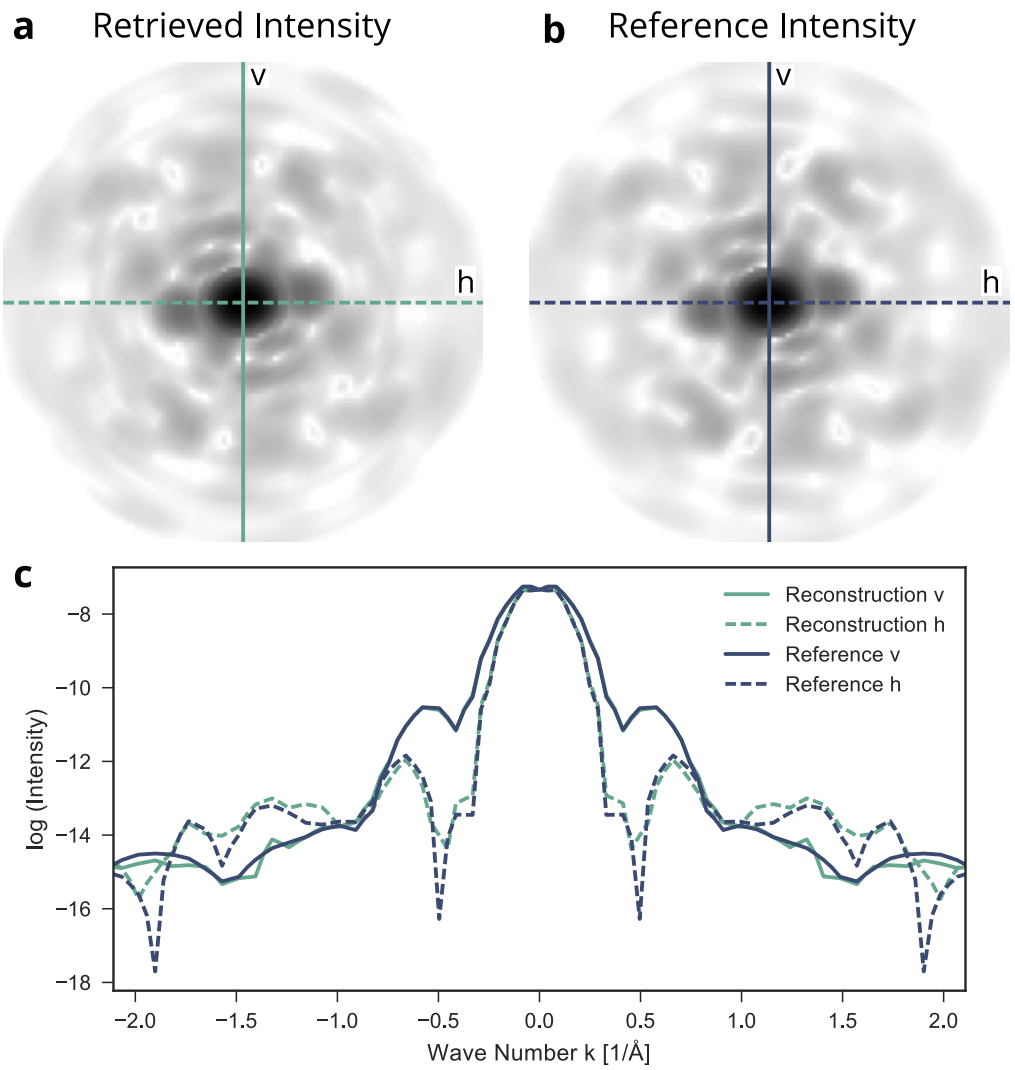


Figure 6.2: Comparison of the retrieved density (green lines and structure) and the reference density of Crambin (blue lines and structure) in Fourier space and real space. Shown are averages over 20 structure determination runs, each using the same $3.3 \cdot 10^9$ images with an average of 10 photons per image yielding $3.3 \cdot 10^{10}$ photons. A cutoff in reciprocal space of $k_{\mathrm{cut}} = 2.15\,\text{Å}^{-1}$ was used and the intensity was expanded with $K = 26$ shells using an expansion order of $L = 18$. **a,b:** Comparison of the the retrieved intensity (a) and the reference intensity (b) in the $k_x k_y$-plane (logarithmic shading). **c:** Comparison of two orthogonal linear cuts (vertical,v and horizontal,h) through the $k_x k_y$-planes shown in panels (a) and (b).

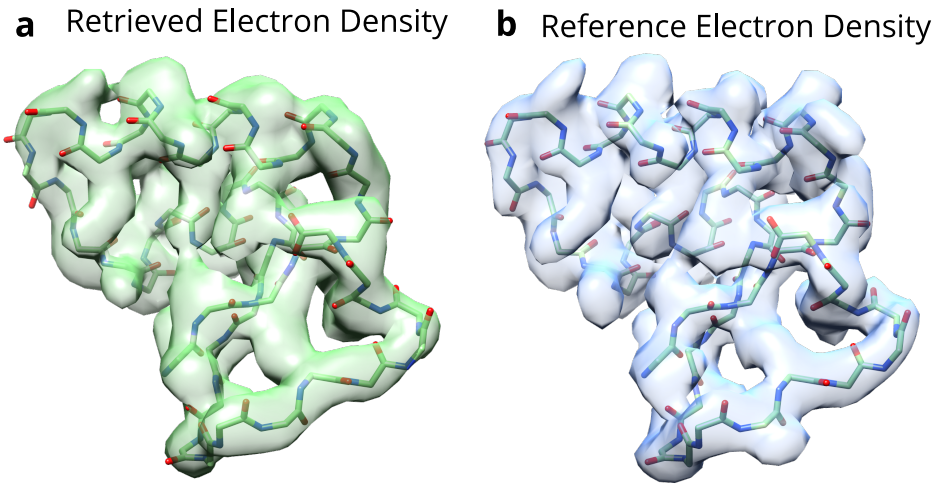


Figure 6.3: Comparison of the retrieved electron density (a) and the reference electron density (b) corresponding to the intensities shown in Fig. 6.2. The reference density was calculated from the known Fourier density using the same cutoff $k_{\text{cut}} = 2.15\,\text{Å}^{-1}$ in reciprocal space as (a). The resolution of the retrieved density is 3.3 Å, the resolution of the reference density is 2.9 Å and the cross-correlation between the two densities is 0.9.

6.2 Impact of Total Number of Recorded Photons on Resolution

Next I explored how the achieved resolution depends on the number of observed photons (and triplets respectively) and, hence, the number of recorded images. To this end, electron densities were calculated and averaged as above from $1.3 \cdot 10^7$ up to $3.3 \cdot 10^{10}$ photons gathered from images with 10 photons on average ($4.7 \cdot 10^8$ up to $1.2 \cdot 10^{12}$ triplets).

Figure 6.4 shows the FSC curves of all retrieved (averaged) densities along with the 0.5 cutoff (vertical dashed line) and the corresponding resolutions (inset). In Figure 6.5 visualizes how the resolution improves with the increasing number of detected photons by comparing four electron densities that were retrieved from histograms with $2.0 \cdot 10^8$ to $3.3 \cdot 10^{10}$ photons.

As mentioned before, the best electron density was retrieved with a near-atomic resolution of 3.3 Å (Fig. 6.5d) from the histograms that was derived from a total of $3.3 \cdot 10^{10}$ photons.

Decreasing the number of photons by a factor of 10 decreased the resolution only slightly by 0.4 Å to 3.7 Å (Fig. 6.5c), which indicates that very likely fewer than $3.3 \cdot 10^{10}$ photons suffice to achieve near-atomic resolution. If much fewer

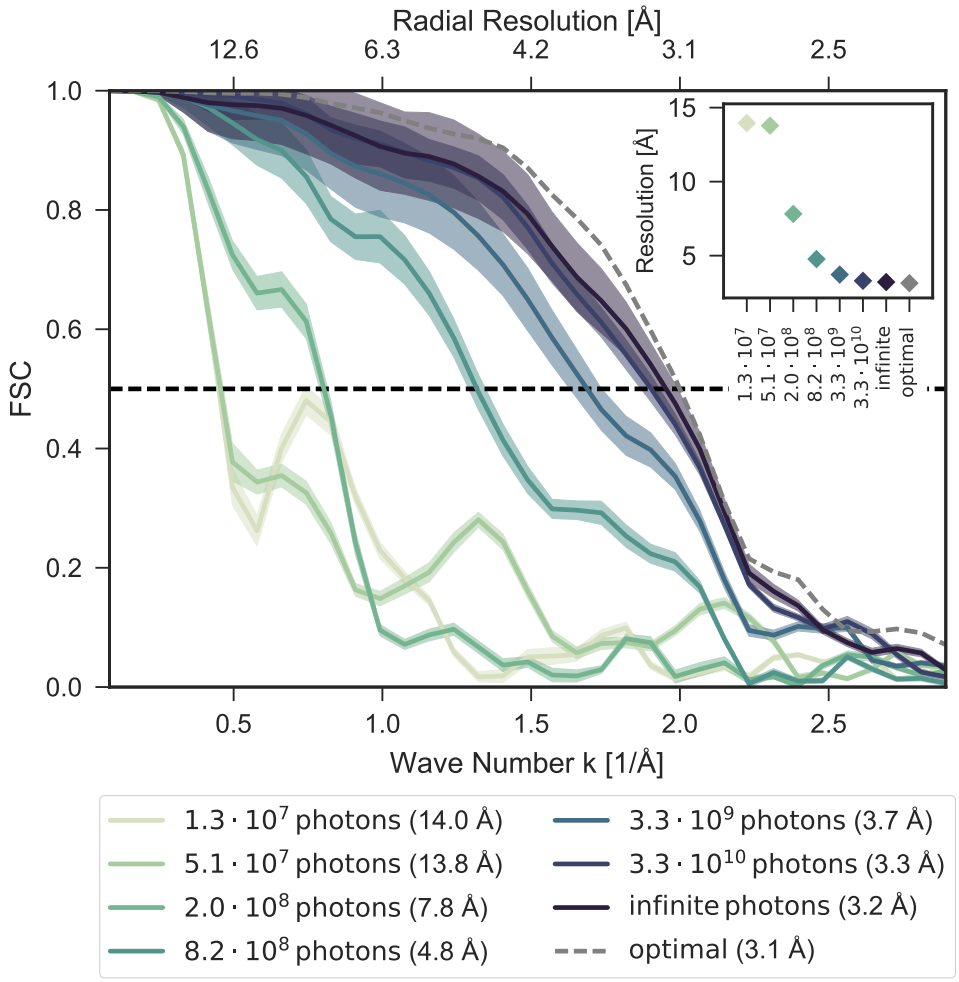
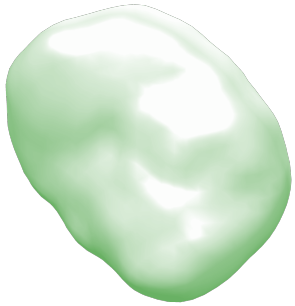


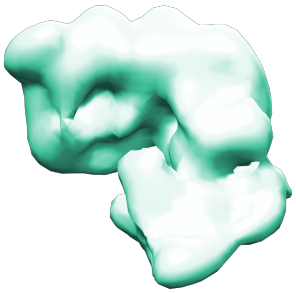
Figure 6.4: Fourier shell correlations (FSC) of densities retrieved from $1.3 \cdot 10^7$ to $3.3 \cdot 10^{10}$ photons ($4.7 \cdot 10^8$ to $1.2 \cdot 10^{12}$ triplets) and infinite photon number. As a reference, the "optimal" FSC is shown (dashed grey), which was calculated directly from the known intensity using the same expansion parameters. The inset shows the corresponding resolutions estimated from $\mathrm{FSC}(k_{res}) = 0.5$.

photons are recorded, e.g. $2.0 \cdot 10^8$, the resolution decreased markedly to 7.8 Å (Fig. 6.5a) and even 14 Å resolution for $1.3 \cdot 10^7$ photons. For comparison, the diamter of Crambin is 17 Å.

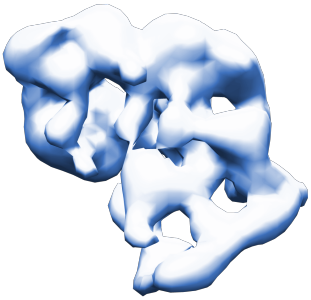
To address the question how much further the resolution can be increased, I mimicked an experiment with infinite number of photons by determining the intensity from the analytically calculated three-photon correlation. As can be seen in Fig. 6.4 (purple line), the resolution only slightly improved by 0.1 Å to about



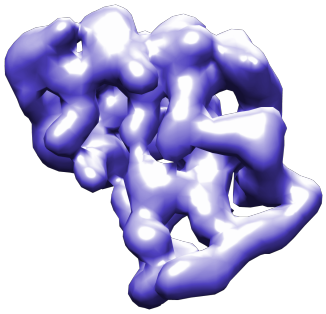
(a) $2.0 \cdot 10^8$ photons, 7.8 Å resolution



(b) $8.2 \cdot 10^8$ photons, 4.8 Å resolution



(c) $3.3 \cdot 10^9$ photons, 3.7 Å resolution



(d) $3.3 \cdot 10^{10}$ photons, 3.3 Å resolution

Figure 6.5: Electron densities retrieved from $2.0 \cdot 10^8$ to $3.3 \cdot 10^{10}$ photons.

3.2 Å indicating that at this point either the expansion order L or insufficient convergence of the Monte Carlo based structure search became resolution limiting. To distinguish between these two possible causes, I phased the electron density directly from the reference intensity, using the same expansion order $L = 18$ as in the other experiments.

The reference intensity is free from convergence issues of the Monte Carlo structure determination and the resulting electron density only includes the phasing errors introduced by the limited angular resolution of the spherical harmonics expansion in Fourier space. The FSC curve of the "optimal phasing" (grey dashed) shows only a minor increase in resolution to 3.1 Å indicating that the Monte Carlo search decreases the resolution by 0.1 Å. The remaining 0.2 Å difference to the optimal resolution of 2.9 Å at the given k_{cut} (not shown) is attributed to the finite expansion order L and the corresponding phasing errors.

I have also independently assessed the overall phasing error by calculating the intensity shell correlation (ISC) between the intensities of the phased electron densities $I_{\text{phased}} = |\mathcal{F}[\rho_{\text{retrieved}}]|^2$ and the intensities before phasing $I_{\text{retrieved}}$. As discussed independently in the Sec. 6.5, the phasing method does not markedly deteriorate the structures.

6.3 Impact of the Photon Counts per Image

In my histogram approach, the maximum number of triplets T that can be collected from an image with P photons is $T = P \cdot (P+1) \cdot (P+2)/6$. However, these triplets are not all statistically independent; instead, starting from 3 photons, each additional photon adds only two real numbers to the triple correlation: a new angle β (with respect to another photon) and a new distance k to the detector center.

The sampling of the three-photon correlation is improved by either collecting more photons per image P or by collecting more images I. However, because for each image, the orientation (3 Euler angles) needs to be inferred, the total amount of information that remains available for structure determination increases with the number of photons per image. Therefore, for every structure determination method, including ours, increasing P is preferred over increasing I, especially at low photon counts. For larger photon counts, the ratio between the 3 Euler angles and P becomes small and hence also the information asymmetry between P and I.

