UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1905

Towards Fast and Robust Algorithms in Flash X-ray single-particle Imaging

JING LIU

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2020

ISSN 1651-6214
ISBN 978-91-513-0877-7
urn:nbn:se:uu:diva-403878

Dissertation presented at Uppsala University to be publicly examined in BMC B41, Dag Hammarskjölds väg, Uppsala, Tuesday, 31 March 2020 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Helmut Grubmüller (Max Planck Institute for Biophysical Chemistry, Göttingen, Germany).

Abstract

Liu, J. 2020. Towards Fast and Robust Algorithms in Flash X-ray single-particle Imaging. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1905. 79 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0877-7.

Modern X-ray Free Electron Laser (XFEL) technology provides the possibility to acquire a large number of diffraction patterns from individual biological nano-particles, including proteins, viruses, and DNA. Ideally, the collected data frames are high-quality single-particle diffraction patterns. However, unfortunately, the raw dataset is noisy and also contains patterns with scatterings from multiple particles, contaminated particles, etc. The data complexity and the massive volumes of raw data make pattern selection a time-consuming and challenge task. Further, X-rays interact with particles at random and the captured patterns are the 2D intensities of the scattered waves, i.e. we cannot observe the particle orientations and the phase information from the 2D diffraction patterns. To reconstruct 2D diffraction patterns into 3D structures of the desired particle, we need a sufficiently large single-particle-pattern dataset. The computational methodology for this reconstruction task is still under development and in need of an improved understanding of the algorithmic uncertainties.

In this thesis, we tackle some of the challenges to obtain 3D structures of sample molecules from single-particle diffraction patterns. First, we have developed two classification methods to select single-particle diffraction patterns that are similar to provided templates. Second, we have accelerated the 3D reconstruction procedures by distributing the computations among Graphics Processing Units (GPUs) and by proposing an adaptive discretization of 3D space. Third, to better understand the uncertainties of the 3D reconstruction procedure, we have evaluated the impact of the different sources of resolution-limiting factors and introduced a practically applicable computational methodology in the form of bootstrap procedures for assessing the reconstruction uncertainty. These technologies form a data-analysis pipeline for recovering 3D structures from the raw X-ray single-particle data, which also analyzes the uncertainties. With the experimental developments of the X-ray single-particle technology, we expect that the data volumes will be increasing sharply, and hence, we believe such a computational pipeline will be critical to retrieve particle structures in the achievable resolution.

Keywords: X-ray Free Electron lasers (XFELs); 3D electron density determination; Machine learning; Image processing; High-performance computing; GPUs; Uncertainty quantification; X-ray single-particle Imaging; Flash X-ray single-particle diffraction Imaging (FXI);

Jing Liu, Department of Cell and Molecular Biology, Molecular biophysics, Box 596, Uppsala University, SE-75124 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0877-7
urn:nbn:se:uu:diva-403878 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-403878)

Dedicated to Gang & Elin

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **J. Liu**, G. van der Schot, and S. Engblom. (2019). Supervised classification methods for flash X-ray single particle diffraction imaging. Optics express, 27(4), 3884-3899. https://doi.org/10.1364/OE.27.003884 [49].
Contribution: J. Liu proposed and implemented the methods, and performed the numerical experiments. J. Liu also drafted the paper, and participated in revising the paper.

II T. Ekeberg, S. Engblom, and **J. Liu**. (2015). Machine learning for ultrafast X-ray diffraction patterns on large-scale GPU clusters. The international journal of high performance computing applications, 29(2), 233-243. http://doi.org/10.1177/1094342015572030 [25].
Contribution: J. Liu conducted the implementations and the numerical experiments, drafted parts of the paper, and participated in revising the paper. The authorship is in alphabetical order.

III **J. Liu**, S. Engblom, and C. Nettelblad. (2018). Assessing uncertainties in X-ray single-particle three-dimensional reconstruction. Physical Review E, 98(1), 013303. http://doi.org/10.1103/PhysRevE.98.013303 [50].
Contribution: J. Liu implemented the proposed methods, performed the numerical experiments, drafted parts of the paper as well as revised the paper.

IV (Manuscript) **J. Liu**, S. Engblom, and C. Nettelblad. Flash X-ray Imaging in 3D: A Proposed data analysis pipeline.
Contribution: J. Liu conceived and developed the major ideas, and also performed the experiments, as well as drafted and revised the manuscripts.

V (Not included) S. Engblom, and **J. Liu**.(2013, September). X-ray laser imaging of biomolecules using multiple GPUs. In

International Conference on Parallel Processing and Applied Mathematics (pp. 480-489). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-55224-3_45 [28]
Contribution: J. Liu conceived and developed the presented idea, and also drafted and revised the paper.

Abbreviations

AMO	Atomic, Molecular and Optical Sciences
API	Application Programming Interface
CPU	Central Processing Unit
CXI	Coherent X-ray Imaging
CXIDB	Coherent X-ray Imaging Data Bank
DFT	Discrete Fourier transform
cryo-EM	Cryo-electron Microscopy
DM	Difference Map algorithm
EM	Expectation Maximization
EMC	Expansion Maximization Compression algorithm
ER	Error Reduction algorithm
FFT	Fast Fourier Transform
FLASH	Free Electron LASer in Hamburg
FLOPs	FLoating Point Operations Per second
FSC	Fourier shell correlation
FT	Fourier transform
FXI	Flash X-ray single-particle diffraction Imaging
I/O	Input and Output
GPU	Graphics Processing Unit
HDF5	Hierarchical Data Format version 5
HIO	Hybrid Input Output
HTC	High Throughput Computing
HPC	High Performance Computing
LCLS	Linac Coherent Light Source
LL	Log Likelihood
ML	Maximum Likelihood
NNMF	None Negative Matrix Factorization
PRTF	Phase Retrieval Transfer Function
RAAR	Relaxed Averaged Alternating Reflections
SASE	Self Amplified Stimulated Emission
SVD	Singular Value Decomposition
MPI	Message Passing Interface
XFEL	X-ray Free-Electron Laser
SACLA	SPring-8 Angstrom Compact Free Electron Laser

Contents

Part I:
Motivation

1. Introduction

The wavelength of the probing light is a fundamental factor that limits our ability to observe small objects. To determine nanoscale structures, X-rays with wavelengths down to a few Angstroms are essential. However, X-rays interact weakly with matter. When illuminating an object with X-rays, only a small fraction of it will be elastically scattered, which provides signals containing the structural information. Other forms of energy transfer, such as Compton scattering and photon absorption, will cause radiation damage and lead to degradation of the scattering signal, resulting in recovered structures in low resolution or with artifacts.

A traditional way to use X-ray in structural biology is X-ray crystallography, and it requires to crystallize samples to a certain minimum size to run out radiation damage and enhance scattering signals. Unfortunately, due to conformational flexibility, not all molecules can be packed, i.e., forming high-quality crystals is hard or even impossible for flexible molecules.

Thanks to the advent of Modern X-ray Free Electron Laser (XFEL) technology [55, 60], we have possibilities to explore biological structure without packing sample molecules into crystals. The so-called "diffract and destroy" [14] strategy uses ultra short and extremely bright X-ray pulses, produced by XFELs, to create interpretable diffraction signals before the samples explode and turn into a plasm [66]. Since then, the approach has caught considerable attention in structural biology [39, 13, 9, 45, 26].

The state-of-the-art method using the "diffract and destroy" strategy is the Flash X-ray single-particle diffraction Imaging (FXI), or sometimes referred to as the X-ray Single-Particle Imaging (SPI) [3]. In an FXI experiment, a stream of particles is injected into the X-ray beam, and hit by the extremely intense X-ray pulses, producing 2D diffraction patterns showing the illuminated objects at random orientations. Due to the high repetition rate of XFELs and the stochastic nature of FXI experiments, the readouts from the digital detectors are of varying qualities, and a large portion consists of empty frames without any scatterings from sample particles. We also obtain a substantial amount of scatterings from contaminants and multiple sample particles. The most interesting readouts are the single-particle diffraction patterns, i.e. the frames containing scatterings from just one sample particle, and

unfortunately, most of the readouts are not single-particle diffraction patterns. Further, the readouts from digital detectors are intensities without phase information and have varying beam intensities from shot to shot.

Since FXI studies relatively small sample particles and capture diffraction intensities in the far-field, the diffraction patterns on the detectors are continuous signals and are in the Fourier domain. Oversampling is therefore used to retrieve phase information [62, 31, 78, 61]. Further, considering that many biological particles exist in identical copies at the resolution scales of relevance, the 2D diffraction patterns can be treated approximately as differently oriented exposures of the same particle, and hence 3D structures can be obtained by averaging the 2D diffraction patterns with the recovered particle orientations. To obtain 3D intensities of sample particles in the real-space, we can perform a two-stage procedure — reconstruct the 3D Fourier intensity first and then retrieve the 3D phase information [53, 16, 7, 76]. Alternatively, it is also possible to combine the phasing algorithms with the rotation determination [22, 48].

Owing to the XFELs facilities and the efforts made for FXI technology, a lot of FXI experiments have been performed with both artificial and non-artificial samples. The first FXI experiment took place at the soft X-ray Free-electron LAser in Hamburg (FLASH) (formerly known as the VUV-FEL) using artificial samples [14]. Later in 2011 [81], FXI experiments with a higher photon flux and harder X-rays succeeded on Mimivirus particles at the LINAC Coherent Light Source (LCLS). Although the resolution achieved was limited to 32 nm for the 2D project images [81], this experiment was still encouraging as a proof-of-concept. Since then, FXI attracted more and more attention in the community. In 2014, another promising 2D structure of carboxysomes [37] was published, and its best resolution was better than the detector-edge resolution. In 2015, a follow-up study on the mimivirus dataset [81] reported the first 3D FXI structure [26] at a full-period resolution of 125 nm. The 3D resolution was remarkedly inferior to the one achieved in 2D from individual patterns, due to the heterogeneity of the mimivirus and the limited number of diffraction patterns. Later, smaller and more homogeneous viruses, the Rice Dwarf Virus (RDV) [64], the PR772 virus [75] and the Paramecium bursaria Chlorella virus (PBCV-1) [72], were selected as sample particles, and then reconstructed to 3D structures [42, 48, 76, 72].

Other than building huge facilities and experimenting on different samples, scientists and engineers are seeking possibilities to improve FXI for higher resolution and quality. The beam focus and wavefront are one of the key issues [5] for FXI experiments, and by optimizing the positions and the angles of mirrors [40, 20], it is possible to have

nano-focusing and Gaussian-like beam shape. The sample delivery technology is also developing as the size of the droplets shrink down to inject single particles more efficiently [21, 82, 46, 37, 79]. Various types of digital detectors are in use to cope with the different beam characteristics at different facilities [84, 8, 35, 1, 44, 43], and software libraries are available for optimizing FXI experiments parameters [94], simulating diffraction patterns [36], monitoring online data [19], data converting [18], data management [88], data bank [57], data preprocessing [4], phase retrieving [56], etc.