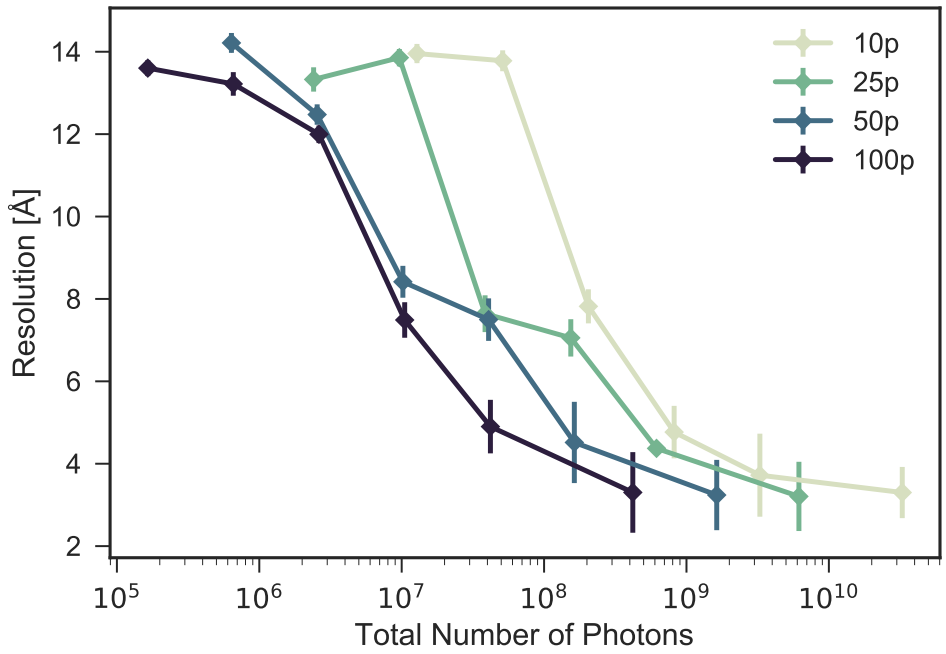


Figure 6.6: The resolution as a function of the total number of photons collected from images with 10, 25, 50 and 100 photons on average.

To assess this effect, I asked how the resolution depends on the number of images I and the photons per image P and therefore carried out additional synthetic experiments using image sets with 10, 25, 50 and 100 average photons P per shot

at different image counts yielding different total number of photons. In Figure 6.6, the achieved resolutions are shown as a function of the number of collected photons for four different $P = [10, 25, 50, 100]$. For the best achievable resolution of 3.3 Å, e.g., the total number of required photons decreases by a factor of 100 from $3.3 \cdot 10^{10}$ to $3.3 \cdot 10^8$ photons (and the number of images decreased by a factor of 1000 from $3.3 \cdot 10^9$ to $3.3 \cdot 10^6$ images) when increasing the photons per image from 10 to 100, thus substantially decreasing the data acquisition time from over 20.000 minutes to only 30 minutes (see Fig. 6.1).

6.4 Structure Determination in Presence of Additional Noise

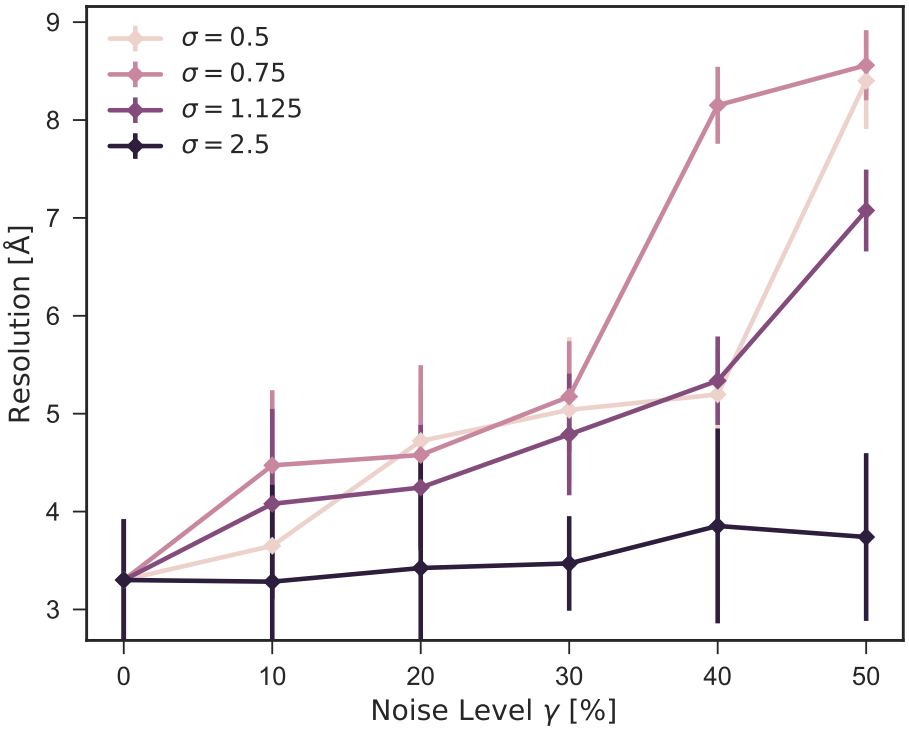


Figure 6.7: Achieved resolution for synthetic experiments with $3.3 \cdot 10^9$ images including an additional fraction γ of random photons ($\gamma = 0 - 50\%$) following a Gaussian distribution with varying width σ. For noise from incoherent scattering (indicated as grey background), I assumed a width $\sigma = 2.5\,\text{Å}^{-1}$, but also included narrower widths $\sigma = [0.5, 0.75, 1.125]\,\text{Å}^{-1}$ as a model for other sources of noise.

I finally assessed how robust my approach is in the presence of additional experimental noise due to, e.g., incoherent scattering, background radiation, detector noise, or scattering at the unstructured fraction of water molecules that may adhere to the surface of the macromolecules [8]. Since only very few single molecule scattering experiments have been carried out so far, quantitative noise models are available only for incoherent scattering, for which a noise level of ca. $\gamma = 25\%$ [109] is expected. Here I modeled the noise as a Gaussian distribution, $G(k, \sigma) = \gamma(2\pi\sigma^2)^{-1/2} \exp(-k^2/2\sigma^2)$. Depending on the width σ, different signal-to-noise ratios are expected in the low-resolution and high-resolution regions of the image respectively. For incoherent scattering (indicated as grey background) a width of $\sigma = 2.5\,\text{Å}^{-1}$ was assumed [121] (see Supplement) which correspond to a relatively uniform noise distribution. Figure 6.7 (black line) shows a moderate decrease in resolution to approx. 3.5 Å when this noise is included within my synthetic experiments (as described in Supplement). Additional noise with a uniform distribution from, e.g. background radiation or detector noise, slightly decreased the resolution to 3.8 Å at 50% noise level.

For scattering at disordered water molecules that are attached to the macromolecular surface, a narrower intensity distribution is expected (see Supplement). To also investigate this effect and the effect of other potential noise sources with non-uniform distribution, in Fig. 6.7, I considered noise with widths of $\sigma = [0.5,\ 0.75,\ 1.125]\,\text{Å}^{-1}$ and noise levels γ between 10% and 50%, the latter corresponding e.g. to up to 100 disordered water molecules per Crambin molecule. The resolution remained better than 5 Å within the 25% noise level but decreases markedly to 9 Å with $\gamma = 50\%$, in particular for narrow noise widths of $\sigma = [0.5,\ 0.75]\,\text{Å}^{-1}$.

In Figure 6.8, the electron densities from the discussed runs are compared to each other.

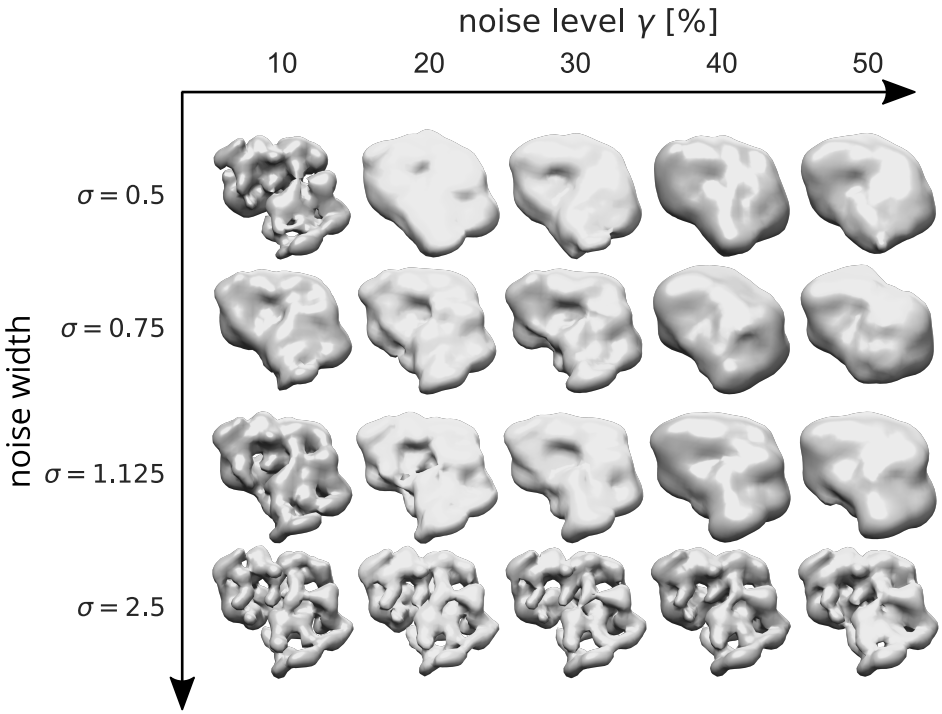


Figure 6.8: Comparison of the electron densities retrieved from images containing noise of different levels $\gamma \in [10\%, ..., 50\%]$ and widths $\sigma \in [0.5, 0.75, 1.125, 2.5]$.

6.5 Evaluation of Phasing Errors

To asses the phasing error, I compared the intensities of the phased electron densities $I_{\text{phased}} = |\mathcal{F}[\rho_{\text{retrieved}}]|^2$ with the intensities $I_{\text{retrieved}}$ before phasing. To this end, the intensity shell correlation (ISC) was calculated as

$$\text{ISC}(k) = \frac{\sum_{k_i \in k} (I_{\text{res}}(k_i) - \overline{I_{res}(k_i)})(I_{\text{ref}}(k_i) - \overline{I_{\text{ref}}(k_i)})}{\sqrt{\sum_{k_i \in k} (I_{res}(k_i) - \overline{I_{res}(k_i)})^2} \sqrt{\sum_{k_i \in k} (I_{\text{ref}}(k_i) - \overline{I_{\text{ref}}(k_i)})^2}}. \quad (6.1)$$

In analogy to the Fourier shell correlation, I considered $\text{ISC}(k) = 0.5$ as a measure for the resolution. As can be seen in Fig. 6.9, the phasing shifted this crossover from approx. 2.8 Å to 3.1 Å, but it does not distort the shapes and relative heights of the ISC curves. Assuming that the phasing error can be estimated from the shift of this crossover, for the high-resolution density result with 3.3 Å resolution (retrieved from $3.3 \cdot 10^{10}$ photons), a decrease in resolution of ca. 0.3 Å is expected to be due to phasing.

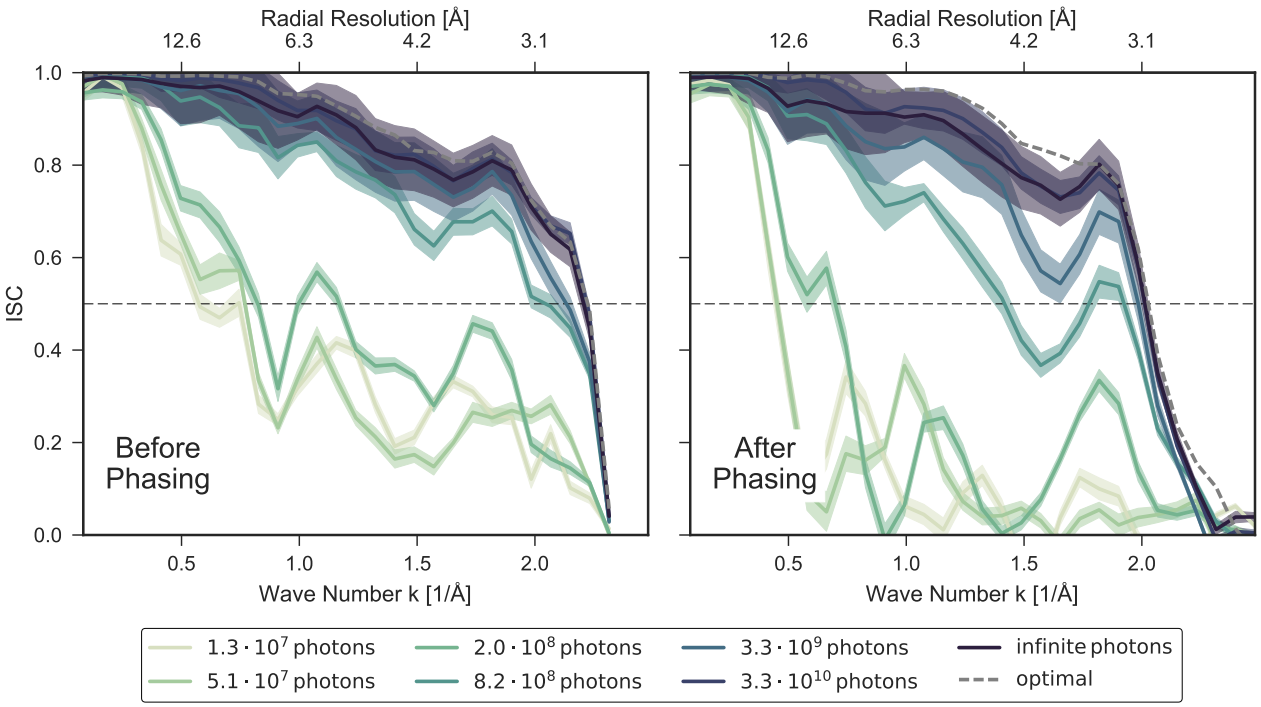


Figure 6.9: Comparison between the intensity shell correlation (ISC) of the retrieved intensities (left) and the ISC calculated from the phased electron densities (right).

6.6 Evaluation of Over-Fitting

Because a large expansion order L requires a larger number of shells K and, therefore, much larger numbers of unknowns, the question remains at which point over-fitting occurs. To quantify this effect for the used sets of images, I calculated the achieved resolution as a function of expansion order L for four different total photon counts $5.1 \cdot 10^7$, $2.0 \cdot 10^8$, $8.2 \cdot 10^8$ and $3.3 \cdot 10^{10}$ ($1.8 \cdot 10^9$, $7.1 \cdot 10^9$, $2.8 \cdot 10^{10}$ and $1.2 \cdot 10^{12}$ triplets respectively) at a fixed number of shells $K = 26$.

Indeed, as shown in Fig. 6.10, for up to $2.0 \cdot 10^8$ photons, the obtained three-photon correlation is too noisy to yield an improved resolution when increasing the model detail and for larger L, the probability p of the intensity model still increases whereas the resolution decreases again, indicating over-fitting.

In contrast, for larger photon counts ($> 8.2 \cdot 10^8$), the resolution improves even up to the expansion order $L = 18$ and no over-fitting is expected here. Even in the outer shells of these correlation histogram, e.g., $k_1 = k_2 = k_3 = 26$, on average 32 triplets per bin are detected. This signal equals to an error of $\sim 17\%$ in each bin entry, a number which seems tolerable for this method.

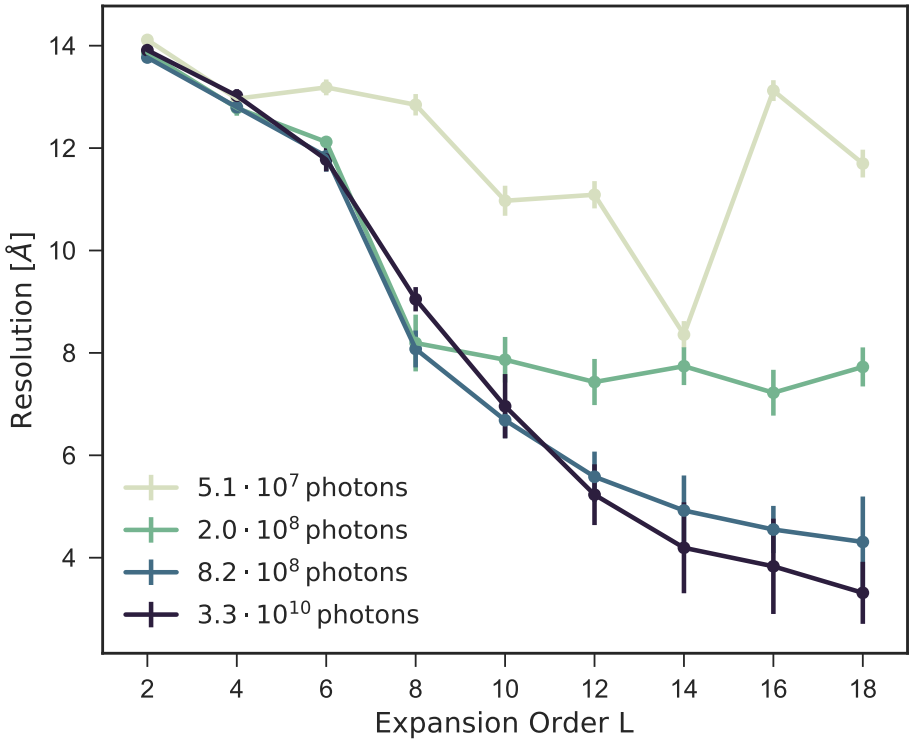


Figure 6.10: Achieved resolution as a function of the expansion order L using $5.1 \cdot 10^7$, $2.0 \cdot 10^8$, $8.2 \cdot 10^8$ and $3.3 \cdot 10^{10}$ photons ($1.8 \cdot 10^9$, $7.1 \cdot 10^9$, $2.8 \cdot 10^{10}$ and $1.2 \cdot 10^{12}$ triplets respectively).