In this thesis, we address the following challenges of FXI technology and contribute methods and software to solve them:

- Newer facilities, such as the European XFEL, aim to increase data rates and qualities. The European XFEL [2] is capable of acquiring 27,000 patterns per second — 225 times faster than the Linac Coherent Light Source (LCLS) [12], and more than 450 times faster than the Spring-8 Ångström Compact free-electron LAser (SACLA) [86]. With the increasing XFEL repetition rate and interest in FXI experiments, we foresee huge volumes of FXI data. Unfortunately, due to the stochastic nature of the FXI experiments, the large volumes of the readouts from detectors will include scatterings from contaminated samples to a varying degree with low hit rates [85]. Further, the heterogeneity of the sample biomolecules is unavoidable, especially at high resolution, as biologically essential functions proceed via structural and conformational changes [3]. To study the sample structure, we need to extract single hits from the massive data volumes and deal with sample heterogeneity. Moreover, FXI diffraction patterns are randomly oriented, and most rotation recovery approaches rely on the assumptions of identical objects in different projections [37, 73], i.e. we have to handle the structure heterogeneity before aligning 3D objects. In this thesis work, we made efforts in handling data complexity and volumes by selecting high-quality and homogeneous single-particle diffraction patterns using machine learning algorithms.
- The particle orientations are unobservable in current FXI experiments, and hence extra steps to compute relative rotations are necessary for 3D alignment. The current preferable method for orientation determination is the Expectation-maximization-compression (EMC) [53, 52] algorithm. This method is computationally very demanding. Firstly, a large number of computations is needed to fit each diffraction pattern into the discretized rotational space. Second, a large number of FXI patterns is fundamental for achieving high resolution and balancing the weak photon signal when studying smaller objects. There is an imminent need for a massively parallel implementation of the 3D reconstruction algorithm to keep up with the

increased data rates. Our proposed distributed computation scheme allows EMC to run on many nodes with a nearly linear speedup.

- To better understand the 3D structure of the sampled particles would require high-resolution 3D reconstructions along with a comprehensive understanding of the uncertainty propagation in the reconstructing procedure. Adopting bootstrap methodology, we may analyze uncertainties in both the Fourier domain and in the real-space. With a proper cutoff, we can determine the resolution of a 3D structure directly from the uncertainty analysis.
- Manual data analysis is no longer an option for FXI experiments, with the high XFEL repetition rate and the growing interests of different samples. We such proposed a (semi-)automatic data analysis pipeline with multiple components including selecting desired patterns, aligning patterns into 3D volumes, converting Fourier intensity into the real domain, analyzing uncertainties, removing unwanted features, etc. The idea is use the pipeline for determining 3D structures directly from the detector readouts during an FXI experiment in the near future.

To summarize, this thesis brings an improved multi-step data-analysis pipeline into FXI experiments. With our pipeline, we can now process raw FXI data to reconstruction 3D models of the sample molecules quickly and robustly, along with a better understand of uncertainties in the models.

Part II:
Concepts and Technology

2. Diffractive Imaging with XFELs

A photon interacts with an atom in different ways, for example, elastic scattering, photon absorption, and Compton scattering, etc. Among them, elastic scattering is the most "useful" form, in that it gives scattering signals containing structural information. Unavoidably and unfortunately, a significant fraction of X-rays is not elastically scattered and will lead to structural damage [15, 95], and the scattering signals are also degraded, which may lead to poor resolution and artifacts in the structure.

The concept of "diffraction and destruction", which was brought up in 2000 [66], and experimentally demonstrated in 2006 [14], makes use of the fact that the time scale of X-ray diffraction is much shorter than the radiation-induced sample degradation. Today, femtosecond X-ray pulses produced from modern XFELs achieve the power density of more than 10^{16} W/cm^2 per pulse with a micron-sized focus. And hence, modern XFELs allow capturing interpretable diffraction signals from single particles before a significant structural change occurs. The technology that captures 2D diffraction images from single particles using XFEL pulses is typically named Flash X-ray single-particle diffraction Imaging (FXI) or X-ray single particle imaging (SPI).

Thanks to the high repetition rate of XFELs, we obtain enormous data from FXI experiments. For example, SPring-8 Angstrom Compact Free Electron Laser (SACLA) operates at 60 Hz [86], Linac Coherent Light Source (LCLS) at 120 Hz [12], and the European XFEL [2], is capable of acquiring up to 27,000 diffraction patterns per second. This incredible rate leads to sharply increasing data volumes, and allows study single molecules either in 2D or in 3D by aligning multiple diffraction patterns into their relative rotations.

In this chapter, we introduce the key terminologies of XFEL in §2.1 and the theory behind diffractive imaging in §2.2. We also briefly describe the methods to solve the phasing problems and retrieve the rotations of the diffraction patterns in §2.3 and §2.4, respectively.

2.1 XFELs

The XFELs are X-ray light sources that allow researchers to study the structure of matter at the atomic level. All XFELs are large facilities

equipped with a long accelerator and a huge undulator. Typically, as shown in Figure 2.1, electrons are accelerated to nearly the speed of light by an accelerator. Then those electrons bunches pass an undulator, which is a periodic array of magnets with alternating poles, to be enhanced and shortened until an extremely intense X-ray flash is finally created.

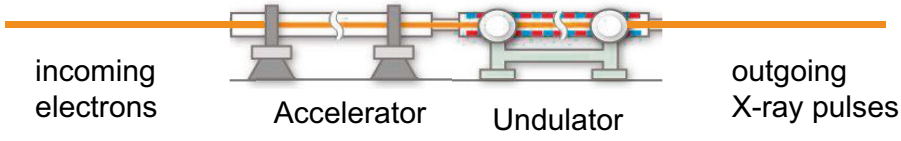


Figure 2.1. An illustration of XFEL components. Modern XFEL facilities use either linear or circular accelerator, followed by a periodic array of magnets (undulator). The accelerated and coherent electrons generated by the undulator hit a target generating the X-ray pulses with the desired wavelength and intensity.

2.1.1 Accelerator

An accelerator typically has three major components. To start with, it needs an electron gun to produce bunches of electrons. Electrons can be generated by a cold cathode, a hot cathode, a photocathode, or radio frequency (RF) ion sources, etc. Then, electrons travel through a vacuum chamber, accelerating by the electromagnetic field.

Modern XFEL facilities use either a linear or a circular accelerator. The design of different types of accelerators give different properties and advantages. In a linear accelerator, electrons travel down a long, straight track and the electromagnet keeps the particles confined in a narrow beam. Typically linear accelerators are huge and are kept underground. An example of a linear accelerator is Linac at the Stanford Linear Accelerator Center (SLAC) in California, which is about 1.8 miles (3 km) long.

Circular accelerators do mainly the same jobs as linear accelerators. However, instead of using a long straight track, they move the electrons around a circular track many times. The advantage of circular accelerators is that the ring topology allows continuous acceleration, and they are therefore typically smaller than a linear accelerator of comparable power. However, electrons accelerated radially may emit synchrotron radiation leading to undesired energy loss. For this reason, many XFELs use linear accelerators.

2.1.2 Undulator Radiation

Although an accelerator is powerful enough to speed up electrons to nearly light speed, the problem is that the lengths of electrons bunches from an accelerator are much longer than the desired wavelength of an XFEL pulse. To shorten the wavelengths of electrons bunches, XFEL uses a periodic array of magnets with alternating poles across the beam path to wiggle and concentrate the electrons in the magnetic field generated by the XFEL undulator.

The XFEL undulator is a periodic array of magnets with alternating poles across the beam, see Figure 2.2. Due to the Lorentz force of the magnetic field, the undulator forces electrons in the beam to wiggle transversely, traveling along a sinusoidal path about the axis of the undulator. The transverse acceleration of electrons will release monochromatic but incoherent photons, and the power of the radiation scales linearly with the number of the electrons. The wavelength of the undulator radiation λ is a function of the undulator period λ_u,

$$\lambda(\theta) = \frac{\lambda_u}{2\gamma}\left(1 + \frac{1}{2}K^2 + (\gamma\theta)^2\right), \tag{2.1}$$

where θ is the divergence angle, and γ is the relativistic Lorentz factor,

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2}. \tag{2.2}$$

Further, K is the undulator strength,

$$K = \frac{\lambda_u e B_0}{2\pi m_e c} \approx 0.9337 B_0 \lambda_u, \tag{2.3}$$

where B_0 is the magnetic field strength of the undulator, e is the electron charge, m_e is the electron mass, c is the speed of light, and v is the speed of the electrons.

Eq. (2.1) is often referred to as the undulator equation. By manipulating the parameters: the speed of electrons v , the magnetic filed B_0, the divergence angle θ, and the unduator period λ_u, XFELs may produce X-ray pulses with the desired wavelength,

$$\lambda_r = \frac{\lambda_u}{2\gamma}\left(1 + \frac{1}{2}K^2\right). \tag{2.4}$$

2.1.3 Microbunching and Self Amplified Stimulated Emission (SASE)

Self Amplified Stimulated Emission (SASE) [63] allows the pulse energy to grow exponentially in the XFEL undulators. Initially, electrons enter

an undulator with random phases, and hence the emitted radiation is incoherent as illustrating in Figure 2.2 *(left)*. In the magnetic fields, electrons interact with the emitted radiation, forcing faster electrons to lose energy and slower ones to gain energy. The result is a modulation of the longitudinal velocity, which eventually leads to a concentration of the electrons in slices. The electron slices are called microbunches and located close to the positions where maximum energy transfer to the light wave can happen. The microbunching process leads to an exponential growth of the radiation power along the undulator, and the exponential growth stops when the electrons are strongly bunched, and begin to debunch, showing in Figure 2.2 *(right)*. Moreover, the SASE process is stochastic, and hence, the emitted radiations have a shot-to-shot intensity fluctuation.

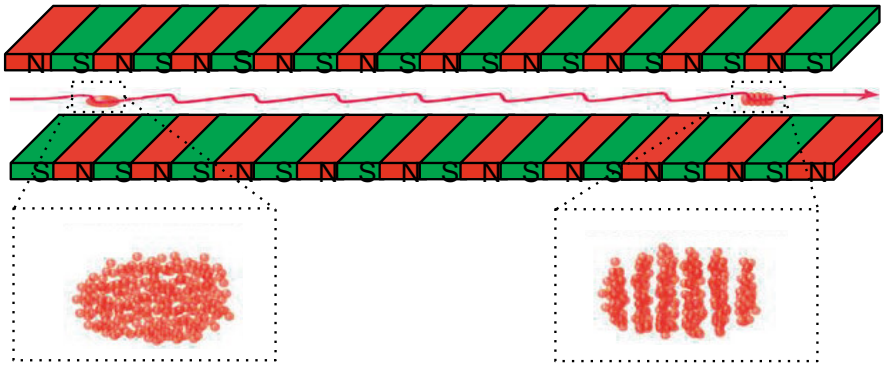


Figure 2.2. An illustration of the XFEL undulator. Initially, electrons are in random phases, and the radiation is incoherent. At the end of the undulator, electrons are microbunched and the radiation is coherent.

2.2 Diffractive Imaging

2.2.1 2D Diffraction Pattern Acquisition

For a typical FXI experiment, the diffraction data is acquired according to the scheme in Figure 2.3. As in the illustration, a stream of inflow samples of biomolecules is injected into the X-ray beam by injectors, such as gas injector [10], liquid injector [90], lipidic cubic phase (LCP) injector [91], etc. X-ray pulses generated from XFELs interact with sample biomolecules at random, and the captured signals on the detector can be diffraction signals from single biomolecules or a cluster of biomolecules, or from a background, or even from contaminants. To proceed further data analysis, we need fast and accurate algorithms to

classify and select high-quality diffraction patterns from the raw frames (see more details in ***Paper I*** and in §4). Further, we lose the following information during the process: the phases of the diffraction signals, the orientations of samples, and the X-ray pulses instantaneous intensities at the interactions. Moreover, we may have missing pixel values due to the physical arrangement of the detector, saturation, and faulty pixels. The direct beam — the unscattered wave passes through the hole at the center of the detector and is then collected by a beam stop. Sometimes pixels around the hole are saturated and may hence be considered as missing information. The algorithms for retrieving missing phases and orientations are discussed in §2.3 and §2.4, respectively.

2.2.2 Data Recording

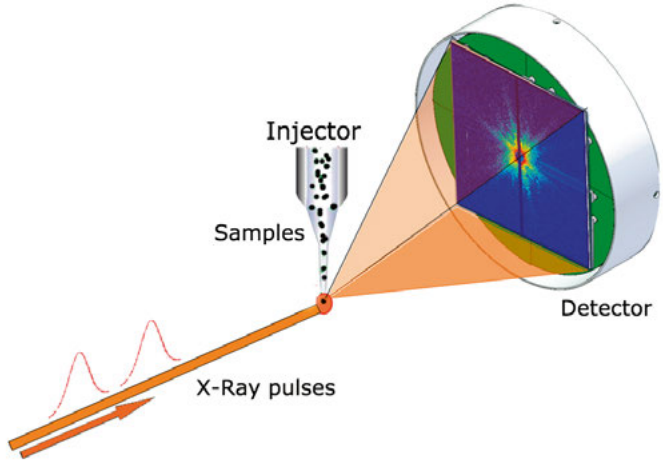
As shown in Figure 2.3, we typically use 2D digital detectors containing massive pixels, such pnCCD cameras. In a general FXI setup, we may include two pairs of detectors, namely a front and a back detector. The back detector, the one illustrated in Figure 2.3, captures the signals from the lower scattering angles, and allows the direct beam to pass through. Comparing with the back detector, the front detector is much closer to the interaction region, and opened to allow the passing of the scattered waves to the back detectors. The data used in this thesis were from the back detector, which consists of two moveable halves, with an empty strip in between, and a central hole for the unscattered signal (the direct beam).

Other than the detector geometry, the saturation of the detector also leads to missing information. Pixels of digital detectors can only hold a certain amount of electrical charge. Charges can overflow to the neighbouring pixels if the incoming changes exceed the maximal value, a phenomenon called saturation. Saturated pixels are usually located in the center of the detector, as the scattered signals are much stronger there.

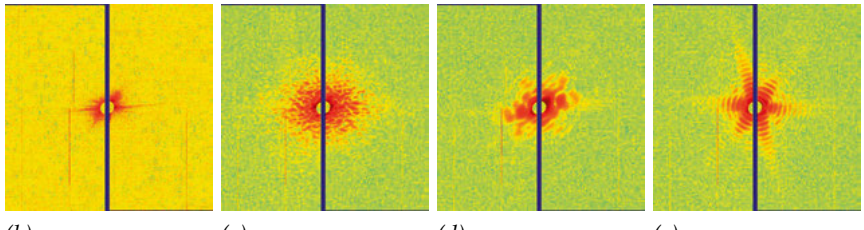
Given that the small angle approximation holds for FXI experiments, we can approximate the pixel size in real space (scattering potential space) d_r as follows:

$$d_r = \frac{\lambda d}{d_p}, \tag{2.5}$$

where d is the object-detector distance, and λ is the wavelength. Further, d_p is the physical Euclidean distance of a pixel to the detector center.



(a)



(b) *(c)* *(d)* *(e)*

Figure 2.3. An illustration of a typical FXI experiment. A flow of incoming samples interacts with X-ray pulses, and the detector captures the intensities of the scattered waves. At the center of the detector, there is a hole for the unscattered waves. This figure is adapted from ***Paper III***. The readouts from detectors ([(b)–(e)]) are scatterings to varying degrees. (b) was a blank frame which contains only background scattering. Due to the low hit rate, most readouts from the detector are blank frames. [(c) and (d)] were frames from multiple particles or with contaminants. (e) was a single-particle frame from an icosahedron virus with a relatively strong fluence. Currently, single particle patterns are the most interesting patterns, which can be used to assemble 3D structures.

2.2.3 Prediction of 2D Diffraction Patterns

According to classical diffraction theory [71], we may consider a sample particle as a collection of infinitesimal point scatterers for coherent diffractive imaging (CDI). Let an XFEL pulse (a coherent-plane wave $\mathbf{k_0}$) illuminate the samples at position $\mathbf{x}$ and the scattered wave propagate along the wave vector $\mathbf{k_1}$. The detector then captures the scattered wave at position $\mathbf{x}'$. Vectors $\mathbf{k_0}$ and $\mathbf{k_1}$ have the same length since we only consider elastic scattering. The scattering vector $\mathbf{q} = \mathbf{k_1} - \mathbf{k_0}$ lies on the Ewald sphere in the diffraction / reciprocal / Fourier space. Figure 2.4 illustrates the simplified geometry of FXI (plane-wave CDI) experiments in both Fourier and real space.

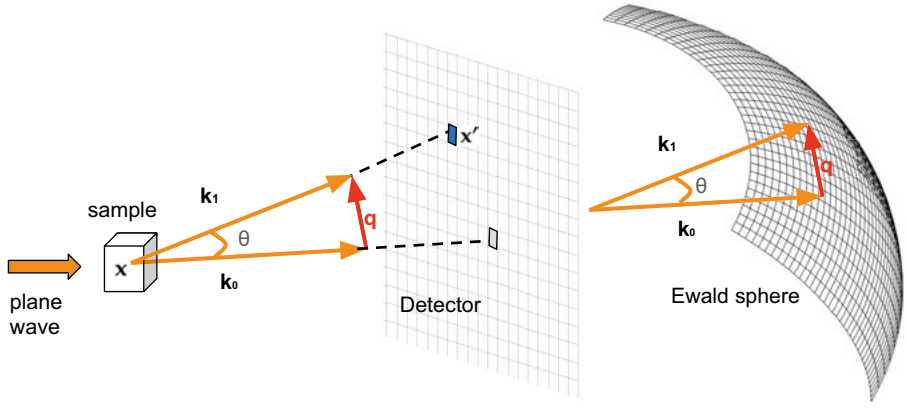


Figure 2.4. Geometry for plane-wave CDI in the real space *(left)* and the Fourier space *(right)*. The wave vectors $\mathbf{k_1}$ and $\mathbf{k_0}$ have the same length, and the scattering vector $\mathbf{q}$ lies on the Ewald sphere in the Fourier space.

To simulate FXI experiments with single particles with relative small sizes, we only consider the single scattering events. By applying the first order Born approximation (the single-scattering approximation) to the Maxwell's equations, we may write the scattered wave $\boldsymbol{\Psi}(\mathbf{x}')$ as the sum of the incoming plane wave $\boldsymbol{\Psi}^{(0)}(\mathbf{x}')$ and the scattering wave $\boldsymbol{\Psi}^{(1)}(\mathbf{x}')$:

$$\begin{aligned}\boldsymbol{\Psi}(\mathbf{x}') &= \boldsymbol{\Psi}^{(0)}(\mathbf{x}') + \boldsymbol{\Psi}^{(1)}(\mathbf{x}') \\ &= \Psi_0 \exp(\mathrm{i}\mathbf{k_0}\mathbf{x}') + \Psi_0 \iiint \psi(\mathbf{x}) \exp(\mathrm{i}\mathbf{k_0}\mathbf{x}) \frac{\exp(-\mathrm{i}k|\mathbf{x}' - \mathbf{x}|)}{|\mathbf{x}' - \mathbf{x}|} \mathrm{d}\mathbf{x}, \quad (2.6)\end{aligned}$$

where $\mathbf{x}$ is the position of scatterers inside a sample, $\mathbf{x}'$ is the projected position of the scattered wave $\mathbf{k_1}$ on the detector as shown in Figure 2.4, and k is the wave number. Further, Ψ_0 is the amplitude of the incoming

wave $\mathbf{\Psi}^0$, and the scattering potential $\psi(\mathbf{x})$ is

$$\psi(\mathbf{x}) = \frac{k^2}{4\pi}[1 - n^2(\mathbf{x})], \tag{2.7}$$

with $n(\mathbf{x})$ is the refractive index.