In all cases where no over-fitting occurred, convergence of the simulated annealing became computationally demanding due to the large parameter space.

7 Conclusion

The method developed in this thesis demonstrates *de novo* macromolecular structure determination from as few as three photons per XFEL scattering image at near-atomic resolution. Given that currently available refinement methods require at least 100 photons per image, this finding is quite unexpected. In particular, because two photons per image do not uniquely determine the structure [46], here I have reached the fundamental limit.

My synthetic scattering experiments with subsequent structure determination have shown that, for the most challenging case of small biomolecules, a resolution better than 3.3 Å should be achievable with available technology at realistic beam times; specifically, as my estimate rests on a conservative estimate for the beam fluence of $5.0 \cdot 10^{11}$ transversely coherent photons per X-ray pulse.

The high-resolution structure was derived from $3.3 \cdot 10^9$ images which were generated for the 48 residues mid-sized Crambin protein. The images contained 10 coherently scattered photons, a number which is expected given the European XFEL beam parameters, but in principle also lower photon counts are possible. With only 10% of the scattering images, I have shown that a resolution of 3.7 Å can still be achieved.

Notably, according to Barty *et al.*[126], the beam fluence may be as high as 10^{13} photons per pulse, which is more than 10 times higher than my estimate. I have shown that with a 10-fold increase of the photon count to 100 photons per image, the total number of required images and data acquisition time, respectively, decreases by a factor 1000, putting the beam time required to measure a 3.3 Å resolution structure in the order of minutes. Therefore, even higher resolutions are conceivable for larger molecules [31] at realistic data acquisition times because the number of scattered photons per image increases quickly with the number of atoms in the molecule.

My analysis also suggests that the method is robust against noise from incoherent scattering, and that removing disordered water (or other contaminants) from the molecule in the experiment as much as possible is crucial. Further, fluctuations of the beam intensity – both in time and space, which are a limiting factor for image-wise orientation based methods – should not deteriorate the resolution in my approach, as the correlations are insensitive to such fluctuations. Clearly, further experimental data and improved noise models are required to study the effect of these and other potential noise sources such as background radiation from the evaporated water and detector noise.

Since the experiment measures only intensities, the missing phases were retrieved with a non-convex iterative phasing algorithm which is not guaranteed to produce the optimal solution and there adds errors to the density. However, I have shown that with my spherical harmonics implementation of the phasing, the resolution of the densities deteriorates only by approx. 0.3 Å for the high-resolution case and the phasing step is reliable and stable, in particular for the high-resolution cases.

Over-fitting effects were only seen in structure determination results from sparse histograms with fewer than 10^7 images. The high-resolution results, however, were retrieved from histograms that still had an average signal count of more than 32 triplets per bin even in the outer shells (e.g., $k_1 = k_2 = k_3 = 26$ which only involves photons with high scattering angles) and for these cases no over-fitting was observed.

The Crambin structure determination of 3.3 Å resolution used $L = 18$ and $K = 26$ as spherical harmonics expansion parameters and the Monte Carlo run time was in the order of days. Going towards higher resolutions will require the use of higher expansion parameters which increases the computational effort and eventually makes the computational resources a limiting factor for my approach in the current version. Despite the fact that the most expensive matrix multiplication for the three-photon calculation profits from the high-parallelization of modern graphics cards, it most-like will become memory-limited for larger parameter spaces but solutions are available for this class of problems.

I identified $K = 26$ and $L = 18$ as the optimal spherical harmonics parameters to achieve approx. 3 Å resolution at minimal computational effort. However, fundamentally, the resolution is only limited by the data collection, i.e. the number of images and the photon counts per image.

Note that in this work, multiple gradient based method were tried for the structure optimization but failed due to the non-convex shape of solution space which is comprised of many local minima. The Monte Carlo simulated-annealing approach was a good solution to escape these minima and sufficed to determine near-optimal solutions for the inversion of the three-photon correlation. In particular the use of the hierarchically-distributed stepsizes substantially improved the convergence of the optimization method because it exploited the hierarchy in the high-dimensional search space. This hierarchy is an important advantage of using spherical harmonics expansions and similar stepsize distributions or other iterative refinement approaches will most likely also be used in other optimization schemes.

Assuming a conservative 10% hit rate, my method requires only ca. 10^{10} molecules, which is, compared to nano-crystallography, smaller by a factor of 10 (10^5 nano-crystals with $10^6 \mathrm{nm}^3$ volume) [23]. However, methods that are used to increase the hit rate in nano-crystallographic experiments, e.g., embedding the nano-crystals within polymers in order to decrease the flow velocity of the bulk material, most likely will not improve single molecule scattering experiments because

these sample delivery methods introduce a substantial amount of background noise which negates the advantage of collecting more images.

I have demonstrated that the use of photon correlations does not require the explicit classification of the orientations of the images. However, with the study of the variation of the photon counts I have also shown that it is more beneficial to collect more photons per image than to record more images. This information asymmetry indicates that the lack of information on the orientations for each image reduces the net information available for structure determination. In contrast to previous non-correlation-based structure determination methods, however, my three-photon correlation approach allows for compensation of fewer photons per image P with more images I, in particular in the extreme Poisson regime.

Using only correlations up to third order is sufficient for *de novo* structure determination, but the use of higher-order correlations (four-photons, etc...) may further improve the resolution, albeit most likely rendering the calculation of the forward model extremely expensive. Due to their universal nature, purely Bayesian methods, such as the expectation maximization compression algorithm (EMC) [32], should be optimal and cover the full information content. However, in contrast to my approach, structure determination with EMC has not been demonstrated with very few photons, probably because the method cannot handle the large number of images required for low photon counts due to the exponential computational effort required.

In this work, I have only considered single molecule shots but Kam suggested [46] that the two-photon correlation of scattering images of multiple particles convergence to the two-photon correlation of single particle images, and this occurs even if an "extremely large number" of biomolecules is imaged in each shot (omitting the potentially higher background noise for these scattering events). This fact has been exploited already in several structure determination approaches that use angular (two-photon) correlations measured in fluctuation X-ray scattering experiments (CXS) on proteins in solution [99, 113, 115].

I expect that this 'convergence theorem' also holds for higher-order correlations, because all correlated doublets, triplets, etc... that are comprised of photons from different particles will eventually average out and appear as background signal, similar to noise from incoherent scattering. If, in fact, this is the mechanism responsible for the convergence, this would imply that the speed of convergence is slower for e.g., three-photon correlations because the ratio between triplets from the same molecule and mixed triplets is much smaller than in the two photon case. However, it would allow the illumination of multiple particles in a single shot from time to time which would greatly ease the sample preparation and sample delivery and result in a higher yield of usable scattering images.

It remains to be shown if this assumption holds for the full three-photon correlations, e.g., numerically with the method at hand, by comparing the correlations of synthetic images with superimposed signals of two particle orientations with the

correlations of single particles generated for this thesis. If the sampling of three-photon correlations are indeed robust to multi-particle shots, this may imply that my method may also be applied for structure determination with CXS experiments and potentially enable *ab initio* structure determination of system that are difficult to asses with single molecule experiments, such as proteins that are embedded in a 2D membrane or densely packed disordered systems.

Overall, my results suggest that near-atomic structure determination by single molecule X-ray scattering is within experimental reach. The method is potentially also useful to extract as much as possible information from other types of scattering experiments, in particular when 3D structures are inferred from noisy two dimensional projections, such as cryo-EM, X-ray microscopy, sub-diffractive optical microscopy [127, 128], and X-ray scattering experiments in material science.

8 Outlook

The structure determination results in this thesis suggest that a unique 3D intensity solution exists for a given three-photon correlation because all runs with were started independently from random starting points converged towards the same intensities. However, a direct analytic inversion of the correlations has not been achieved yet. Such a solution could provide an analytic estimate for the final resolution of the electron density for a given set of images. Yet, also the direct inversion will have to address the errors and inaccuracies introduces by the Poissonian shot noise and my method may already by a good solution for this problem. In this Chapter, I will therefore discuss potential improvements and possible validation scenarios of the presented probabilistic three-photon correlation method.

8.1 Improving the Probability Maximization

The Monte Carlo simulated annealing scheme is one of many approaches to optimize the probability of observing the recorded triplets. Among others are, e.g., gradient-based methods that require the derivative of the matrix equation $\mathbf{T} = \mathbf{F} \cdot \mathbf{A}(\mathbf{U})$ with respect to the rotations $\{\mathbf{U}_l\}$ [129] (note that local minima need to be addressed), Markov chain Monte Carlo (MCMC) using Hamilton dynamics [130] or replica exchange MCMC [131, 132].

Along with the optimization method, other energy functions (or probability functions) are possible, that use e.g., additional constraints in the intensity search space (e.g., the positivity and smoothness of the intensity) or different weighting of the triplets and shells $k_1k_2k_3$.

Also different metrics could be used for evaluating the statistical difference between the probability distributions given by expected three-photon correlation $t(k_1, k_2, k_3, \alpha, \beta)$ and the recorded (histogrammed) three-photon correlation $h(k_1, k_2, k_3, \alpha, \beta)$, e.g., by Kullback–Leibler divergence [133] or Hellinger distance [134].

8.2 Improvements and Alterations to the Spherical Harmonics Expansion

I have used the same expansion parameters on all spherical harmonics shells but, in an extension of the method, lower shells could be described with a lower expansion

limit L which could improve computation time and compensate for increasing spatial resolution in lower shells due to the smaller radius.

The decomposition of 3D Fourier space in shells is motivated by the spherical nature of the averaging on the detector. Here, the intensity is only expanded in angular direction but it could also be expanded in radial direction through the use of Zernike polynomials [115]. A radial expansion potentially eradicates the problem of the histogramming which I addressed by normalizing $k_1k_2k_3$-shells individually. In addition, a radial expansion would allow to extend the hierarchical approach, which goes from low to high angular resolution, to also go from low to high radial resolution in the iterative refinement and thus further improve convergence.

As an alternative to the spherical harmonics expansion, the 3D intensity may also be described with an infinite Gaussian mixture model [135], using Gaussian spheres with a width that corresponds to the smallest structure in real space. The three-photon correlation may be calculated in the Gaussian base, most likely yielding basis functions $f(..., \alpha, \beta)$ with triple products of Gaussian functions.

8.3 Real Space Optimization

Combining the structure determination and the phasing into one step may give better structure results with fewer computational effort, as demonstrated e.g. by Donatelli *et al.* [36] with a method that determines the structure of test molecule by classifying the orientations of 24 images with 8000 photons on average.

Such a combined approach (see Fig. 8.1) would start with a random "two-photon conform" intensity and Fourier-transforms the corresponding amplitudes with random phases to real space. After applying the real space constraints (e.g., finite support and positivity of $\rho(\mathbf{x})$), the density is Fourier-transformed and its intensity is projected onto the closest intensity that fits to the measured two-photon correlation [136]. A Monte Carlo simulated annealing step then drives the intensity towards a *better fit* with the measured three-photon correlation (my proposed method) and the resulting new amplitudes $A_{2,3\ \mathrm{pc}}(\mathbf{k})$ are used for the back-transform to real space with the phases $\varphi(\mathbf{k})$ of the structure before two-photon projection and three-photon optimization.

Beyond the typical real space constraints (positivity and support), chemical information or information about the sequence would further reduce the size of the search space. It was also proposed by Bhamre *et al.* [136] to retrieve the phases from similar structures by *orthogonal matrix retrieval* recently developed for cryo-EM, thus further increasing the accuracy and speed of the method.

All transformations and projections described in this proposed method can be expressed in the spherical harmonics framework.

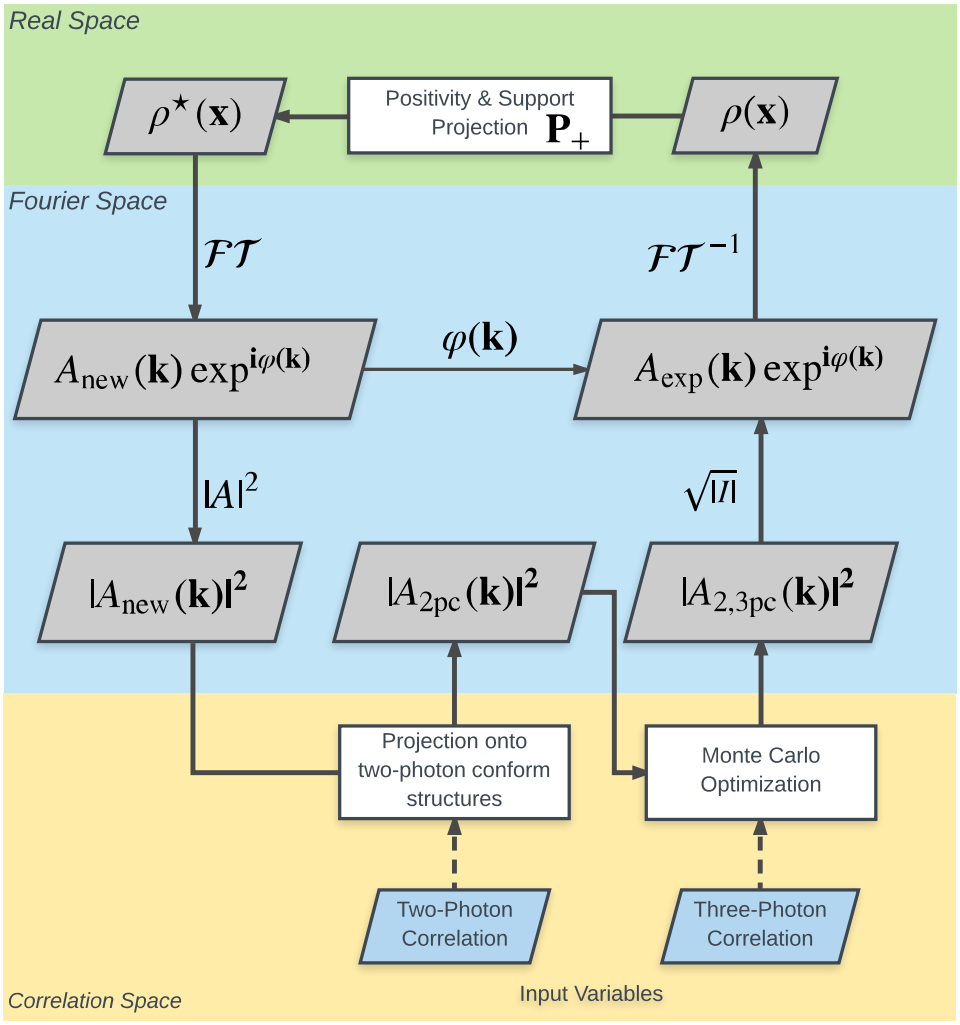


Figure 8.1: Flowchart of a correlation-based structure determination method that combines the phase-retrieval and the structure determination into one iterative scheme.