Since the detector captures the scattered wave at the far field, i.e. the propagation distance $r = |\mathbf{x}' - \mathbf{x}|$ is much larger than the size of the sample, we may simplify Eq. (2.6) as follows:

$$\begin{aligned}\mathbf{\Psi}^{(1)}(\mathbf{q}) &= \Psi_0 r^{-1} \iiint \psi(\mathbf{x}) \exp(-\mathrm{i}\mathbf{q}\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \Psi_0 r^{-1} (2\pi)^{3/2} \mathscr{F}[\psi(\mathbf{x})](\mathbf{q}),\end{aligned} \tag{2.8}$$

with $\mathbf{q} = \mathbf{k_1} - \mathbf{k_0}$ the scattering vector as shown in Figure 2.4. Further, $\mathscr{F}$ is the continuous Fourier transformation (FT). For any well-behaved function $h(\mathbf{x})$ in l Euclidean dimensions,

$$\mathscr{F}[h(\mathbf{x})](\mathbf{q}) = \tilde{h}(\mathbf{q}) = (2\pi)^{-l/2} \int_{\mathbb{R}^l} h(\mathbf{x}) \exp(-\mathrm{i}\mathbf{q}\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{2.9}$$

For FXI experiments, the diffraction pattern does not contain information for the full 3D structure of the sample, but has the structural information from the Ewald sphere, which is orthogonal to the plane wave vector $\mathbf{k_0}$. Without loss of generality we assume that the diffraction wave propagates along the z axis, and the 2D scattered wave of Eq. (2.8) is now

$$\begin{aligned}\mathbf{\Psi}_{\perp}^{(1)}(q_x, q_y) &= \Psi_0 r^{-1} \iint \psi_{\perp}(x, y) \exp(-\mathrm{i}(q_x x + q_y y)) \mathrm{d}x \mathrm{d}y, \\ &= \Psi_0 r^{-1} (2\pi)^{1/2} \mathscr{F}[\mathbf{\Psi}^{(1)}(x, y)](q_x, q_y)\end{aligned} \tag{2.10}$$

with

$$\psi_{\perp}(x, y) = \int \psi(x, y, z) \exp(-\mathrm{i}q_z z) \mathrm{d}z. \tag{2.11}$$

Moreover, the detector can only measure the intensity of the scattered wave, and hence we write the expectation value of scattered photons measured in pixels (without noise and assuming no signal loss) as follows:

$$I = |\mathbf{\Psi}_{\perp}^{(1)}(q_x, q_y)|^2 P(\Theta) \Omega(\Theta), \tag{2.12}$$

with $\Omega(\Theta)$ is the solid angle covered by the detector pixels, and $P(\Theta)$ is the polarization factor of the incoming beam. For XFELs, the undulator radiation is linearly polarized, therefore $P(\Theta) \approx 1$. A detailed derivation of the scattered wave from single particles can be found in [38].

2.3 Phase Retrieval

The captured diffraction signals are intensities of the scattered wave, hence excluding the phase information of the scatterings. To retrieve structures from diffraction intensities, we need to robustly recover the phase information.

2.3.1 Phase Retrieval Algorithms

Back in 1972, Gerchberg and Saxton [33] proposed an iterative algorithm for retrieving phases from scattering intensities, and it became the foundation of today's iterative phase retrieval algorithms. Their strategy is to apply the FT and its inverse back and forth in iterations between the real-domain and Fourier-domain constraints. In 1978, Fienup [30] applied Gerchber-Saxton strategy under the condition of CDI imaging, i.e., the amplitudes in the Fourier domain and the support in the real domain are known, and he named his algorithm error reduction algorithm (ER) (see Algorithm 1). In the ER algorithm,the real-space is split into two sub-regions: within a defined boundary and outside of the boundary. The sub-region within the boundary is allowed to have a density and the other parts are forced to be empty. The sub-region allowed to have a density is usually called the object's support. The Fourier domain constraint is that the Fourier transform of the object should match the square root of the measured intensities. The ER algorithm treats the phase retrieving problem as a convex optimization problem, however the set of Fourier-domain constraints is not convex, so ER might be easily trapped by local minima, see graphical illustration of ER algorithm in Figure 2.5.

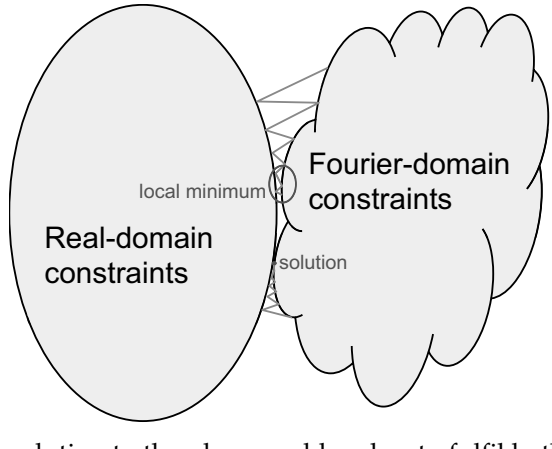


Figure 2.5. The solution to the phase problem has to fulfil both the real-space constraint and the Fourier-space constraint.

Algorithm 1: Error reduction (ER) algorithm.

Input: A diffraction pattern and a support.
Output: The recovered object density and phases.

1: Assign a random phase to every pixel of the amplitude of diffraction pattern.
2: **while** the recovered phase is not accurate enough **do**
3: Inverse Fourier transform the pattern.
4: Set all pixels outside the support to zero in the real domain (apply the real-domain constraint).
5: Fourier transform the object in the real domain.
6: Replace the amplitudes with the experimentally measured amplitudes, but keep the phases (apply the Fourier-domain constraint).
7: **end while**

To escape from local minima, Fienup [30] developed the hybrid input-output algorithm (HIO). The workflow of HIO is identical to ER, except that the pixels outside the support are no longer empty, but implemented via a negative feedback term. Another popular modification of ER is the Relaxed Averaged Alternating Reflections algorithm (RAAR) [54], which has a negative feedback term similar to HIO and slightly modifies the update rule of pixels inside the support. RAAR behaves intermediately between ER and HIO, and escape only shallower local minima. Other iterative phase retrieval algorithms using a support include the difference map [27], Saddlepoint optimization [58], and Hybrid projection reflection[6].

HIO and RAAR requires the support to follow the shape of the actual object tightly, and in practice, the object shape is most-likely unavailable. In 2003 Marchesini developed an algorithm called Shrinkwrap [59] to deduce the shape of the support during the phase retrieve iterations. It update the supports by applying a Gaussian blur to the real space image and selecting the pixels that have a value above a certain threshold periodically.

2.3.2 Validation

Due to the concave constraints in the Fourier domain, we need to identify failed phase searches, which typically have high errors in both

the Fourier domain E_f and the real domain E_r:

$$E_f = \sqrt{\frac{\sum_{i=1}^{M_{\text{pix}}} (|\tilde{h}_i| - \sqrt{I_i})^2}{\sum_{i=1}^{M_{\text{pix}}} I_i}}, \tag{2.13}$$

and

$$E_r = \sqrt{\frac{\sum_{i \in \bar{S}} (|h_i|)^2}{\sum_{\in \bar{S} \cup S} |h_i|^2}}, \tag{2.14}$$

where S is the object's support, $\bar{S}$ is the area outside of S. Further, h is the recovered pixel intensity of the object in the real space, while $\tilde{h}$ is the recovered wave in pixels in the Fourier space, M_{pix} is the number of pixels, and I_i the ith pixel value of a diffraction pattern. Note that a recovered intensity in real space is quite often called reconstruction.

As described in Algorithm 1 *(step 6)*, we are seeking for the best match between the amplitudes of diffraction patterns and the recovered signals (the Fourier-domain constraints) in the ER algorithm, i.e. Eq. (2.13) should be small. Further, the real-domain error E_r Eq. (2.14) integrates the intensity outside the object's support S, and a high E_r normally indicates an incorrect support.

Another issue of the iterative phase retrieval algorithms is that the residuals of the reconstruction fluctuate, and hence the reconstruction which fits the constraints the best may not be the desired one. To find the most representative reconstruction, we run phase retrieval algorithms from different sets of starting phases, and then average among those successful reconstructions (i.e., with low E_f and E_r). We also measure the reliability of the average reconstruction by the phase retrieval function (PRTF) [14]:

$$\text{PRTF}_i = N^{-1} \sum_{n=1}^{N} \frac{\tilde{h}_i^{(j)}}{|\tilde{h}_i^{(j)}|}, \tag{2.15}$$

where N is number of successful reconstructions, i is again the index of detector pixels, and $|\tilde{h}_i^{(j)}|$ is the same for all reconstructions since they are the pixel intensities of the diffraction pattern, i.e. $|\tilde{h}_i^{(j)}| = \sqrt{I_i}$. If phases from all reconstructions match well, PRTF = 1, and if phases are entirely random, PRTF = $N^{-1/2}$. We also use e^{-1} or 0.5 cut-off [14] in the radial average of PRTF to determine resolution of the retrieved object.

Fourier shell correlation (FSC) [89] is a compensation method to PRTF, which guards against over-fitting to noise. Given a sufficiently

over-sampled diffraction pattern I, we randomly split it into two sets of equal size and then downsample them. The downsampled sets ($I^{(A)}$ and $I^{(B)}$) are different due to the noises in the diffraction pattern I. FSC measures the normalized cross-correlation coefficient over corresponding shells r_k between recovered Fourier-domain images $\tilde{h}^{(A)}$ and $\tilde{h}^{(B)}$. In other words, FSC is a function of the spatial frequency r,

$$\mathrm{FSC}(r_k) = \frac{\sum_{i \in r_k} \tilde{h}_i^{(A)} \tilde{h}_i^{(B)*}}{\sqrt{\sum_{i \in r_k} |\tilde{h}_i^{(A)}|^2 | \sum_{i \in r_k} |\tilde{h}_i^{(B)}|^2}}, \tag{2.16}$$

where $\tilde{h}_i^{(B)*}$ denotes complex conjugation of $\tilde{h}_i^{(B)}$, and FSC ≥ 0.5 indicates overfitting to noise [89].

2.3.3 Additional Constraints

Until now, we have only used the size of the object as an external input for the phase retrieval methods. Two other common extra constraints are the *reality constraint* and the *positivity constraint*.

In the real space, we often assume that the real part of the scattering factor is much larger than the imaginary part. By applying the *reality constraint* in Eq. (2.17), we force the phase of the object to be either 0 or π,

$$\mathrm{Im}(h_x) = 0. \tag{2.17}$$

Typically FXI samples are small enough to keep absorbance and the maximum phase shift sufficiently small, hence the negative scattering factors do not exist and we can restrict the phase of the object between $[0, \pi/2]$ by applying the *positivity constraint*:

$$\mathrm{Im}(h_x) \geq 0, \;\; \mathrm{Re}(h_x) \geq 0. \tag{2.18}$$

Both constraints require an accurate center of the diffraction pattern, since *reality constraint* implies Friedel symmetry and the *positivity constraint* conflicts with large phase ramps in the object domain.

2.4 From 2D diffraction patterns to 3D Volumes

The underlying idea of FXI is "diffract and destroy", which means that all particles will be ruined shortly after being hit by an XFEL pulse, and hence FXI cannot image one particle multiple times[1]. Luckily,

[1] Although it is possible to illuminate a sample from multiple directions at once, the diffraction data is still not enough.

many bioparticles have identical copies, i.e. they are structurally reproducible. We can treat diffraction patterns from those particles as if they came from the same particle, and therefore we can assemble those 2D diffraction patterns into 3D intensities assuming the particle rotations can be recovered.

In this section, we briefly summarize the different algorithms to recover relative rotations of 2D diffraction patterns. The 3D alignment can be done in Fourier space before phase retrieval (see §2.4.1), and the current state-of-the-art algorithm is the Exception-Maximization-Compression (EMC) algorithm. An alternative way is to combine iterative phase retrieval algorithms with orientation determination and directly obtain the 3D intensity in real space, see §2.4.2.