8.4 Potential Challenges in Light of Experimental Data

For now, single molecule scattering experiments have been carried out only on large virus capsids [18, 111]. After post-processing (removal of all shots that were empty, partly illuminated or contained multiple molecules), only a few 100 scattering images are available which are released on the Coherent X-ray Imaging

Data Bank (CXIDB) [137][1] (provided by the Single Particle Initiative (SPI) at Stanford).

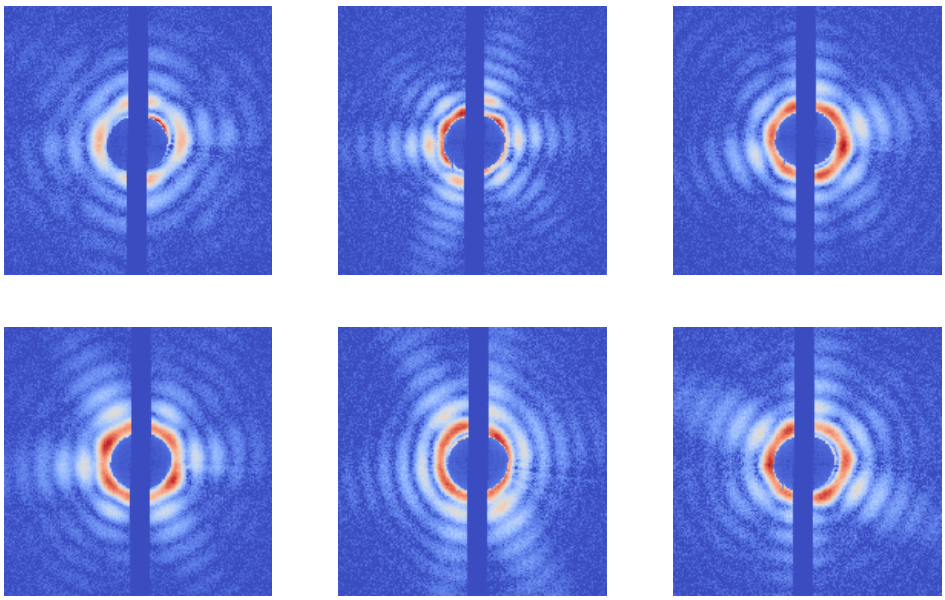


Figure 8.2: Exemplary single molecule X-ray scattering images from the first experiments in 2015 on giant mimivirus particles [18][2]. A wavelength of 1.03 nm was used at LCLS (AMO beamline). The image is cropped down to 300x300 pixel from originally 1000x1000 for visualization purposes. The detector records single photon events.

In these images (see Fig. 8.2), the beam fluence significantly varies from shot to shot, but as mentioned before, this is most likely no problem for my approach. Due to the presence of the beam stop in the experiment, parts of the image are occluded. A close look at the distribution of the noise per pixel throughout the image may give important insights for improving the noise model.

The first single molecule scattering experiments on proteins are scheduled this year at XFEL in Hamburg, most likely with rigid proteins because they are easy to prepare and to inject with existing sample delivery methods. In the beginning, the focus will be on the extensive evaluation of noise, in particular the substantial amount of background radiation which is expected due to the evaporated water molecules in vacuum [138].

As a preparation for experimental data and for further validation, the dependency of the resolution on number of images/photons per image for different

[1] `http://www.cxidb.org/`
[2] `http://www.cxidb.org/id-30.html`

molecule sizes should be checked and in particular it should be verified that larger molecules are simpler to determine. The method should be further validated with scattering images that include some emulation of the beam-stop, although I expect that the shadow of the beam-stop will be averaged out in the photon correlations.

Noise Distributions that Depend on the Structure

I have shown that structure determination in the presence of isotropic noise is possible but in my model, the noise is assumed to be independent of the structure. However, e.g., the noise from scattering on the water molecules on the molecule's surface is anisotropic and depends on the structure of the biomolecule and intensity, respectively. In these cases, an improved model $N(\mathbf{k}|I(\mathbf{k}))$ for the connection between structure and noise is required, including good estimates for the noise levels $\gamma(I(\mathbf{k}))$.

With structure-dependant noise models, the intensities may still be determined with an iterative optimization based on

$$S(\mathbf{k}) = I(\mathbf{k}) + \gamma(I(\mathbf{k}))N(\mathbf{k}|I(\mathbf{k})) \tag{8.1}$$

and a good initial guess for both the noise distributions $N(\mathbf{k}$ and the intensity $I(\mathbf{k})$.

My method can be validated further with different types of noise, maybe with a more realistic model for the noise from water molecules that goes beyond Gaussian distributions.

Conformational Ensembles

Although some proteins are rock-solid, others are very flexible in their physiological environment. The latter pose a challenge for the structure determination because the simple assumption that all molecules are identical and only differ in their orientation does not hold anymore. Instead, in each single molecule scattering shot, one particular conformation will be imaged and the information about this conformation will be lost.

In a first step to address this problem, synthetic scattering images of random representative of the conformational ensemble could be generated for validation. I expect that the structure can still be determined but its density will be "smeared out" in the regions where larger variance occurs. Molecular Replacement methods may still be able to solve the structure in the presence of these uncertainties and the variance gives important insights into the flexibility and dynamics of the protein.

In order to resolve the different conformations, the image sets could be sorted into conformational sub-classes using a Bayesian approach and a coarse model of the biomolecule and its conformations. The correlations of these sub-classes can then be used for structure determination and the retrieved structures could be used for another iterative refinement.

8.5 Assessment of the Information Content in the Scattering Images

I have shown that the achieved resolution depends both on the number of recorded images I and the number of photons per image P. However, the total information available for structure determination does not scale the same way as the total number of photons $M = I \cdot P$. Instead, the total available information c that is contained within I images with P average photons may be calculated as

$$c = I(5 + 2(P - 3)) - 3I. \tag{8.2}$$

As discussed earlier, the first three photons contribute 5 distances and angles and the following photons contributing 2 additional information each. From this information, the three Euler angles required to determine the orientation, are subtracted, which is the reason why additional photos per image are valuable, especially at low photon counts.

The Equation 8.2 can be reformulated to express the total number of required photons M as a function of the photons per images P,

$$M = \frac{cP}{2(P - 2)}. \tag{8.3}$$

If this assumption holds, the trend should be observable in the scattering images of our synthetic experiments. Unfortunately, our generated histograms did not match correct numbers of images and photons per image in order to validate this hypothesis for now but future experiments will be done to clarify this question.

If the expressions above are valid, the total number of images needed to achieve a particular resolution could be calculated for any given signal strength (photon count per image), a prediction which is very important for the experiment.

A Appendix

A.1 Supplementary Theory

A.1.1 Spherical Harmonics Expansions

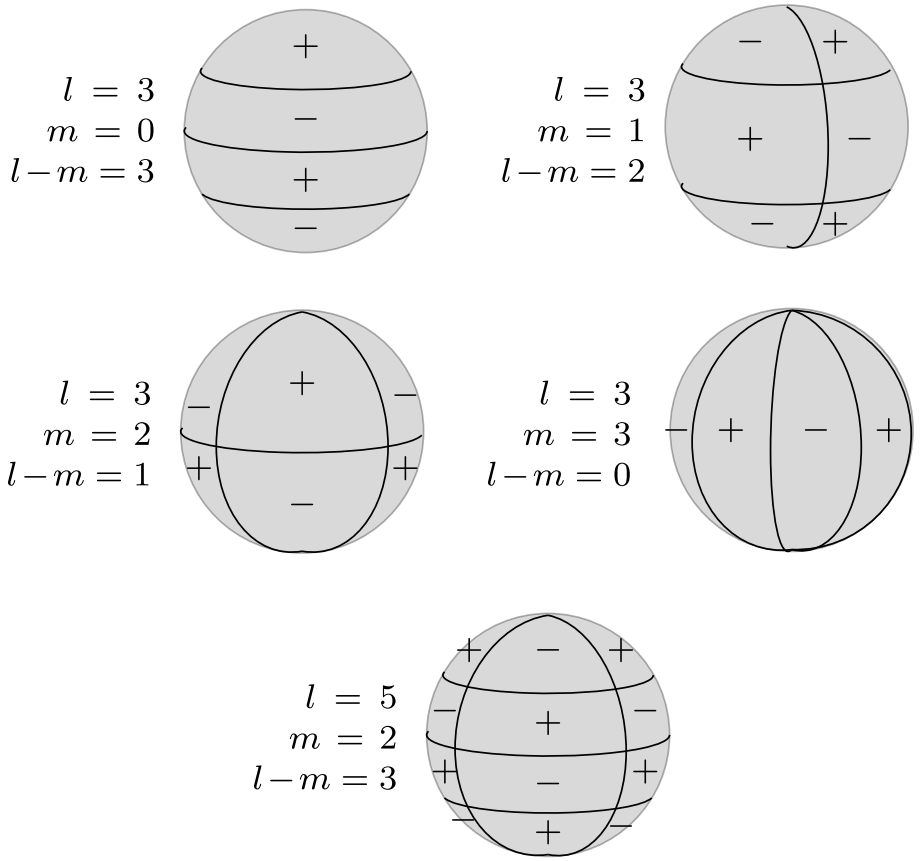


Figure A.1: Schematic depiction of the nodal lines of the spherical harmonics basis functions.[1]. A nodal line marks the crossing of zero (transition between positive and negative). In latitudinal direction, the real and imaginary components of the associated Legendre polynomials cross zero $l - |m|$ times, whereas in longitudinal direction, the trigonometric sin and cos functions have $2|m|$ zeros.

In this work, the 3D Fourier intensity $I(\mathbf{k})$ is decomposed into a set of concentric shells with increasing radii k. The intensity on each shell is **expanded** in spherical harmonics basis functions $Y_{lm}(\theta, \varphi)$:

$$I(\mathbf{k}) = \sum_{l=0}^{L} \sum_{m=-l}^{l} A_{lm}(k) Y_{lm} \varphi, \theta) \tag{A.1}$$

with the spherical harmonics coefficients $A_{lm}(k)$ specifying the contributions of the corresponding Y_{lm}. The set of spherical harmonics coefficients $\{A_{lm}(k)\}$ for $k = n \cdot \Delta k$ with $1 < n < K$ then describe the complete 3D structure. With increasing orders l and m the spherical harmonics exhibit higher modes and express finer structures. This allows to control the angular resolution of the description by limiting the orbital momentum number l with an expansion limit L.

Incorporating the symmetry of the molecule's Fourier intensity with Friedel's rule $I(\mathbf{k}) = I(-\mathbf{k})$ [42] leads to a reduction of descriptors in spherical harmonics representation. This is easily shown by expressing the rule in the spherical harmonics expansion of the intensity:

$$\begin{aligned} \sum_{lm} A_{lm}(k) Y_{lm}\left(\hat{\mathbf{k}}\right) &= \sum_{lm} A_{lm}(k) Y_{lm}\left(-\hat{\mathbf{k}}\right) \\ &= \sum_{lm} A_{lm}(k) (-1)^l Y_{lm}\left(\hat{\mathbf{k}}\right). \end{aligned} \tag{A.2}$$

This expression only holds for even l which is why spherical harmonics coefficients for $l = 1, 3, 5, \ldots$ will be zero when describing symmetric functions such as the intensity.

Fourier Transformation using Spherical Hankel Transform

$$\begin{array}{ccccc} \rho(x) & \longrightarrow & \mathcal{F}(\rho(x)) & \longrightarrow & SH(\mathcal{F}(\rho(x))) \\ \| & & & & \| \\ SH(\rho(x)) & & \longrightarrow & & SB(SH(\rho(x))) \end{array}$$

Figure A.2: The spherical Hankel transform (also denoted as spherical Bessel transform, or SB in contrast to the spherical harmonics transform, SH) calculates the spherical harmonics coefficients of the Fourier transform from the coefficients of the density in real space.

[1] https://de.wikipedia.org/wiki/Kugelflächenfunktionen

The spherical harmonics coefficients of the expansion of the Fourier transformation are calculated using the spherical Hankel transform (see Fig. A.2). The Hankel transform of the SH coefficients of the electron density $f_{lm}(r)$ is given by,

$$a_{lm}(k) = (i)^l \sqrt{\frac{1}{k}} \int_0^\infty f_{lm}(r) j_l(kr) r^2 \mathrm{d}r, \tag{A.3}$$

with the spherical Bessel function $j_l(z) = \sqrt{pi/2z} J_{n+1/2}(z)$. The resulting SH coefficients $a_{lm}(k)$ describe the expansion of the Fourier transformation. The back-transform is calculated as

$$f_{lm}(r) = (-i)^l \sqrt{\frac{1}{r}} \int_0^\infty a_{lm}(k) j_l(kr) k^2 \mathrm{d}k. \tag{A.4}$$

In this Thesis, the integral of the Hankel transform from Eq. A.3 is implemented using direct integration where the sampling points are assumed to have been taken at points related to the zeroes of a Bessel function of fixed order; just like in the case of the discrete Fourier transform, where samples are taken at points related to the zeroes of the sine or cosine function. Other methods can calculate the integral more efficiently (see GNU GSL library) but they would require an uneven distribution of the sampling points, i.e., an uneven spacing between the spherical shells of the expansion. This aspect might be used in future implementations to further reduce the errors in the back-and-forth transformation required, in particular during the phasing.

The numerical integration is performed as a matrix product of the integration kernel $\mathbf{j}_l(kr)$ with the integrand $I(r) = r^2 f_{lm}(r) \Delta r$,

$$\mathbf{a}_l = \sum_{m=0}^{M} j_l(kr_m) I_m \tag{A.5}$$

The integrand is the product of the single with the ring area between neighboring mid-points for $r \in [\Delta r, ..., r_{\max}]$

$$\begin{aligned} I_0 &= \frac{f_{lm}(r_0)}{4}(r_1 - r_0)^2 \\ I_n &= \frac{f_{lm}(r_n)}{4}(r_{n+1}^2 - 2r_n(r_{n+1} - r_{n-1}) - r_{n-1}^2) \forall n \in [1, N-1] \\ I_N &= \frac{f_{lm}(r_N)}{4}(2r_N r_{N-1} - r_{N-1}^2) \end{aligned} \tag{A.6}$$

The integration kernels are precomputed in my implementation which is based on previous work by Leutenegger (MPI-BPC)[2] but was modified to work with spherical harmonics coefficients. Note that both real and imaginary parts of the coefficients need to be transformed.

[2] `http://de.mathworks.com/matlabcentral/fileexchange/13371-hankel-transform`, `http://documents.epfl.ch/users/l/le/leuteneg/www/MATLABToolbox/HankelTransform.html`

Calculating the Intensity

The intensity ($I(\mathbf{k}) = |\mathcal{F}[\rho(\mathbf{x})]|^2$) is calculated in spherical harmonics coordinates from the coefficients of the Fourier amplitudes $F_{lm}(k)$ with [42, 100]

$$A_{lm} = (-1)^m \sum_{l_1 m_1 l_2 m_2} G(l, m, l_1, m_1, l_2, m_2) F_{l_1 m_1}(k) F^*_{l_2 m_2}(k) \tag{A.7}$$

where

$$G(l, m, l_1, m_1, l_2, m_2) = (-1)^{m_1} \tag{A.8}$$

$$\times \left[\frac{(2l_1+1)(2l_2+1)(2l+1)}{4\pi}\right]^{1/2} \begin{pmatrix} l_1 & l_2 & l \\ 0 & 0 & 0 \end{pmatrix} \tag{A.9}$$

$$\times \begin{pmatrix} l_1 & l_2 & l \\ m_1 & \bar{m}_2 & \bar{m} \end{pmatrix}. \tag{A.10}$$

Expressing Rotations

Rotations are expressed as operations on spherical harmonics coefficients. To describe all possible rotations of a point on a sphere, two angular coordinates are necessary. The angle α describes a rotation about the y-axis which is in plane with the equator, while the angle β denotes the rotation about the polar z-axis. With the help of the rotation factors $R^l_{m'm}(\omega)$ [117]

$$R^l_{m'm}(\alpha, \beta) = e^{im\beta} r^l_{m'm}(\alpha) \tag{A.11}$$

$$= e^{im\beta} \int_0^{2\pi} \int_0^{\pi} Y_{lm}(R_y(\alpha)(\theta, \varphi)) \overline{Y_{lm'}(\theta, \varphi)} \sin(\theta) \, \mathrm{d}\theta \mathrm{d}\varphi \tag{A.12}$$

the spherical harmonic coefficients are expressed after an arbitrary rotation as

$$A^{\mathrm{rot}}_{lm} = \sum_{m'=-l}^{l} A^{\mathrm{unrot}}_{lm} R^l_{m'm}. \tag{A.13}$$

The new coefficients are then a linear combination of the unrotated coefficients for a fixed index l.