2.4.1 Aligning in Fourier space

Maximum Likelihood Imaging

Given a probability models **P** of the expected diffraction intensities, we can improve the 3D Fourier intensity of an object from a sufficiently large number of diffraction patterns using a Maximum-Likelihood (ML) estimator. With i.i.d. frames $K = (K_k)_{k=1}^{M_{\text{data}}}$, the Maximum Likelihood estimator is given by

$$\hat{W} = \arg_W \max\, M_{\text{data}}^{-1} \sum_{k=1}^{M_{\text{data}}} \log \mathbf{P}(K_k|W). \tag{2.19}$$

This optimization is incomplete for FXI experiments for two reasons: firstly, the true *rotation* R_k of the diffraction pattern K_k is unknown and consequently the frame cannot be directly associated with a definite Ewald sphere W; secondly, the *photon fluence* ϕ_k, the X-ray intensity at the time and location when hitting the sample particle, is also unknown. Hence, we need to redefine the optimization problem in Eq. (2.19) with a marginal probability,

$$\hat{W} = \arg\max_W M_{\text{data}}^{-1} \sum_{k=1}^{M_{\text{data}}} \sum_{j=1}^{M_{\text{rot}}} \log \mathbf{P}(K_k|W, R, \phi). \tag{2.20}$$

The Expectation Maximization (EM) algorithm finds the ML estimator of the marginal likelihood Eq. (2.20) by iteratively applying the Expectation step (*E step*) and the Maximization step (*M step*).

In the E step, we calculate the expectation of the log likelihood function with respect to the conditional distribution of the rotation R, given the diffraction pattern K and the current estimation of $W^{(n)}$ and $\phi^{(n)}$ at

iteration n,

$$Q(W,\phi|W^{(n)},\phi^{(n)})^{(n+1)} = \mathrm{E}_{R|K,W^{(n)},\phi^{(n)}} \log \mathbf{P}(K_k|W^{(n)},\phi^{(n)},R), \tag{2.21}$$

and the rotational probability is explicitly available by taking the exponential of $Q^{(n+1)}$.

The M step freezes the Q at the $(n+1)$th iteration, so that ϕ and W may be obtained as a solution of the following optimization problem:

$$[\phi^{(n+1)}, W^{(n+1)}] = \arg\max_{\phi,W} Q(\phi, W|\phi^{(n)}, W^{(n)}) \tag{2.22}$$

Further, we denote by $(R_j)_{j=1}^{M_{\mathrm{rot}}}$ the sample rotations of the rotational space. Since sampled rotations are generally non-uniformly distributed, we denote by w_j the prior weight for the jth rotation, and normalize all sample rotations such that $\sum_j w_j = 1$. In other words, selecting R_j with probability w_j implies a practically uniform sampling of the rotational space. A suggestion for sampling rotational space uses the quaternions encoded rotations and a suitable geometric object for this purpose is the 600-cell (or *hexacosichoron*) [53, Appendix C]. With this suggestion, we can calculate the number of sampled rotations by:

$$\begin{aligned} M_{\mathrm{rot}}(d) &= 10 \cdot (5d^3 + d) \\ &= [6\,300, 10\,860, 25\,680, 50\,100, 86\,520] \quad \text{for } d = [5,6,8,10,12,\cdots], \end{aligned} \tag{2.23}$$

with d a free integer parameter, and we typically use $d = 10$ or $d = 12$.

Since detectors are pixelized, we can write the kth diffraction pattern as $(K_{ij})_{i=1}^{M_{\mathrm{pix}}}$, where M_{pix} is the number of pixels on the detector. With a set of points q_i defined on an Ewald sphere, we write the 2D sampled Fourier intensity at position $R_j q_i$ as W_{ij}. Moreover, we denote by ϕ_{jk} the estimated *photon fluence* of the diffraction pattern K_k, given that the object was rotated according to R_j. With these, we may rewrite the log likelihood function as follows:

$$Q_{ijk} = \log \mathbf{P}(K_{ik}|W_{ij}, R_j, \phi_{jk}), \tag{2.24}$$

The joint log likelihood function is therefore

$$Q_{jk} = \sum_i Q_{ijk} = \sum_i \log \mathbf{P}(K_{ik}|W_{ij}, R_j, \phi_{jk}), \tag{2.25}$$

and this is also the solution of the E-step in equation Eq. (2.21). The normalized rotational probability is now obvious,

$$\begin{aligned} P_{jk}^{n+1} &= P_{jk}^{n+1}(W^n, \phi^n) =: \mathbf{P}(R_j|K_k, \phi^n, W^n) \\ &= \frac{w_j \exp(Q_{jk}(W^n))}{\sum_{j'=1}^{M_{\mathrm{rot}}} w_{j'} \exp(Q_{j'k}(W^n))}. \end{aligned} \tag{2.26}$$

We now briefly introduce some different probability models. Since the photon counting process is Poisson [32, 53], it is natural to assume that the ith pixel of the kth measured diffraction pattern K_{ik} is Poissonian around the unknown Fourier intensity W_{ij}, i.e.,

$$\mathbf{P}(K_{ik} = \kappa | W_{ij}, R_j) = \prod_{j=1}^{M_{\text{rot}}} \frac{(W_{ij})^{\kappa} \exp(-W_{ij})}{\kappa!}. \tag{2.27}$$

In this Poisson modele, we need to normalize the diffraction patterns before applying this model, Since the photon fluence ϕ is not identical for every diffraction pattern in practice.

In stead of normalizing images, some attempts [52, 26] have been made to better estimate the photon fluence ϕ by approximating the Poisson distribution by a Gaussian distribution for high-intensity FXI diffraction patterns, i.e.,

$$\mathbf{P}(K_{ik} = \kappa | W_{ij}, R_j, \phi_{jk}) \approx \prod_{j=1}^{M_{\text{rot}}} \exp\left(\frac{(\kappa/\phi_{jk} - W_{ij})^2}{2\delta^2} \right), \tag{2.28}$$

with δ a free noise parameter, and is difficult to define in practice.

We proposed the scaled Poissonian model in ***Paper III***. More precisely, we assume that K_{ik} is Poissonian around the scaled unknown Fourier intensity $\phi_{jk} W_{ij}$, i.e.,

$$\mathbf{P}(K_{ik} = \kappa | W_{ij}, R_j) = \prod_{j=1}^{M_{\text{rot}}} \frac{(\phi_{jk} W_{ij})^{\kappa} \exp(-\phi_{jk} W_{ij})}{\kappa!}. \tag{2.29}$$

Since the photon fluences ϕ, the values of the slices W and the diffraction pattens K cannot be negative, we may borrow ideas from Non-negative matrix factorization (NNMF) to solve the scaled-Poissonian ML problem, see more details in ***Paper III***.

The EMC algorithm

We have now discussed EM with FXI diffraction patterns K and a set of unknown Fourier intensities W. Since both K and W are 2D images and the wanted Fourier intensity is in 3D, we need to apply extra steps to interpolate data between the 2D and the 3D space. More specifically, we first, in the expansion step(*e step*), expand the 3D volume $\mathbb{W}$ into slices W, and then update W by EM, finally assemble them back into 3D volume, in the Compression step (*c step*), for every iteration, until the algorithm converges. This algorithm is called the

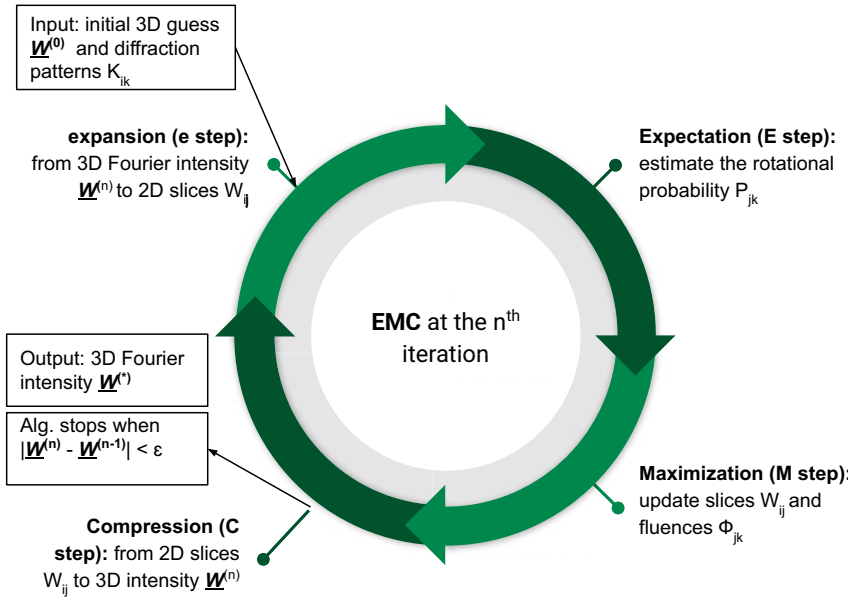


Figure 2.6. A graphical illustration of the EMC algorithm. In this figure, the symbol $\underline{\boldsymbol{W}}$ is the 3D Fourier intensity $\mathbb{W}$. $\underline{\boldsymbol{W}}^{(*)}$ is a converged model from the EMC algorithm. The algorithm contains four steps at each iteration. i) The expansion step (*e step*) interpolate a 3D volume into 2D slices, according to Eq. (2.31). ii) The Expectation step (*E step*) seeks for the rotational probability P via Eq. (2.26). iii) The Maximization step (*M step*) updates the photon fluence ϕ and slices W by solving the optimization problem stated in Eq. (2.22). iv) The Compression step (*C step*) assembles updated slices W_{ij} back into a 3D volume $\mathbb{W}$, according to Eq. (2.31).

expansion-Expectation-Maximization-Compression (EMC) algorithm, see the illustration of the algorithm in Figure 2.6.

The e step interpolates a 3D Fourier intensity into 2D slices and the c step reverses the procedure. Let $\mathbb{W} = \{\mathbb{W}_l\}_{l=1}^{M_{\text{grid}}}$ be a 3D discrete model, an estimation of the 3D Fourier intensity of a biomolecule, where $M_{\text{grid}} = M_{\text{pix}}^{3/2}$. We define interpolation weights f and interpolation abscissas $(p_l)_{l=1}^{M_{\text{grid}}}$ for some smooth function g,

$$g(q) \approx \sum_{l=1}^{M_{\text{grid}}} f(p_l - q) g(p_l). \tag{2.30}$$

An e step slices W_j from the 3D model $\mathbb{W}$

$$W_{ij} = \sum_{l=1}^{M_{\text{grid}}} f(p_l - R_j q_i) \mathbb{W}_l. \tag{2.31}$$

The C step inverses the interpolation of the e step by averaging the 2D slices W_{ij} back into the 3D grid $\mathbb{W}$,

$$\mathbb{W}_l = \frac{\sum_{i=1}^{M_{\text{pix}}} \sum_{j=1}^{M_{\text{rot}}} f(p_l - R_j q_i) W_{ij}}{\sum_{i=1}^{M_{\text{pix}}} \sum_{j=1}^{M_{\text{rot}}} f(p_l - R_j q_i)}. \tag{2.32}$$

Further, we may use the stopping criterion for the EMC algorithm as follows:

$$\sum_{l}^{M_{\text{grid}}} |\mathbb{W}_{l=1}^{(n)} - \mathbb{W}_l^{(n-1)}| \leq \epsilon, \tag{2.33}$$

where ϵ is a small positive number, and in practice we use $\epsilon = 0.001$.