A.1.2 Inversion of The Two-Photon Correlation

The two photon correlation can be inverted analytically [46], which is best demonstrated by rewriting the expression of the two-photon correlation in vector and matrix form. To that end, we first define a $(2l+1)$-dimensional coefficient vector with all the spherical harmonics coefficients A_{lm} of fixed l and $\{m \in \mathbb{Z} | -l < m < l\}$ as follows,

$$\mathbf{A}_l(k) = \begin{pmatrix} A_{l-m}(k) & \ldots & A_{lm}(k) \end{pmatrix}^T. \tag{A.14}$$

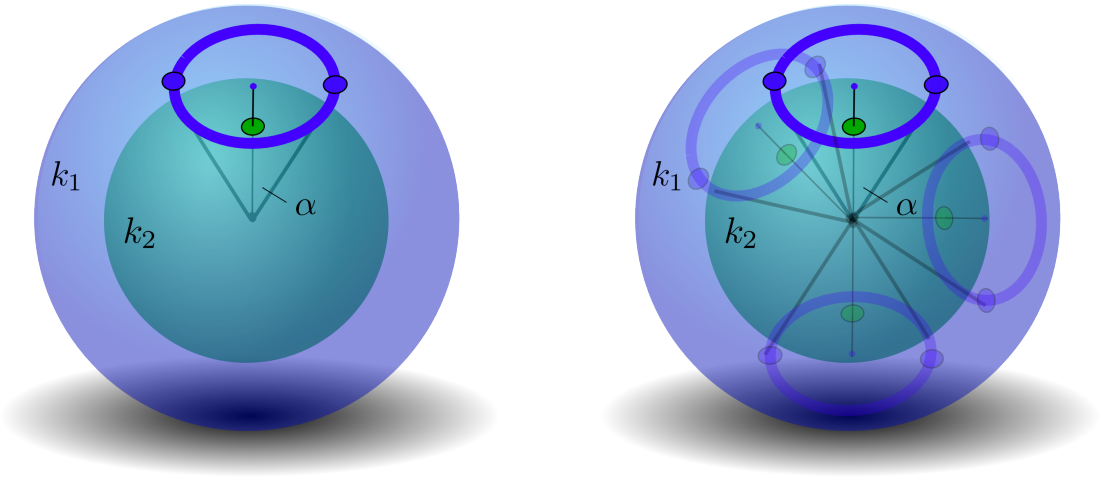


Figure A.3: Sketch of the two photons in the 3D intensity (left) and the orientational average of two intensities as a visualization for the integration of the two-photon correlations (right).

By expressing the sum over m as a scalar product of coefficient vectors, the correlation can be written partly in vector form,

$$c_{k_1,k_2,\alpha} = \sum_l P_l\left(\cos\left(\alpha\right)\right) \underbrace{\mathbf{A_l}\left(k_1\right)\mathbf{A_l^*}\left(k_2\right)}_{S_l(k_1,k_2)}. \tag{A.15}$$

If one further uses the vectors with correlation entries

$$\mathbf{c}_{k_1,k_2} = \left(\begin{array}{ccc} c_{k_1,k_2,\alpha_1} & \ldots & c_{k_1,k_2,\alpha_N} \end{array} \right) \tag{A.16}$$

and scalar product entries

$$\mathbf{S}\left(k_1,k_2\right) = \left(\begin{array}{ccc} S_0\left(k_1,k_2\right) & \ldots & S_l\left(k_1,k_2\right) \end{array} \right) \tag{A.17}$$

as well as the matrix

$$\mathbf{P} = \left(\begin{array}{ccc} P_0\left(\cos\left(\alpha_1\right)\right) & \ldots & P_0\left(\cos\left(\alpha_N\right)\right) \\ \ldots & \ldots & \ldots \\ P_{l_{max}}\left(\cos\left(\alpha_1\right)\right) & \ldots & P_{l_{max}}\left(\cos\left(\alpha_N\right)\right) \end{array} \right), \tag{A.18}$$

the two-photon correlation is written as,

$$\mathbf{c}_{k_1,k_2} = \mathbf{PS}\left(k_1,k_2\right). \tag{A.19}$$

In Equation A.19 the vector $\mathbf{c}$ is known from the experiment and the matrix $\mathbf{P}$ is constant, which is why the linear equation can be solved for the unknown scalar products $S_l\left(k_1,k_2\right)$. For a fixed orbital momentum number l and $K_{\max} \geqq$

$2l+1$, sufficiently many scalar products $S_l(k_i, k_j)$ are available to express the *Gram* matrix known from Distance Geometry,

$$\mathbf{G}_l = \mathbf{F}_l^T \mathbf{F}_l \tag{A.20}$$

$$= \begin{pmatrix} S_l(k_1,k_1) & ... & S_l(k_1,k_{K_{\max}}) \\ ... & ... & ... \\ S_l(k_{K_{\max}},k_1) & ... & S_l(k_{K_{\max}},k_{K_{\max}}) \end{pmatrix}. \tag{A.21}$$

The *Gram* matrix is comprised of the known scalar products $S_l(k_j,k_i)$ between the coefficients vectors belonging to a shell k and index l calculated from Eq. A.19.

The matrix $\mathbf{F}_l$ with $\mathbf{F}_l = \begin{pmatrix} \mathbf{S}_l(k_1) & \mathbf{S}_l(k_2) & ... & \mathbf{S}_l(k_{K_{\max}}) \end{pmatrix}$ contains the spherical harmonics coefficients for a fixed l. By diagonalizing $\mathbf{G}_l$ the system is solved for the matrix $\mathbf{F}_l$ as follows,

$$\mathbf{L}_l = \mathbf{Y_l} \mathbf{G_l} \mathbf{Y_l^T}. \tag{A.22}$$

Solving this eigenvalue problem gives

$$\mathbf{F}_l = \sqrt{\mathbf{L}_l} \mathbf{Y}_l \tag{A.23}$$

with the diagonal matrix $\mathbf{L}_l$ and the transformation matrix $\mathbf{Y}_l$.

The solution matrix $\mathbf{F}_l$ contains a valid set of spherical harmonics coefficients $\{A_{lm}\}$ that yield the experimentally measured two-photon correlation. However, this is an eigenvalue problem and the solution is only obtained up to an arbitrary rotation $\mathbf{U}_l$ in the independent $2l+1$-dimensional eigenspaces,

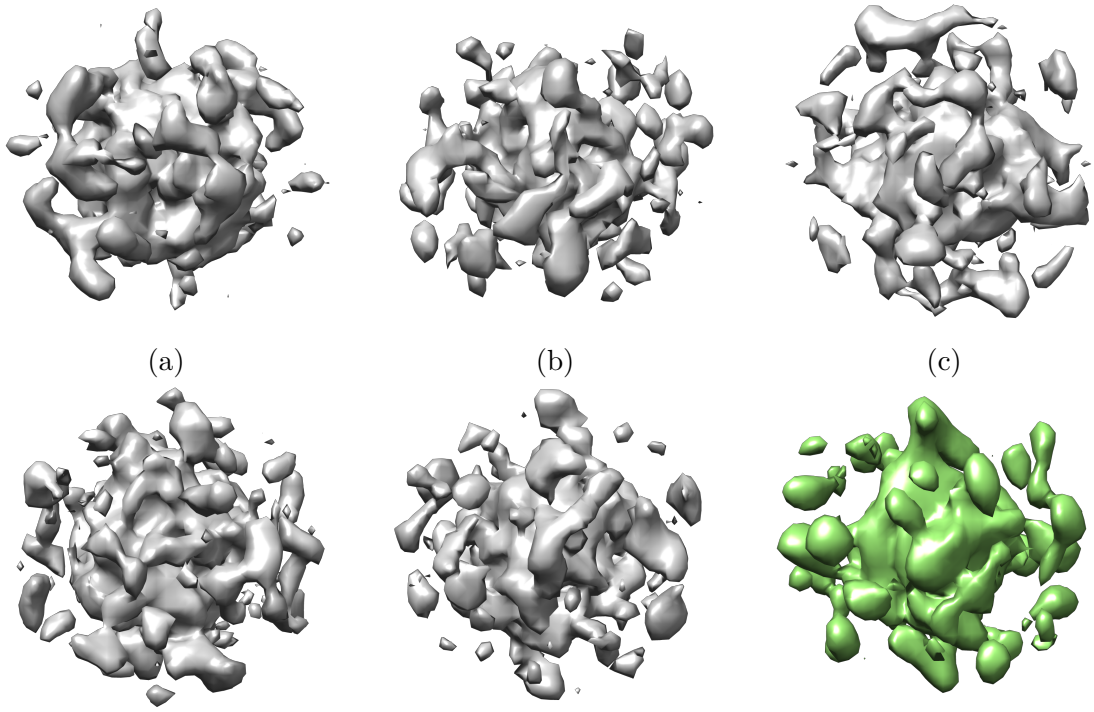
$$\mathbf{A}_l(k) = \mathbf{U}_l \mathbf{A}_l^0(k). \tag{A.24}$$

In Figure A.4 randomly-chosen intensities (a-e, gray), represented by $\{A_{lm}\}$ and calculated from random rotations $\{\mathbf{U}_l\}$, that fit only the two-photon correlation are compared with the reference intensity (f,green) which also fits the three-photon correlation.

The numerical implementation of the inversion was calculated from the doublet histogram, which itself was collected in analogy to the triplet histogram. Note, that during histogramming, doublets with $k_1 \neq k_2$ occur twice as often and the coefficients A_{lm}^0 are retrieved as real values. Real spherical harmonics coefficients were transformed into complex spherical harmonics coefficients according to Ref. [117, 139, 140].

A.1.3 Phase Retrieval

As discussed in Sec. 2.3.2, on the detector the phases are not measured and as a result, 50% of complex Fourier structure information is lost. However, the missing



(a) (b) (c)

(d) (e) (f) Reference Intensity

Figure A.4: Comparison between different intensities that share the same two-photon correlation. The intensities (a-e) were chosen randomly and (f) is the reference intensity of Crambin.

information can be retrieved with iterative algorithms that use additional constraints to retrieve the phases. Starting from a set of random phases $\{\varphi(\mathbf{k})\}$ and the known amplitudes $A(\mathbf{k}) = \sqrt{I(\mathbf{k})}$, an electron density is calculated using the inverse Fourier transformation. In real space, the electron density $\rho(\mathbf{x})$ is projected onto a set of constraints P_{S_+} such that $\rho^\star(\mathbf{x}) = P_{S_+}\rho(\mathbf{x})$. The constraints account for the fact that the:

- protein size is known (electron density support): $\rho(r) = 0 \mid r > r_{max}$
- density is positive: $\rho(\mathbf{x}) \geq 0 \mid \forall \mathbf{x} \in \mathbb{R}^3$
- density is real: $\rho(\mathbf{x}) \in \mathbb{R}^3$

The new electron density $\rho^\star(\mathbf{x})$ is transformed back into Fourier space and the phases are used together with the experimental amplitudes (amplitude projection P_M) for the next iteration.

Phasing algorithms slightly differ in the way they implement the projections P_{S_+} and P_M and how these projections are applied to the densities, thus varying in stability and speed of convergence Here I use the well-established *relaxed averaged*

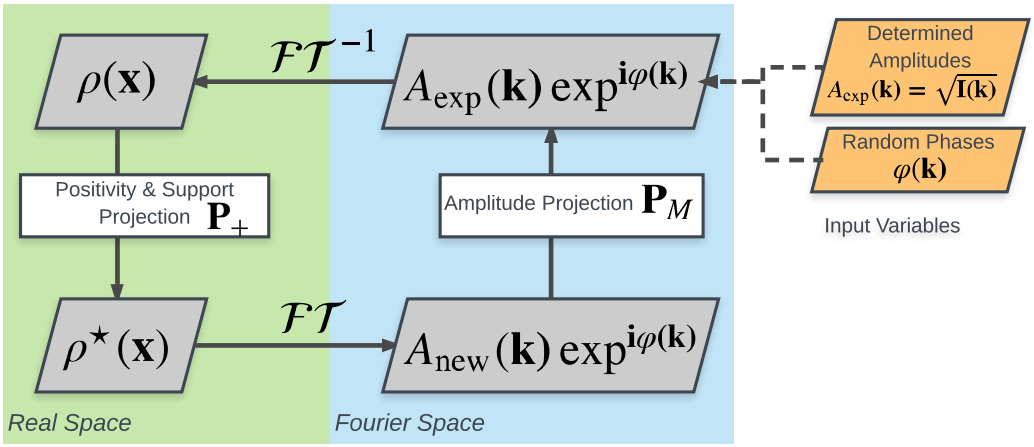


Figure A.5: Schematic flowchart of the phase retrieval algorithm. A back-and-forth transformation between real space and Fourier space is used to fullfill the constraints in both spaces. In real space, the support (size) of the density is known and the constraint that $\rho(\mathbf{x}) > 0$ everywhere is imposed. In Fourier space the retrieved amplitudes are imposed as constraints.

alternating reflections (RAAR) algorithm developed by Luke[107]. The density between consecutive steps are calculated as:

$$\rho_{n+1} = \left(\frac{\beta_n}{2}(P_{S_+}P_M + I) + (1-\beta)P_M \right) \rho_n \tag{A.25}$$

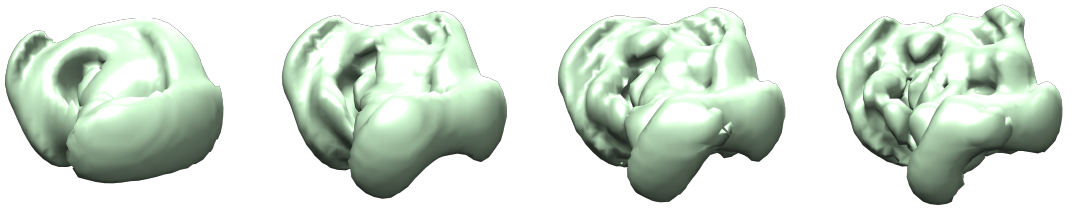
and a contraction factor β_n,

$$\beta_{n+1} = \beta_0 + (1-\beta_0)(1-\exp(-(n/7)^3)), \tag{A.26}$$

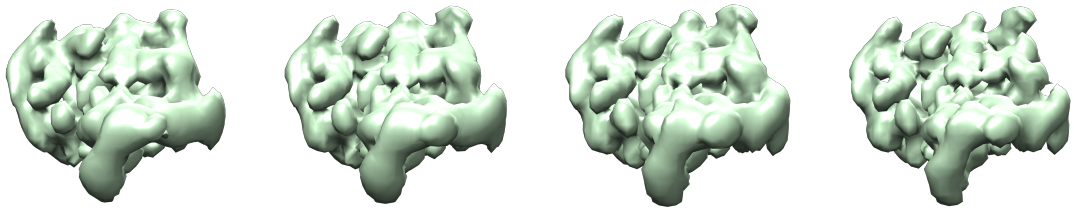
typically starting from $\beta_0 = 0.75$ and increasing towards 1.0. Luke proposes the convergence criteria

$$E_{S_+}(x_n) \quad = \quad \frac{\|P_{S_+}(P_M(\rho_n)) - P_M(\rho_n)\|^2}{\|P_M(\rho_n\|^2}, \tag{A.27}$$

however, in the implementation used in this Thesis, it sufficed to end the phasing after $\sim$ 1000 iterations.



(a) $L = 4$ (b) $L = 6$ (c) $L = 8$ (d) $L = 10$



(e) $L = 12$ (f) $L = 14$ (g) $L = 16$ (h) $L = 18$

Figure A.6: Electron density of a Crambin molecule with increasing $L = 4..18$ expanded with $K = 18$ shells.