To achieve high resolution and balance the low signal-to-noise ratio, we need a fine grid (a large number of rotations) and a massive number of diffraction patters. This makes the EMC algorithm a compute-intensive and memory-intensive algorithm. Implementation details for distributing/ parallelizing the EMC algorithm are discussed in §5 (or ***Paper II***), and modelling details for the EMC algorithm in §6 (and ***Paper III***).

Common arc

Instead of iteratively improving the quality of the 3D Fourier intensity, the *common arcs* algorithms [11] tries to determine the relative rotations

from cross-section images of the same object directly. Assume that two diffraction patterns come from identical copies of a sample particle, that each diffraction pattern is a sampled Ewald sphere from the same 3D Fourier intensity, and the two Ewald spheres intersect through the center via a shared a curve, see Figure 2.7. In other words, the relative rotation of the two diffraction patterns can be found by finding the shared curve. The method is straightforward and easy to parallel, but it can be quite sensitive the pattern noise due to that we only use pixels on the common arcs to determine the relative rotations.

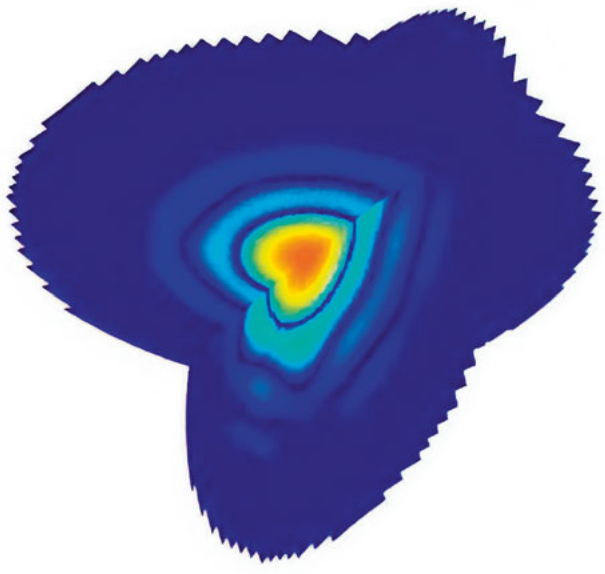


Figure 2.7. A demonstration of the *common arc* algorithm with the two diffraction patterns from the same objects interacting through a common arc. In brief, the common arc algorithm, firstly, finds out the relative orientations of all pattern pairs by choosing the maximum correlations of the intensities along the shared curves. It then fits all relative orientations together to to choose the "absolute" rotation of each pattern.

Manifold embedding

Finding the relative rotations of diffraction patterns can also be considered as a manifold embedding problem in a very high dimensional vector space, as rotations are 3D (or 4D) manifolds and diffraction patterns can be described in a vector space. To find the map between the observed data (diffraction patterns) and the manifolds (rotations), we can use mapping algorithms such as diffusion maps (DM) [16, 34, 80], Self Organizing Map [47], Generative Topographic Mapping (GTM) [7]. It is hard to draw a general conclusion about this group of methods,

but they are more robust than the *common arc* since it makes use of all pixels.

Further, working in high-dimension vector space can be hard and time-consuming, and hence in practise, people seek for dimensional reduction methods for the mapping algorithms. Assuming a Gaussian statistics, [7] used the GTM method with a vector space computed by Factor Analysis (FA). To adjust the parameters of the mapping function, Expectation–Maximization (EM) algorithm was used. With the Wigner D-functions and the eigenfunctions computed via diffusion maps method, [41] also limited the degrees of freedom according to the point groups (symmetries) as well as reduced the dimension of image space. In 2017, [42] illustrated this method on a real FXI dataset, and recovered the structure of the PR772 virus.

Other methods

Similar to the EMC algorithm, the correlation maximization method [87] begins with a random 3D intensity and a new intensity is constructed in each step from all diffraction patterns rotated to their best-fitting orientation. Instead of working on a Cartesian grid, the correlation maximization method first transfers all 2D patterns into a 3D polar grid. It then samples the 3D intensity into polar sections in different orientations and computes the correlations among the polar sections and diffraction patterns. To form a new 3D intensity, all patterns are orientated to their best-fit orientations, which are determined by the maximal value of the correlations.

The angular correlation [77] method also works on a polar grid, it adapts Icosahedral Harmonics and calculates the average angular correlations among the different frequency bins of diffraction patterns. The rotational information of diffraction patterns can therefore be found in the spherical harmonic expansion coefficients.

2.4.2 Real space reconstruction

Other than aligning 2D diffraction patterns to 3D volumes in Fourier space and then phasing the 3D Fourier intensity, one could phase 2D diffraction patterns while aligning them into a 3D real-space volume. The most successful algorithm in this category is the multi-tiered iterative phasing (M-TIP) [22]. Briefly, the M-TIP method [22] is an extension of standard iterative phasing algorithms, and can recover the 3D internal intensity directly from fluctuation X-ray scattering data. Instead of working on a Cartesian grid , the M-TIP considers 2D and 3D Fourier transformations on a polar grid, and defines several projection operators to enforce constraints and assumptions in a fluctuation scattering

experiment. By combining the angular correlation to the iterative phasing algorithm, such as ER, RAAR, HIO with projection operators, the method merges 2D diffraction patterns into 3D intensities in real space.

[48] demonstrated the M-TIP method with reconstructions to 3D intensity of rice dwarf virus (RDV) and PR772 viruses. Similar to the results in [42, 76], the obtained intensities deviated from an ideal icosahedron and have non-uniform distribution of internal structures. However, the resolutions were slightly worse than the detector-edge resolution. Another use of M-TIP method is the 3D structure of the Paramecium bursaria Chlorella virus [72], which also had icosahedral capsid with asymmetrical interior.

To summarize, the rapid developments of the modern XFELs provide possibilities to study the structures of non-periodic bio-samples. Relied on XFELs, the FXI experiments will capture 2D diffraction patterns of single particles before X-ray pulses ruin the samples into plasma, ideally. However, in practise, we may also get a large amount of empty frames, patterns from multiple particles and contaminants from the detector, etc. Further, the phase information of the scattering wave, the particle orientations, and the beam intensity information at the time and location when hitting the particles are unobservable for FXI. Later in §III, we will summarize methods to handling challenges mentioned above. Briefly, the methods for selecting high-quality homogeneous single-particle diffraction patterns is in ***Paper I*** (§4). The method to speed up computations of rotation determination is in ***Paper II*** (§5). The scaled Possionian model the speedup of rotation determination (Eq. (2.29)) and uncertainties measurement of the reconstruction object are in ***Paper III*** (§6). We also demonstrate the above methods together with phase retrieval method, and pattern healing methods, etc. in in ***Paper IV*** (§7) to recover the 3D structure of sample particle from a real FXI experiment.

3. Accelerated Computing

With the rapidly increasing volumes of data and computations, data scientists are in high need of computing resources to produce high-quality results in time. The need for scalable storage and computational resources is fulfilled at a supercomputing facility, or by using a cluster, or even by the cloud. The computational servers handle computational workloads by a variety of processing elements, and here we list some:

1. CPUs are short for Central Processing Units, which carry out the instructions of a computer program. Modern computers often employ multicore processors, which contain two or more CPUs. All modern general-purpose CPUs can support both instruction-level parallelism and thread Level parallelism.
2. Coprocessors are many-core processors that are used in tandem with CPUs. Some coprocessors, such as Floating-point units, rely on direct control via coprocessor instructions, embedded in the CPU's instruction stream. Others are independent of the CPUs and work asynchronously via a limited instruction set focussed on accelerating specific tasks. Although CPUs absorb the functionality of most popular coprocessors over time, the specialized coprocessors are developed to boost computational power and allow for the continued evolution of processor units.
3. Accelerators are also many-core processors, and the most used accelerator is the Graphics Processing Unit (GPU), which may consist of thousands of cores. For example, NVIDIA GTX680 GPU consists of 1536 cores. A GPU is a specialized electronic circuit designed to manipulate and alter memory blocks rapidly, and hence it can efficiently handle highly parallel structures, such as images and computer graphics.

In this section, I will summarize some popular computational paradigms in §3.1, and parallel and distributed computing schemes commonly used in High-Performance Computing in §3.2.

3.1 Popular Computational Paradigms

The rapid developments of modern society, including industries, research, social media, and personal life, etc, rely much on data-driven technologies. Indeed, the amounts of computations and data storage

are enormous and are still increasing rapidly. Researches and engineers move their works from stationary computers to modern supercomputers, clusters, and clouds to deal with big data and heavy computations. Many platforms are available on the market, such as Hadoop [92], Spark [96], Amazon Web Services (AWS), Google Cloud, IBM Cloud, and Microsoft's Azure [23], etc. Those platforms are one-stop, high-level, and easy-access solutions, which may integrate hardware and software for files/data management, data analytics, etc. If one would like to have full control of parallel/distribution computing paradigms in a private computer cluster, the High-Throughput Computing (HTC) and High-performance Computing (HPC) are the key terminologies.

HTC
HTC [51] concerns the computation power over a longer period, in other words, how many floating-point operations can be obtained from the computing environment per month or per year. The HTC tasks are *loosely-coupled* [83], i.e. the communications among tasks are very limited. Therefore, HTC jobs can be executed on physically distributed resources using grid-enabled technologies.

HPC
HPC concerns *floating point operations per second (FLOPS)*. HPC programs use aggregated high-end computing resources along with parallel or distributed processing techniques to solve both compute- and data-intensive problems. HPC computing typically requires communications and synchronizations among HPC servers that are connected by a fast and efficient network. In an HPC platform, a computational task is broken down to many similar subtasks that can be processed independently and simultaneously on different processors, so that the overall execution time is reduced. For efficiently using the underlying processing units and accelerating applications, both parallel and distributed computing schemes can be used.

Since this work involved the HPC platform in various ways, we hereby briefly introduce parallel, distributed, and mixed HPC computing schemes.