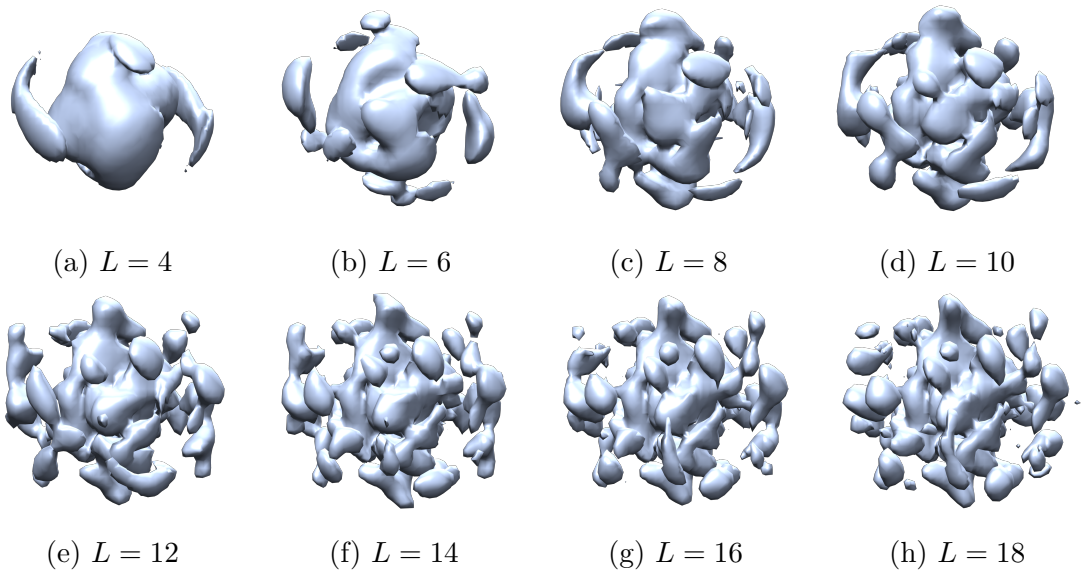
A.2 Implementation Details

A.2.1 Implementation of the Spherical Harmonics Expansion

All Fast Spherical Harmonics Transformations were performed using the S2Kit framework [141, 142][3]. The same spherical harmonics expansion order L was used for all shells. For the structure determination, $L = 18$ was used, which yields $(2L)^2 = 1296$ sample points on the sphere with an even sampling in $\phi \in [0, 2\pi]$ and $\theta \in [-\pi/2, \pi/2]$ direction. The angular resolution of the expansion is $\Delta\theta = \pi/(2L)$ or $\Delta\varphi = 2\pi/(2L)$ respectively which in our case for $L = 18$ corresponds to an angular resolution of $\Delta\theta = 5.0°$ in longitude direction and $\Delta\varphi = 10.0°$ in latitude direction. The density $\rho(\mathbf{x})$, expanded with a spherical harmonics basis, was Fourier transformed by applying the spherical Bessel transform (Hankel transform) to the coefficients according to Ref. [143–145] All Wigner matrices were calculated as described in Ref. [139, 140] and the absolute square of the Fourier density was calculated according to Ref. [100] by transforming the coefficients directly.

As a visual demonstration for the scaling of the angular resolution with the expansion order L, Figure A.6 and Figure A.7 show the Crambin electron density and the intensity respectively for $L = 2...18$ ($\Delta\theta = 45°...5.0°$ and $\Delta\varphi = 90°...10.0°$).

[3] http://www.cs.dartmouth.edu/~geelong/sphere



(a) $L = 4$ (b) $L = 6$ (c) $L = 8$ (d) $L = 10$

(e) $L = 12$ (f) $L = 14$ (g) $L = 16$ (h) $L = 18$

Figure A.7: Intensity of a Crambin molecule with increasing $L = 4..18$ expanded with $K = 38$ shells.

In Figure A.8, the intensity shell correlations (ISC) between different intensity models expanded with a spherical harmonics order $L \in [8, 10, ..., 20]$ and the reference intensity with expansion order $L = 35$ (mimicking high-resolution) is shown. The intensity model expanded with $L = 18$, as used in this Thesis, is already very close to reference model, indicating that $L = 18$ captures almost all structural details.

A.2.2 Efficient Computation of the Three-Photon Correlation

Our method requires the fast evaluation of the three-photon correlation for a proposed set of spherical harmonics coefficients $\{A_{lm}(k)\}$. For this purpose, we combined the Wigner-3j symbols and spherical harmonics basis functions into the three-photon basis function $f(l_1, l_2, l_3, m_1, m_2, m_3, \alpha, \beta)$ which is non-zero only for $m_1 + m_2 + m_3 = 0$ and $|l_1 - l_2| \leq l_3 \leq l_1 + l_2$ as inherited by the Wigner-3j symbols. Here, we denote the number of non-zero index combinations $(l_1, l_2, l_3, m_1, m_2, m_3)$ as B and the number of discrete angles $\alpha, \beta \in [0, \pi]$ in one dimension as N, as further described in Sec. 4.5.

The entire three-photon correlation $\mathbf{T}$ is calculated by the matrix product

$$\mathbf{T} = \mathbf{F} \cdot \mathbf{A}, \tag{A.28}$$

with the matrix $\mathbf{A} \in \mathbb{R}^{B \times K^3}$ that consists of the triple products of coefficients, $A_{ij} = A^{k_1(j)}_{l_1 m_1(i)} A^{k_2(j)}_{l_2 m_2(i)} A^{k_3(j)*}_{l_3 m_3(i)}$, the matrix $\mathbf{F} \in \mathbb{R}^{N^2 \times B}$ that consists of three-photon

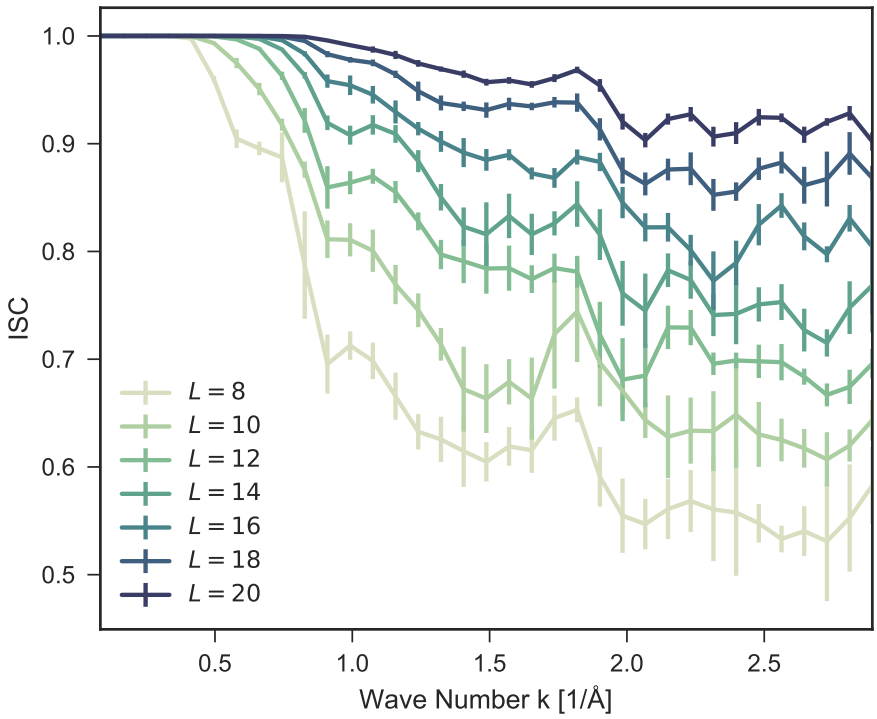


Figure A.8: Intensity shell correlation (ISC) between a spherical harmonics model of the reference intensity with large expansion order $L = 31$ (and high angular resolution respectively) and a model with a lower expansion limit $L = [8, .., 20]$.

basis functions, $F_{ij} = f((l_1 l_2 l_3 m_1 m_2 m_3)(i), (\alpha, \beta)(j))$ and the full three-photon correlation matrix $\mathbf{T} \in \mathbb{R}^{N^2 \times K^3}$ with the entries $T_{ij} = t((k_1, k_2, k_3)(i), (\alpha, \beta)(j))$. This vectorized expression can be calculated with a high degree of parallelism, but nevertheless becomes the limiting factor in the computation. In particular, the number B of three-photon basis functions $f(l_1, l_2, l_3, m_1, m_2, m_3, \alpha, \beta)$ grows quickly with $B \sim L^4$ (e.g., $B = 11841$ for $L = 10$ and $B = 163153$ for $L = 18$). See Sec. 4.6 on how this scaling affects our choice of spherical harmonics parameters.

In our implementation, we calculated both $\mathbf{A}$ and $\mathbf{T}$ using a custom CUDA kernel, which significantly improved (> 100x) the performance over CPU-based implementations and thus rendered the optimization computationally tractable.

A.2.3 Usage of ThreePhotons.jl Software Package

The electron density, Fourier density and intensity can be calculated from any PDB structure as follows:

Listing A.1: Example code for loading a Crambin pdb file and calculating electron densities and Fourier densities.

```
#loading a Crambin structure and generating spherical
#harmonics represenations of its electron density,
#Fourier transform and intensity

using ThreePhotons

#Maximum expansion order of SH expansion
LMAX = 25
#Maximum number of shells used in the expansion
KMAX = 30

#Description of the Crambin electron density,
#Fourier density, and Fourier intensity expanded
#in Spherical Harmonics
density,fourier,intensity =
createSphericalHarmonicsStructure(
"data/structures/crambin.pdb", LMAX, KMAX, float(KMAX))
#Same Crambin structure on a cubic grid
densityCube,fourierCube,intensityCube =
createCubicStructure(
"data/structures/crambin.pdb", 2*KMAX+1, float(KMAX))
```

The synthetic scattering images are calculated as:

Listing A.2: Example code for generating 10^6 synthetic scattering images in parallel with 10 photons per image.

```
using ThreePhotons
include("jobs/runs.jl")

generate_histograms(;
    #Number of images to generate
    num_images        = Integer(1e6),
    #Or maximum number of triplets
    max_triplets      = Integer(0),
    #Number of CPU cores used for the data generation
    Ncores            = 8,
    #Alpha/beta discretization
    N                 = 32,
```

```
        photons_per_image = 10,
        batchsize         = round(Int64,1e6/8),
        successive_jobs   = 1,
        #Use cubic or SH description for data generation
        use_cube          = false,
        #Fraction of maximum wave number
        qcut_ratio        = 1.0,
        #total number of shells
        K                 = 38,
        #maximum radius in real space
        rmax              = float(38),
        histogram_method  =
        "histogramCorrelationsInPicture_alltoall",
        structure_pdb_path= "data/structures/crambin.pdb")
```

jobs/runs.jl includes helper functions to spawn data generation and structure determination runs in various environments (including cluster systems).

Given a histogrammed two- and three-photon correlation, the structure can be retrieved de novo:

Listing A.3: Example code for determining the structure from a two- and three-photon correlation histogram.

```
using ThreePhotons
include("jobs/runs.jl")

#number of images
num_images::Int64 = Integer(1e6)
#Maximum shell number used for two-photon inversion
KMAX::Int64       = 38
#Alpha/beta discretization
N::Int64          = 32
#Maximum expansion order
L::Int64          = 18
#Number of shells used for structure determination
K::Int64          = 26
#Photons per image used for the histogram
ppi::Int64        = 10
#Maximum radius of the reference structures
rmax              = float(KMAX)
#histogram file name
name              = "histogram.dat"
```

```
run_determination(
    #Name of the run
    "runname",
    #Path to the histogram file
    histograms                = name,

    #Expansion parameters (see above)
    K                         = K,
    L                         = L,
    KMAX                      = KMAX,
    rmax                      = rmax,
    N                         = N,

    #Monte Carlo simulated annealing parameters
    initial_stepsize          = pi/4.0,
    optimizer                 = "rotate_all_at_once",
    initial_temperature_factor=0.1,
    temperature_decay         = 0.99998,
    stepsizefactor            = 1.01
    measure                   = "Bayes",

    #Misc parameters
    range                     = 1000:1019,
    postprocess               = true,
    gpu                       = true,
    Ncores                    = 20)
```

Bibliography

[1] Rosalind E. Franklin and R. G. Gosling. "Molecular Configuration in Sodium Thymonucleate". In: *Nature* 171.4356 (Apr. 1953), pp. 740–741. DOI: 10.1038/171740a0. arXiv: arXiv:1011.1669v3.

[2] M. F. Perutz et al. "Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis". In: *Nature* 185.4711 (Feb. 1960), pp. 416–422. DOI: 10.1038/185416a0.

[3] C. C. F. Blake et al. "Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 Å Resolution". In: *Nature* 206.4986 (1965), pp. 757–761. DOI: 10.1038/206757a0.

[4] M Kurplus and J A McCammon. "Dynamics of Proteins: Elements and Function". In: *Annual Review of Biochemistry* 52.1 (June 1983), pp. 263–300. DOI: 10.1146/annurev.bi.52.070183.001403.

[5] Eric Martz et al. *Nobel Prizes for 3D Molecular Structures*. 2017.

[6] Kim D. Pruitt et al. "NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy". In: *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D130–D135. DOI: 10.1093/nar/gkr1079.

[7] H M Berman et al. "The protein data bank." In: *Nucleic acids research* 28.1 (Jan. 2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[8] R Neutze et al. "Potential for biomolecular imaging with femtosecond X-ray pulses." In: *Nature* 406.6797 (Jan. 2000), pp. 752–757. DOI: 10.1038/35021099.

[9] Janos Hajdu. "Single-molecule X-ray diffraction". In: *Current Opinion in Structural Biology* 10.5 (Jan. 2000), pp. 569–573. DOI: 10.1016/S0959-440X(00)00133-0.

[10] G Huldt, A. Szoke, and J Hajdu. "Diffraction imaging of single particles and biomolecules". In: *Journal of Structural Biology* 144.1-2 (2003), pp. 219–227. DOI: 10.1016/j.jsb.2003.09.025.

[11] Richard Neutze et al. "Potential impact of an X-ray free electron laser on structural biology". In: *Radiation Physics and Chemistry*. Vol. 71. 3-4. Oct. 2004, pp. 905–916. DOI: 10.1016/j.radphyschem.2004.04.121.

[12] Kelly J Gaffney and Henry N Chapman. "Imaging Atomic Structure and Dynamics with Ultrafast X-ray Scattering". In: *Science* 316.5830 (June 2007), pp. 1444–1449. DOI: 10.1126/science.1135923.

[13] Stephan Stern et al. "Toward atomic resolution diffractive imaging of isolated molecules with X-ray free-electron lasers". In: *Faraday Discuss.* 171 (Aug. 2014), pp. 393–418. DOI: 10.1039/C4FD00028E. arXiv: 1403.2553.

[14] J. Miao et al. "Beyond crystallography: Diffractive imaging using coherent x-ray light sources". In: *Science* 348.6234 (2015), pp. 530–535. DOI: 10.1126/science.aaa1394.

[15] Benedikt J. Daurer et al. "Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses". In: *IUCrJ* 4.3 (May 2017), pp. 251–262. DOI: 10.1107/S2052252517003591.

[16] R.W. Schoenlein. *New Science Opportunities Enabled By LCLS-II X-Ray Lasers*. Tech. rep. 2015, pp. 1–189.

[17] M Marvin Seibert et al. "Single mimivirus particles intercepted and imaged with an X-ray laser". In: *Nature* 470.7332 (2011), pp. 78–81. DOI: 10.1038/nature09748. arXiv: NIHMS150003.

[18] Tomas Ekeberg et al. "Three-dimensional reconstruction of the giant mimivirus particle with an X-ray free-electron laser". In: *Physical Review Letters* 114.9 (Mar. 2015), p. 98102. DOI: 10.1103/PhysRevLett.114.098102.

[19] Chun Hong Yoon et al. "A comprehensive simulation framework for imaging single particles and biomolecules at the European X-ray Free-Electron Laser." In: *Scientific reports* 6.24791 (Apr. 2016), p. 24791. DOI: 10.1038/srep24791.

[20] C Fortmann-Grote et al. "SIMEX: Simulation of Experiments at Advanced Light Sources". In: *arXiv:1610.05980 [physics.comp-ph]* (2016). arXiv: 1610.05980.