3.2 Parallel Computing in HPC

Parallel computing and distributed computing have a lot in common, and it is hard to draw a clear and extinct border between them. In this chapter, we roughly distinct parallel computing and distributed computing via the usage of memory. In parallel computing, all processors or threads have access to shared memory to exchange information. On

the other hand, processors/threads of a distributed computing system will have their private memory, and can only exchange information via passing messages.

A typical parallel computing application uses a language or a library that supports spawning of multiple threads. All threads run concurrently with the runtime environment allocating threads to different processors, and they can access both the private memory and the shared memory space. Figure 3.1 illustrates an example of memory usage for parallel computing.

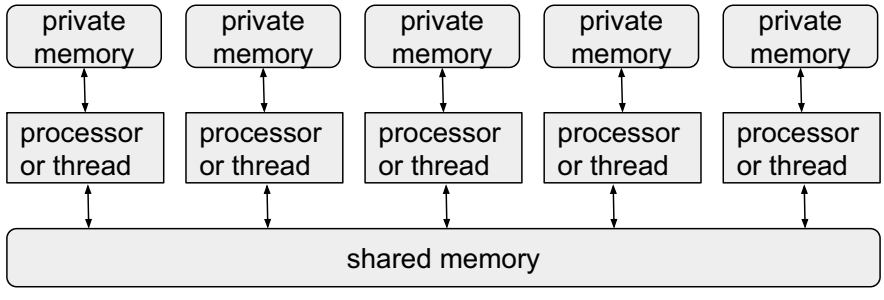


Figure 3.1. An example of memory usage for parallel computing.

POSIX Threads (PThreads) [67] and (Open Multi-Processing) OpenMP [17] are two major CPU implementations of the multithreaded shared-memory parallel paradigm.

PThreads

For a UNIX-like system, PThreads has been specified by the IEEE POSIX 1003.1c Standard. Although the Standard exists independently from a language, PThreads is quite often referred to as Pthreads in C/C++ programming languages. The implemented Pthreads library contains more than 100 PThreads procedures, such as thread management, mutexes, locks, critical section, condition variables, etc. PThread allows one to spawn a new concurrent process flow and allows scheduling of a process flow on a different processor. A typical PThreads program starts with creating threads with specified tasks to each thread and ends up with threads joining.

OpenMP

OpenMP is an explicit programming model, and it offers the programmer full control over parallelization. Similar to PThreads, it also uses the fork-join parallel execution model, which means a master thread creates a specified number of slave threads and the *system* divides the tasks among them. OpenMP has straightforward syntax compared

with Pthreads, for example in C/C++, we parallelize a piece of serial code by adding a command starting with"# pragma omp parallel". OpenMP has full support from many compilers such as GCC, Intel Fortran, and C/C++ compilers.

3.2.1 Parallization on GPUs

Modern GPUs are now widely used in HPC applications since they are very efficient at manipulating highly parallel structures. Usually, a GPU-accelerated application is running on the GPU by offloading some of the compute-intensive portions of the code, such as dense linear algebra and Fast Fourier transforms (FFTs). The rest of the application still runs on the CPU, such as I/O operations. From a user's perspective, the application runs faster because it's using the massively parallel processing power of the GPU to boost performance. Currently, OpenCL and CUDA are the two important and dominating frameworks for writing programs that execute across heterogeneous platforms consisting of CPUs and GPUs.

OpenCL

OpenCL [65] is an open standard framework that views a computing system as a collection of computing devices. Nowadays, it provides support for CPUs, GPUs, and even digital signal processors (DSPs), field-programmable gate arrays (FPGAs) and other processors or hardware accelerators. The OpenCL API is defined in C with a C++ Wrapper, and other languages such as Java or Python also provide third-party bindings. Further, OpenCL is intended to use run-time compilations, which allows OpenCL applications to be portable between implementations for various host devices.

CUDA

CUDA was introduced by NVIDIA, the largest GPU manufactory, in 2006 and it leverages the parallel compute engine in NVIDIA GPUs to solve complex computational problems. CUDA comes with a software environment that allows developers to use C/C++ as a high-level programming language. Cuda also supports other languages, application programming interfaces, or directives-based approaches, such as FORTRAN, DirectCompute, OpenACC.

CUDA also uses the shared-memory paradigm. It can arrange its threads in 1D, 2D, and 3D in threads blocks, and group threads blocks into 1D, 2D, and 3D grids. Every CUDA thread can execute CUDA functions (kernels) in parallel with access to three levels of memory: the per-thread local memory, the per-block shared memory, and the global

memory. As illustrated in 3.2, the CUDA programming model assumes that all CUDA threads execute on a physically separated device (GPU), that operates as a coprocessor to the host (CPU) running the C program. In the illustration, the serial code runs as an ordinary C function on CPU, and the parallel kernel (Kernel0) runs on a GPU grid with 2 by 3 thread blocks, and the next serial code runs again on the CPU. This process continues, hence we can execute multiple parallel kernels and several blocks of CPU code in one program. In this heterogeneous system architecture, the memory management unit of the CPU and the input/output memory management unit of the GPU have to share certain characteristics, like a common address space.

Similar to CPU memories, CUDA allows access to GPU memories at different levels, illustrated in Figure 3.3. Each thread has its registers and a local memory. All threads within a block can access the share memory. The global memory, the texture memory, and the constant memory are accessed by all threads and can exchange data with the host. In this memory hierarchy, registers are the fastest, followed by share memories and local memories. Threads in different blocks communicate each other via global memory.

Further, CUDA provides many useful libraries. CUDA BLAS Library (cuBLAS) implements the standard BLAS specification that is 6x to 17x faster than the latest MKL BLAS [68]. CUDA Sparse (cuSPARSE) [69] provides a collection of basic linear algebra subroutines used for sparse matrices which are 8x faster than Boost. CUDA Fast Fourier Transform Library (cuFFT) is a library that provides a simple interface for computing FFTs. Other libraries can be used for neural network applications (such as cuDNN and TensorRT), image processing (such as NPP and FFmpeg), EM Photonics (CULA Tools), and sequence analysis (NVBIO), etc.

3.3 Distributed Computing in HPC

Unlike the parallel computing, a processor used in distributed computing can only access its private memory, and exchange information with other processors by message passing via communication links. Figure 3.4 illustrates memory usage in a distributed computing system.

Historically, the Parallel Virtual Machine (PVM) and Message Passing Interface (MPI) are two typical approaches for communicating between cluster nodes. PVM is a particular set of libraries, while MPI is a specification with several concrete implementations. MPI provides essential virtual topology, synchronization, and communication functionality between a set of processes in a language-independent way, with language-specific syntax (bindings), and a few language-specific

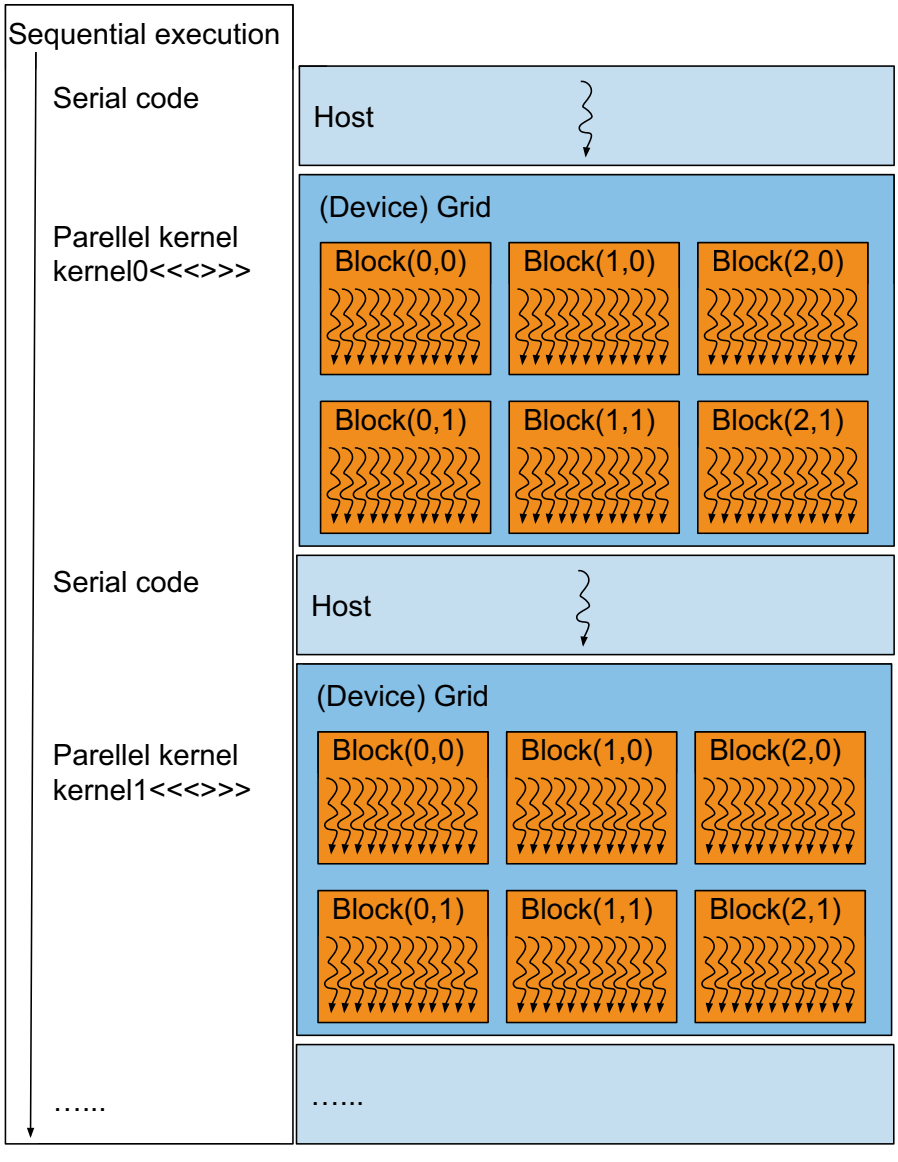


Figure 3.2. An example of CUDA programming model. The serial code executes by a host thread running on CPU and parallel kernel runs on a GPU.

features. MPI programs always work with multiple processes and typically a process is assigned to one CPU (or one core in a multi-core machine) at runtime. The essential functions of the MPI library are point-to-point operations, collective operations, process topology, synchronization. Some implementation of MPI also provides features for