[21] Henry N. Chapman et al. "Femtosecond diffractive imaging with a soft-X-ray free-electron laser". In: *Nature Physics* 2.12 (Dec. 2006), pp. 839–843. DOI: 10.1038/nphys461. arXiv: 0610044 [physics].

[22] Richard A Kirian et al. "Femtosecond protein nanocrystallography—data analysis methods". In: *Optics Express* 18.6 (Jan. 2010), pp. 5713–5723. DOI: 10.1364/OE.18.005713.

[23] Henry N Chapman et al. "Femtosecond X-ray protein nanocrystallography." In: *Nature* 470.7332 (Feb. 2011), pp. 73–77. DOI: 10.1038/nature09750.

[24] Petra Fromme and John C H Spence. "Femtosecond nanocrystallography using X-ray lasers for membrane protein structure determination". In: *Current Opinion in Structural Biology* 21.4 (Aug. 2011), pp. 509–516. DOI: 10.1016/j.sbi.2011.06.001.

[25] Sébastien Boutet et al. "High-resolution protein structure determination by serial femtosecond crystallography." In: *Science (New York, N.Y.)* 337.6092 (July 2012), pp. 362–4. DOI: 10.1126/science.1217737.

[26] Henry N Chapman, Carl Caleman, and Nicusor Timneanu. "Diffraction before destruction." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369.1647 (2014), p. 20130313. DOI: 10.1098/rstb.2013.0313.

[27] Ilme Schlichting. "Serial femtosecond crystallography: The first five years". In: *IUCrJ* 2.2 (Mar. 2015), pp. 246–255. DOI: 10.1107/S205225251402702X.

[28] Philip Roedig et al. "High-speed fixed-target serial virus crystallography". In: *Nature Methods* (June 2017). DOI: 10.1038/nmeth.4335.

[29] V. L. Shneerson, A. Ourmazd, and D. K. Saldin. "Crystallography without crystals. I. The common-line method for assembling a three-dimensional diffraction volume from single-particle scattering". In: *Acta Crystallographica Section A: Foundations of Crystallography* 64.2 (Mar. 2008), pp. 303–315. DOI: 10.1107/S0108767307067621. arXiv: 0710.2561.

[30] Ne Te Duane Loh and Veit Elser. "Reconstruction algorithm for single-particle diffraction imaging experiments". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 80.2 (Aug. 2009), p. 26705. DOI: 10.1103/PhysRevE.80.026705. arXiv: 0904.2581.

[31] Michał Walczak and Helmut Grubmüller. "Bayesian orientation estimate and structure information from sparse single-molecule x-ray diffraction images". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 90.2 (Aug. 2014), p. 22714. DOI: 10.1103/PhysRevE.90.022714.

[32] Julien Flamant et al. "Expansion-maximization-compression algorithm with spherical harmonics for single particle imaging with X-ray lasers". In: *Physical Review E, Volume 93, Issue 5, id.053302* 93.5 (Feb. 2016). DOI: 10.1103/PhysRevE.93.053302. arXiv: 1602.01301.

[33] Stephan Kassemeyer et al. "Optimal mapping of x-ray laser diffraction patterns into three dimensions using routing algorithms". In: *Physical Review E* 88.4 (Oct. 2013), p. 042710. DOI: 10.1103/PhysRevE.88.042710.

[34] Hyung Joo Park et al. "Toward unsupervised single-shot diffractive imaging of heterogeneous particles using X-ray free-electron lasers". In: *Optics Express* 21.23 (Nov. 2013), p. 28729. DOI: 10.1364/OE.21.028729.

[35] Veit Elser. "Three-dimensional structure from intensity correlations". In: *New Journal of Physics* 13.12 (Dec. 2011), p. 123014. DOI: 10.1088/1367-2630/13/12/123014. arXiv: 1107.4030.

[36] Jeffery J Donatelli, James A Sethian, and Peter H Zwart. "Reconstruction from limited single-particle diffraction data via simultaneous determination of state, orientation, intensity, and phase". In: *PNAS* (June 2017), p. 201708217. DOI: 10.1073/pnas.1708217114.

[37] Russell Fung et al. "Structure from fleeting illumination of faint spinning objects in flight". In: *Nature Physics* 5.1 (Jan. 2008), pp. 64–67. DOI: 10.1038/nphys1129. arXiv: 0806.2341.

[38] Brian Moths and Abbas Ourmazd. "Bayesian algorithms for recovering structure from single-particle diffraction snapshots of unknown orientation: A comparison". In: *Acta Crystallographica Section A: Foundations of Crystallography* 67.5 (2011), pp. 481–486. DOI: 10.1107/S0108767311019611. arXiv: 1005.0640.

[39] Peter Schwander et al. "The symmetries of image formation by scattering. II. Applications". In: *Optics Express* 20.12 (2012), pp. 12827–12849. DOI: 10.1364/OE.20.012827. arXiv: 1009.5035.

[40] Dimitrios Giannakis, Peter Schwander, and Abbas Ourmazd. "The symmetries of image formation by scattering. I. Theoretical framework". In: *Optics Express* 20.12 (June 2012), pp. 12799–12826. DOI: 10.1364/OE.20.012799. arXiv: 1009.5035.

[41] Jörg Enderlein. "Maximum-likelihood criterion and single-molecule detection". In: *Applied Optics* 34.3 (Jan. 1995), p. 514. DOI: 10.1364/AO.34.000514.

[42] D K Saldin et al. "Structure of isolated biomolecules obtained from ultrashort x-ray pulses: exploiting the symmetry of random orientations." In: *Journal of physics. Condensed matter : an Institute of Physics journal* 21.13 (Jan. 2009), p. 134014. DOI: 10.1088/0953-8984/21/13/134014.

[43] D. K. Saldin et al. "Beyond small-angle x-ray scattering: Exploiting angular correlations". In: *Physical Review B* 81.17 (Apr. 2010), pp. 1–6. DOI: 10.1103/PhysRevB.81.174105. arXiv: 0505174 [physics].

[44] D K Saldin et al. "Structure of a single particle from scattering by many particles randomly oriented about an axis: a new route to structure determination?" In: *New Journal of Physics* 12 (Jan. 2010), p. 35014.

[45] D. K. Saldin et al. "New light on disordered ensembles: Ab initio structure determination of one particle from scattering fluctuations of many copies". In: *Physical Review Letters* 106.11 (Mar. 2011), p. 115501. DOI: 10.1103/PhysRevLett.106.115501.

[46] Zvi Kam. "The reconstruction of structure from electron micrographs of randomly oriented particles". In: *Journal of Theoretical Biology* 82.1 (Jan. 1980), pp. 15–39. DOI: 10.1016/0022-5193(80)90088-0.

[47] D. Starodub et al. "Single-particle structure determination by correlations of snapshot X-ray diffraction patterns." In: *Nature communications* 3.May (Dec. 2012), p. 1276. DOI: 10.1038/ncomms2288.

[48] Dilano Saldin et al. "Reconstructing an Icosahedral Virus from Single-Particle Diffraction Experiments". In: *Optics Express* 19.18 (Jan. 2011), p. 18. DOI: 10.1364/OE.19.017318. arXiv: 1107.5212.

[49] H. C. Poon and D. K. Saldin. "Use of triple correlations for the sign determinations of expansion coefficients of symmetric approximations to the diffraction volumes of regular viruses". In: *Structural Dynamics* 2.4 (July 2015), p. 041716. DOI: 10.1063/1.4922476.

[50] Ian C Gray and Michael R Barnes. "Amino Acid Properties and Consequences of Substitutions". In: *Bioinformatics for Geneticists.* 4 (2003), pp. 289–304. DOI: 10.1002/0470867302.ch14.

[51] Claire O'Connor and Jill U Adams. "Essentials of Cell Biology". In: *Nature education* (2010), pp. 1–100. DOI: 10.1145/634067.634339.

[52] Hanno Steen and Matthias Mann. "The ABC's (and XYZ's) of peptide sequencing." In: *Nature reviews. Molecular cell biology* 5.9 (Sept. 2004), pp. 699–711. DOI: 10.1038/nrm1468. arXiv: arXiv:1011.1669v3.

[53] Robert Zwanzig, Attila Szabo, and Biman Bagchi. "Levinthal's paradox." In: *Proceedings of the National Academy of Sciences of the United States of America* 89.1 (1992), pp. 20–2. DOI: 10.1073/pnas.89.1.20.

[54] Heinrich Terlau and Frank Kirchhoff. "Ion Channels/Excitable Membranes". In: *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* (2001), pp. 913–916. DOI: 10.1007/3-540-29623-9_5640.

[55] Kuniaki Takata, Toshiyuki Matsuzaki, and Yuki Tajika. "Aquaporins: Water channel proteins of the cell membrane". In: *Progress in Histochemistry and Cytochemistry* 39.1 (2004), pp. 1–83. DOI: 10.1016/j.proghi.2004.03.001.

[56] Daniel Ungar and Frederick M Hughson. "SNARE protein structure and function." In: *Annual review of cell and developmental biology* 19 (2003), pp. 493–517. DOI: 10.1146/annurev.cellbio.19.110701.155609.

[57] The royal Swedish Academy of sciences. *Structure and function of the ribosome*. Tech. rep. October. 2009, pp. –22.

[58] D. W. J. Cruickshank. "Time-Resolved Macromolecular Crystallography: Introductory Remarks and a Little History". In: *Philosophical Transactions: Physical Sciences and Engineering Time-Resolved Macromolecular Crystallography* 340.1657 (1992), pp. 169–173.

[59] Alexi Assmus. "Early history of X rays". In: *Beam Line* 25.2 (1995), pp. 10–24.

[60] Herbert Hauptman. "Phasing methods for protein crystallography". In: *Current Opinion in Structural Biology* 7.5 (Oct. 1997), pp. 672–680. DOI: 10.1016/S0959-440X(97)80077-2.

[61] Isabel Usón and George M Sheldrick. "Advances in direct methods for protein crystallography". In: *Current Opinion in Structural Biology* 9.5 (Oct. 1999), pp. 643–648. DOI: 10.1016/S0959-440X(99)00020-2. arXiv: arXiv:1011.1669v3.

[62] Garry Taylor. "The phase problem". In: *Acta Crystallographica - Section D Biological Crystallography*. Vol. 59. 11. International Union of Crystallography, Nov. 2003, pp. 1881–1890. DOI: 10.1107/S0907444903017815.

[63] Steven E Ealick. "Advances in multiple wavelength anomalous diffraction crystallography". In: *Current Opinion in Chemical Biology* 4.5 (Oct. 2000), pp. 495–499. DOI: 10.1016/S1367-5931(00)00122-8.

[64] C. Lecomte et al. "Ultra-high-resolution X-ray structure of proteins". In: *Cellular and Molecular Life Sciences* 61.7-8 (Apr. 2004), pp. 774–782. DOI: 10.1007/s00018-003-3405-0.

[65] Lynmarie K Thompson. "Solid-state NMR studies of the structure and mechanisms of proteins". In: *Current Opinion in Structural Biology* 12.5 (Oct. 2002), pp. 661–669. DOI: 10.1016/S0959-440X(02)00374-3.

[66] K Wüthrich. "Acids., NMR with proteins and nucleic". In: *Europhysics News* 17 (1986), pp. 11–13.

[67] K Wüthrich. "Protein structure determination in solution by nuclear magnetic resonance spectroscopy". In: *Science 243, 45-50* 243.4887 (1989).

[68] G M Clore and A M Gronenborn. "Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy". In: *Science (New York, N.Y.)* 252.5011 (1991), pp. 1390–1399. DOI: 10.1126/science.2047852.

[69] The Royal Swedish Academy of Sciences. "Mass spectrometry (MS) and nuclear magnetic resonance (NMR) applied to biological macromolecules". In: *October* 1.October (2002), pp. 1–13.

[70] Claus Schneider et al. “Enzymatic synthesis of a bicyclobutane fatty acid by a hemoprotein lipoxygenase fusion protein from the cyanobacterium Anabaena PCC 7120.” In: *Proceedings of the National Academy of Sciences of the United States of America* 104.48 (Nov. 2007), pp. 18941–18945. DOI: 10.1073/pnas.0707148104.

[71] Timothy F Havel, Gordon M Crippen, and D Irwin. “Effects of Distance Constraints on Macromolecular Conformation . 11 . Simulation of Experimental Results and Theoretical Predictions”. In: *Biopolymers* 18 (1979), pp. 73–81. DOI: 10.1002/bip.1979.360180108.

[72] Andreas G. Tzakos et al. “NMR techniques for very large proteins and rnas in solution.” In: *Annu Rev Biophys Biomol Struct* 35.1 (June 2006), pp. 319–342. DOI: 10.1146/annurev.biophys.35.040405.102034.

[73] Remco Sprangers and Lewis E Kay. “Quantitative dynamics and binding studies of the 20S proteasome by NMR.” In: *Nature* 445.7128 (2007), pp. 618–622. DOI: 10.1038/nature05512.

[74] Martin Mechelke and Michael Habeck. “Bayesian weighting of statistical potentials in NMR structure calculation”. In: *PLoS ONE* 9.6 (June 2014). Ed. by Bruce R. Donald, e100197. DOI: 10.1371/journal.pone.0100197.

[75] Liguo Wang and Fred J. Sigworth. “Cryo-EM and single particles.” In: *Physiology* 21.1 (2006), pp. 13–8. DOI: 10.1152/physiol.00045.2005.

[76] Ernst Ruska. “Nobel lecture. The development of the electron microscope and of electron microscopy.” In: *Bioscience reports* 7.8 (July 1987), pp. 607–629. DOI: 10.1103/RevModPhys.59.627.

[77] Manfred von Ardenne. “Das Elektronen-Rastermikroskop, Theoretische Grundlagen”. In: *Z. Physik* 109.9. u. 10 (1938), pp. 553–572.

[78] Manfred von Ardenne. “Praktische Ausführung - Das Elektronen-Rastermikroskop”. In: *Zeitschrift für technische Physik* 19 (1938), pp. 407–416.

[79] Yifan Cheng. “Single-particle Cryo-EM at crystallographic resolution”. In: *Cell* 161.3 (2015), pp. 450–457. DOI: 10.1016/j.cell.2015.03.049. arXiv: 15334406.

[80] Holger Stark, Friedrich Zemlin, and Christoph Boettcher. “Electron radiation damage to protein crystals of bacteriorhodopsin at different temperatures”. In: *Ultramicroscopy* 63.2 (June 1996), pp. 75–79. DOI: 10.1016/0304-3991(96)00045-9.

[81] Robert M Glaeser. “How good can cryo-EM become?” In: *Nature Methods* 13.1 (2016), pp. 28–32. DOI: 10.1038/nmeth.3695.

[82] Xiao chen Bai, Greg McMullan, and Sjors H.W Scheres. "How cryo-EM is revolutionizing structural biology". In: *Trends in Biochemical Sciences* 40.1 (Jan. 2015), pp. 49–57. DOI: 10.1016/j.tibs.2014.10.005.

[83] Niels Fischer et al. "Structure of the E. coli ribosome–EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM". In: *Nature* 520.7548 (Feb. 2015), pp. 567–570. DOI: 10.1038/nature14275.

[84] Niels Fischer et al. "Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy." In: *Nature* 466.7304 (2010), pp. 329–333. DOI: 10.1038/nature09206.

[85] Lars Bock. "Dynamics and Driving Forces of Macromolecular Complexes". In: *PhD Thesis* (2012).

[86] Lars V Bock et al. "Energy barriers and driving forces in tRNA translocation through the ribosome". In: *Nat Struct Mol Biol* 20.12 (Nov. 2013), pp. 1390–1396. DOI: 10.1038/nsmb.2690\rhttp://www.nature.com/nsmb/journal/v20/n12/abs/nsmb.2690.html#supplementary-information.

[87] Michael J. Bogan et al. "Single particle X-ray diffractive imaging". In: *Nano Letters* 8.1 (2008), pp. 310–316. DOI: 10.1021/nl072728k.

[88] Daniel James. "Injection Methods and Instrumentation for Serial X-ray Free Electron Laser Experiments". PhD thesis. 2015. DOI: 10.1007/s13398-014-0173-7.2. arXiv: arXiv:1011.1669v3.

[89] Garrett Nelson. "Sample Injector Fabrication and Delivery Method Development for Serial Crystallography using Synchrotrons and X-ray Free Electron Lasers". PhD thesis. 2015, p. 157.

[90] Uwe Weierstall et al. "Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography." In: *Nature communications* 5 (2014), p. 3309. DOI: 10.1038/ncomms4309. arXiv: NIHMS150003.

[91] Ludger Inhester, Gerrit Groenhof, and Helmut Grubmueller. "Auger Spectrum of a Water Molecule after Single and Double Core-Ionization by Intense X-Ray Radiation". In: *Biophysical Journal* 102.3 (Jan. 2012), 392A–392A.

[92] C Pellegrini. "The history of X-ray free-electron lasers". In: *European Physical Journal H* 37.5 (2012), pp. 659–708. DOI: 10.1140/epjh/e2012-20064-5.

[93] Adrian P. Mancuso et al. *Technical Design Report: Scientific Instrument Single Particles, Clusters, and Biomolecules (SPB)*. Tech. rep. 2013, pp. 1–232. DOI: 10.3204/XFEL.EU/TR-2013-004.

[94] C Pellegrini, A Marinelli, and S Reiche. "The physics of x-ray free-electron lasers". In: *Reviews of Modern Physics* 88.1 (2016). DOI: 10.1103/RevModPhys.88.015006.

[95] Yuichi Inubushi et al. "Determination of the Pulse Duration of an X-Ray Free Electron Laser Using Highly Resolved Single-Shot Spectra". In: *Physical Review Letters* 109.14 (Oct. 2012), p. 144801. DOI: 10.1103/PhysRevLett.109.144801.

[96] P. Emma et al. "First lasing and operation of an ångstrom-wavelength free-electron laser". In: *Nature Photonics* 4.9 (Sept. 2010), pp. 641–647. DOI: 10.1038/nphoton.2010.176.

[97] European X-Ray Free-Electron Laser Facility GmbH. *European XFEL - Facts & Figures*. 2017.

[98] Christoph Bostedt et al. "Linac Coherent Light Source: The first five years". In: *Reviews of Modern Physics* 88.1 (Mar. 2016), p. 015007. DOI: 10.1103/RevModPhys.88.015007.

[99] Richard A. Kirian. "Structure determination through correlated fluctuations in x-ray scattering". In: *Journal of Physics B: Atomic, Molecular and Optical Physics* 45.22 (Nov. 2012), p. 223001. DOI: 10.1088/0953-4075/45/22/223001.

[100] H. B. Stuhrmann. "Interpretation of small angle scattering functions of dilute solutions and gases". In: *Acta Crystallographica Section A* 26.3 (May 1970), pp. 297–306. DOI: 10.1107/S0567739470000748.

[101] D. I. Svergun, H. B. Stuhrmann, and IUCr. "New developments in direct shape determination from small-angle scattering." In: *Acta Crystallographica Section A Foundations of Crystallography* 47.6 (Nov. 1991), pp. 736–744. DOI: 10.1107/S0108767391006414.

[102] Wolfgang Demtröder. "Experimentalphysik 2". In: *Elektrizit{ä}t und Optik, Kapitel* 10 (1995).

[103] Robin Santra. "Concepts in x-ray physics". In: *Journal of Physics B: Atomic, Molecular and Optical Physics* 42.2 (Jan. 2009), p. 023001. DOI: 10.1088/0953-4075/42/2/023001.

[104] R Santra. "Corrigendum for Concepts in x-ray physics". In: *Journal of Physics B: Atomic, Molecular and Optical Physics* 42.16 (Aug. 2009), pp. 169801–169801. DOI: 10.1088/0953-4075/42/16/169801.

[105] J R Fienup. "Phase retrieval algorithms: a comparison." In: *Appl. Opt.* 21.15 (Jan. 1982), pp. 2758–69. DOI: 10.1364/AO.21.002758. arXiv: 1403.3316.

[106] Jim Fienup and C C Wackerman. "Phase retrieval stagnation problems and solutions". In: *Journal of Optical Society of America A* 3.11 (Jan. 1986), pp. 1897–1907. DOI: 10.1364/JOSAA.3.001897.

[107] D. Russell Luke. "Relaxed Averaged Alternating Reflections for Diffraction Imaging". In: *Inverse Problems* 37.1 (Feb. 2004), p. 13. DOI: 10.1088/0266-5611/21/1/004. arXiv: 0405208 [math].

[108] Yoav Shechtman et al. "Phase Retrieval with Application to Optical Imaging: A contemporary overview". In: *IEEE Signal Processing Magazine* 32.3 (May 2015), pp. 87–109. DOI: 10.1109/MSP.2014.2352673. arXiv: 1402.7350.

[109] J. H. Hubbell et al. "Atomic form factors, incoherent scattering functions, and photon scattering cross sections". In: *Journal of Physical and Chemical Reference Data* 4.3 (July 1975), pp. 471–538. DOI: 10.1063/1.555523.

[110] G. Bortel, G. Faigel, and M. Tegze. "Classification and averaging of random orientation single macromolecular diffraction patterns at atomic resolution". In: *Journal of Structural Biology* 166.2 (May 2009), pp. 226–233. DOI: 10.1016/j.jsb.2009.01.005.

[111] Anna Munke et al. "Coherent diffraction of single Rice Dwarf virus particles using hard X-rays at the Linac Coherent Light Source". In: *Scientific Data* 3 (2016), p. 160064. DOI: 10.1038/sdata.2016.64.

[112] Miklós Tegze and Gábor Bortel. "Atomic structure of a single large biomolecule from diffraction patterns of random orientations". In: *Journal of Structural Biology* 179.1 (July 2012), pp. 41–45. DOI: 10.1016/j.jsb.2012.04.014. arXiv: 1204.2102.

[113] Jeffrey J Donatelli, Peter H Zwart, and James A Sethian. "Iterative phasing for fluctuation X-ray scattering". In: *Proceedings of the National Academy of Sciences* 112.33 (Aug. 2015), pp. 10286–10291. DOI: 10.1073/pnas.1513738112. arXiv: arXiv:1404.2263v1.

[114] Derek Mendez et al. "Angular correlations of photons from solution diffraction at a free-electron laser encode molecular structure". In: *IUCrJ* 3.6 (Nov. 2016), pp. 420–429. DOI: 10.1107/S2052252516013956.

[115] Haiguang Liu et al. "Three-dimensional single-particle imaging using angular correlations from X-ray laser data". In: *Acta Crystallographica Section A* 69.4 (July 2013), pp. 365–373. DOI: 10.1107/S0108767313006016.

[116] Veit Elser. "Strategies for processing diffraction data from randomly oriented particles". In: *Ultramicroscopy* 111.7 (Jan. 2011), pp. 788–792. DOI: 10.1016/j.ultramic.2010.10.014. arXiv: 1007.3777.

[117] Natalie Baddour. "Operational and convolution properties of three-dimensional Fourier transforms in spherical polar coordinates". In: *Journal of the Optical Society of America A* 27.10 (Oct. 2010), p. 2144. DOI: 10.1364/JOSAA.27.002144.

[118] E. P. Wigner. "On the Matrices Which Reduce the Kronecker Products of Representations of S. R. Groups". In: *Quantum Theory of Angular Momentum*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1965, pp. 87–133. DOI: 10.1007/978-3-662-02781-3_42.

[119] Francesco Mezzadri. "How to generate random matrices from the classical compact groups". In: *arXiv:math-ph/0609050* (Sept. 2006). arXiv: 0609050 [math-ph].

[120] Jeffrey S Rosenthal. "Random rotations: characters and random walks on SO(N)". In: *The Annals of Probability* 22.1 (Jan. 1994), pp. 398–423. DOI: 10.1214/aop/1176988864.

[121] O. Klein and T. Nishina. "Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac". In: *Zeitschrift für Physik* 52.11-12 (Nov. 1929), pp. 853–868. DOI: 10.1007/BF01366453.

[122] A Hamzawy. "Compton scattering I: Angular distribution and polarization degree". In: *Radiation Physics and Chemistry* 119 (2016), pp. 103–108. DOI: 10.1016/j.radphyschem.2015.10.003.

[123] Andrea Schmidt et al. "Crystal structure of small protein crambin at 0.48 Å resolution". In: *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 67.4 (Apr. 2011), pp. 424–428. DOI: 10.1107/S1744309110052607.

[124] Julian C H Chen et al. "Room-temperature ultrahigh-resolution time-of-flight neutron and X-ray diffraction studies of H/D-exchanged crambin". In: *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 68.2 (Feb. 2012), pp. 119–123. DOI: 10.1107/S1744309111051499.

[125] Marin Van Heel and Michael Schatz. "Fourier shell correlation threshold criteria". In: *Journal of Structural Biology* 151.3 (2005), pp. 250–262. DOI: 10.1016/j.jsb.2005.05.009.

[126] Anton Barty, Jochen Küpper, and Henry N. Chapman. "Molecular Imaging Using X-Ray Free-Electron Lasers". In: *Annual Review of Physical Chemistry* 64.1 (Apr. 2013), pp. 415–435. DOI: 10.1146/annurev-physchem-032511-143708.

[127] Stefan W. Hell and Jan Wichmann. "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy". In: *Optics Letters* 19.11 (June 1994), p. 780. DOI: 10.1364/OL.19.000780. arXiv: NIHMS150003.

[128] Francisco Balzarotti et al. "Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes". In: *Science* 355.6325 (2016). DOI: 10.1126/science.aak9913. arXiv: 1611.03401.

[129] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[130] Radford M Neal et al. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo* 2.11 (2011).

[131] E Marinari et al. "Simulated Tempering: A New Monte Carlo Scheme". In: *Europhys. Lett. Europhys. Lett* 19.196 (July 1992), pp. 451–458. DOI: 10.1209/0295-5075/19/6/002. arXiv: 9205018 [hep-lat].

[132] Bernd A. Berg and Alain Billoire. "Markov Chain Monte Carlo Simulations". In: *Wiley Encyclopedia of Computer Science and Engineering*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Dec. 2007. DOI: 10.1002/9780470050118.ecse696.

[133] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (Mar. 1951), pp. 79–86. DOI: 10.1214/aoms/1177729694. arXiv: 1511.00860.

[134] E. Hellinger. "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen". In: *Journal fur die Reine und Angewandte Mathematik* 1909.136 (1909), pp. 210–271. DOI: 10.1515/crll.1909.136.210.

[135] Carl Edward Rasmussen. "The Infinite Gaussian Mixture Model". In: *Advances in Neural Information Processing Systems 12* (2000), pp. 554–560.

[136] Tejal Bhamre, Teng Zhang, and Amit Singer. "Orthogonal matrix retrieval in cryo-electron microscopy". In: *Proceedings - International Symposium on Biomedical Imaging*. Vol. 2015-July. IEEE, Apr. 2015, pp. 1048–1052. DOI: 10.1109/ISBI.2015.7164051. arXiv: arXiv:1412.0494v1.

[137] Filipe R N C Maia. "The Coherent X-ray Imaging Data Bank". In: *Nature Methods* 9.9 (Aug. 2012), pp. 854–855. DOI: 10.1038/nmeth.2110.

[138] Jochen Küpper et al. "X-ray diffraction from isolated and strongly aligned gas-phase molecules with a free-electron laser". In: *Physical Review Letters* 112.8 (Feb. 2014), p. 083002. DOI: 10.1103/PhysRevLett.112.083002. arXiv: 1307.4577.

[139] Miguel a. Blanco, M. Flórez, and M. Bermejo. "Evaluation of the rotation matrices in the basis of real spherical harmonics". In: *Journal of Molecular Structure: THEOCHEM* 419.1-3 (Jan. 1997), pp. 19–27. DOI: 10.1016/S0166-1280(97)00185-1.

[140] Herbert H.H. Homeier and E.Otto Steinborn. "Some properties of the coupling coefficients of real spherical harmonics and their relation to Gaunt coefficients". In: *Journal of Molecular Structure: THEOCHEM* 368 (Jan. 1996), pp. 31–37. DOI: 10.1016/S0166-1280(96)90531-X.

[141] Peter J Kostelec et al. "Computational Harmonic Analysis for Tensor Fields on the Two-Sphere". In: *Journal of Computational Physics* 162.2 (Aug. 2000), pp. 514–535. DOI: 10.1006/jcph.2000.6551.

[142] Henrik Skibbe et al. "Fast computation of 3D spherical Fourier harmonic descriptors - A complete orthonormal basis for a rotational invariant representation of three-dimensional objects". In: *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009* (Jan. 2009), pp. 1863–1869. DOI: 10.1109/ICCVW.2009.5457509.

[143] Thomas Wieder. "A generalized Debye scattering formula and the Hankel transform". In: *Zeitschrift fur Naturforschung - Section A Journal of Physical Sciences* 54.2 (Feb. 1999), pp. 124–130.

[144] Masayuki Toyoda and Taisuke Ozaki. "Fast spherical Bessel transform via fast Fourier transform and recurrence formula". In: *Computer Physics Communications* 181.2 (Jan. 2010), pp. 277–282. DOI: 10.1016/j.cpc.2009.09.020.

[145] L Yu et al. "Quasi-discrete Hankel transform." In: *Optics letters* 23.6 (Mar. 1998), pp. 409–411. DOI: 10.1364/OL.23.000409.

Acknowledgement

First and foremost, I would like to thank my supervisor Prof. Dr. Helmut Grubmüller for giving me the opportunity to work on this project. He provided the initial idea of using photon correlations and offered an inspiring collaboration with many interesting and helpful discussions that motivated me throughout the project. I greatly appreciate the time he invested to also teach me the virtues of good scientific conduct, including valuable techniques for thinking, presenting and writing. The work with him was always productive and immensely enjoyable and I am very thankful that I could learn from his experience and expertise.

A big thanks goes to the two member of my thesis committee, Prof. Dr. Holger Stark and Prof. Dr. Marcus Müller, for providing helpful questions and comments during the meetings. Holger Stark further shared his knowledge on structure determination with cryo-EM and gave helpful insights on noise and structure determination algorithms.

A big thanks goes to Michal Walczak for all the feedback and idea discussions and for sharing his know-how on single molecule scattering. I also like to thank Martin Mechelke who continuously and productively challenged ideas and concepts and helped me greatly with improving my concepts and ideas. The numerous discussions led to many detailed improvements without the project wouldn't have turned out this way.

I also like to thank Christian Blau, Carsten Kutzner, Lars Bock, Carl Burmeister, Andreas Volkhardt, Leonard Heinz and Frank Wiederschein for their practical tips with all kinds of computer and math-related problems and Petra Keller for sharing her knowledge on scientific writing and for proofreading manuscripts.

Further I would like to thank Eveline Heinemann for providing so much help with organizing the scientific everyday business including going to conferences and scheduling meetings. I would like to also thank our system administrations Martin Fechner and Ansgar Esztermann for providing and maintaining the computer resources that were necessary for this project. Numerous bugs and problems occurred that they always readily and happily fixed.

My final thanks goes to my colleagues and former colleagues in the department of Computational and Theoretical Biophysics at the Max-Planck Institute for Biophysics. I enjoyed a kind and welcoming atmosphere that brought a constant exchange of knowledge on science, computer techniques and non-work related topics.

Last but not least I want to acknowledge all contributors to free and open source projects and the authors of the Wikipedia articles that helped me on a daily basis, in particular the authors of the Julia scripting language, the Jupyter development environment, the Matplotlib library, the s2kit spherical harmonics library, ShareLaTeX, Inkscape and Chimera.

Vita

Personal Data

Benjamin von Ardenne
Brauweg 22
37073 Göttingen

E-Mail: benjamin.von.ardenne@gmail.com

born 01/19/1988 in Dresden

Education

2012 - 2017	PhD at Max-Planck Institute for Biophysical Chemistry, Göttingen, International Max Planck Research Schools (IMPRS), "Physics of Biological and Complex Systems"
2010 - 2012	Master of Science, Physics, University of Göttingen, Max-Planck Institute for Biophysical Chemistry
2007 - 2010	Bachelor of Science, Physics, University of Göttingen
1994 - 2006	Abitur at Evangelisches Kreuzgymnasiums Dresden

Göttingen, 10 August 2017