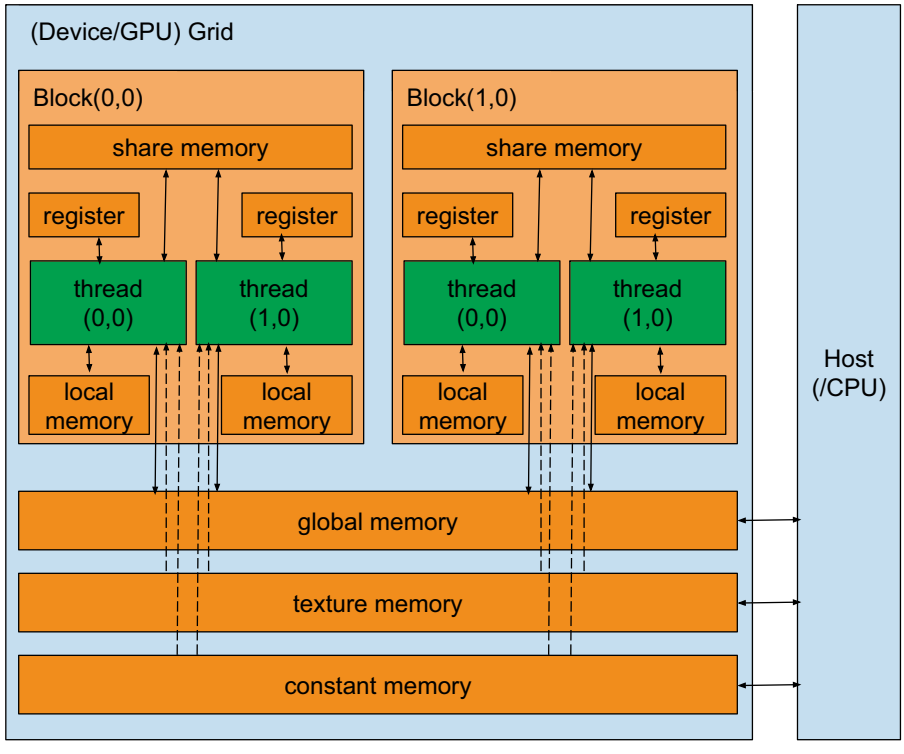


Figure 3.3. An example of GPU memory hierarchy.

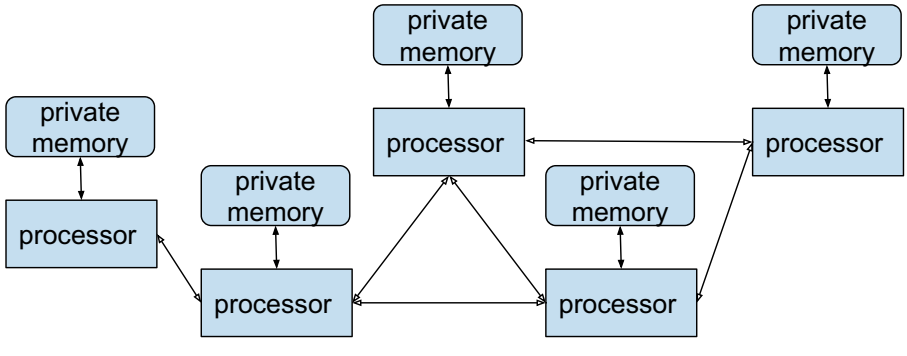


Figure 3.4. An example of memory usage for distributed computing.

parallel I/O, one-sided communication, dynamic process management, etc.

3.4 Mixed Parallel and Distributed Computing

A computing cluster allows using the merits of both parallel computing and distributed computing. In this mixed computing model, information exchange among nodes is accomplished by programming interfaces such as MPI, and via shared memory within a node. In this way, compute-intensive applications use the computation powers of cores from different nodes, and memory-intensive applications use the memory volumes of different nodes.

For HPC applications, a typical communication model in mixed computing model is the master/slave (sometimes also referred to as primary/replica) model, in which the master processor has unidirectional control over one or more other processors. For example, application developers can use OpenMP or Pthreads on a local server and exchange information via MPI among servers on a CPU cluster. The master/slave approach is also attractive for HPC applications on a heterogeneous platform. For example, in heterogeneous platforms equipped with both CPUs and GPUs, we can use a star-like communication topology, illustrated in Figure 3.5 and Figure 3.6.

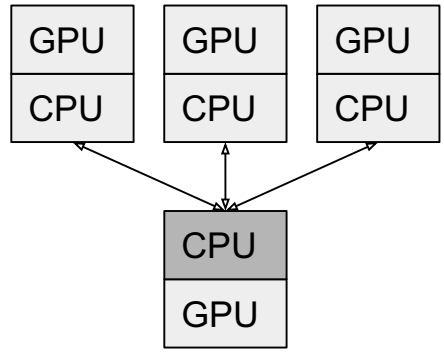


Figure 3.5. The star communication topology of a small heterogeneous platform equipped with both CPUs and GPUs.

In these two examples, all CPUs have private memories and are in charge of launching GPU kernels. CPUs and GPUs might be located physically on different servers. For a smaller topology as showed in Figure 3.5, one specified CPU works as master and takes control of all communications. For a larger topology showing in Figure 3.6, several CPUs work as master and manage information exchange within a node group, and one specified CPU works as the master of masters which is in charge of communications among node groups.

From 2013, Nvidia GPUs can exchange information via PCI express devices without interfering with any CPUs, and this technology is called NVIDIA GPUDirect™[70]. Using GPUDirect, multiple GPUs, third

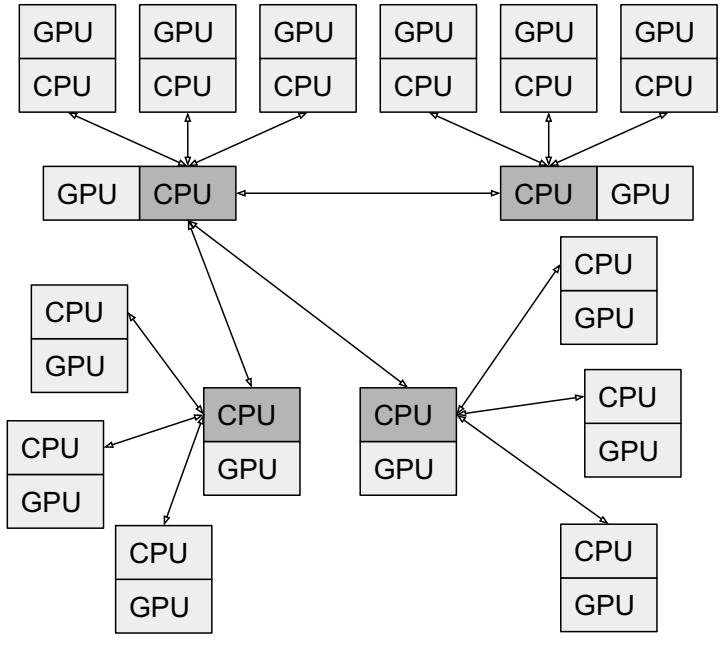


Figure 3.6. The extend star-like communication topology of a big heterogeneous platform. This example contains 4 node groups, and each group is equipped with 4 CPUs and 4 GPUs.

party network adapters, solid-state drives (SSDs) and other devices can directly read and write to CUDA memories, and hence application performances can be improved.

To sum up, the dramatical growth of the FXI data volumes make manual data analysis impossible, and hence we need to parallel and distribute our dataset and computations. Among all computation paradigms, we find our applications fit well into HPC content, in which we can utilize the advantages of the parallel and the distributed computing at the same time. The merits of start-like topology (in Figure 3.6) allows us, in ***Paper II*** (§5), to distribute dataset into a cluster with many computational nodes and parallelize computations using GPUs with minimal communications among different nodes.

Part III:
Contributions

In this thesis, we have tried to solve the challenges mentioned in §1. Figure 3.7 illustrates the pipeline of analyzing FXI data — from the raw diffraction patterns to the 3D electron structures. Our proposed analysis pipeline first selects high-quality single-particle diffraction patterns (§4) in the pre-processing stage and pushes limited but enough patterns into next stage. The reconstruction stage aligns the selected 2D patterns into a 3D Fourier intensity via our accelerated EMC implementations, which are summarized §5. The last stage, the post-analysis stage, quantitatively measures the reconstruction uncertainties (explained in §6), and transfers Fourier domain information into real domain knowledge. In §7, we illustrate this pipeline with an FXI dateset of PR772 viruses [75, 93], which was downloaded from [57].

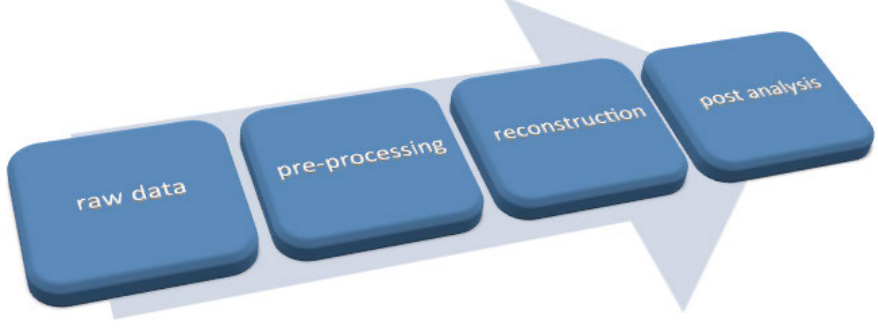


Figure 3.7. The pipeline to analyze the FXI data — from the raw diffraction patterns to the 3D electron structures. In the pre-processing stage, we select high-quality homogeneous single-particle diffraction patterns for the EMC algorithm. In the reconstruction stage, we use our accelerated EMC implementations to reconstruct 3D Fourier intensities. In the post-analysis unit, we do uncertainty analysis together with phasing, etc.

4. ***Paper I***: Classification [1]

The classification procedure prepares input data for the 3D reconstruction step, i.e. the EMC algorithm. Two supervised template-based machine learning algorithms are proposed in the paper — the Eigen Image method (EI) and the Log-Likelihood method (LL). The EI method assesses the similarity between the template diffraction patterns and the incoming testing patterns by analyzing eigenvector projections, and the LL method works on the log-likelihood function. Both methods are independent of access to the full dataset, and consequently they are easy to parallelize for achieving XFEL repetition rate. With our methods, we thus aim to select high-quality homogeneous single-particle diffraction patterns in a fast and robust way. Such datasets may hopefully help 3D assembling algorithms [53, 50] to converge more quickly and improve on the final 3D resolution.

In ***Paper I***, we tested our methods systematically, by gradually increasing the data complexity of the testing dataset, i.e. from noiseless homogeneous patterns to noisy heterogeneous patterns in both particle shapes and sizes. Moreover, we have also evaluated our methods for the mimivirus FXI data [26, 24], which downloaded from [57].

In practice, the EI classifier gave a slightly smaller pattern distances and fluence distances, and the lowest error were obtained around the template size (180 *nm*) for a synthetic icosahedral testing dataset of particle sizes between 150 nm and 210 nm, see Figure 4.1. Further, it also gave a better estimation of particle sizes — on average we obtained a minimum absolute error of 1 nm around 180 nm from both methods, and a maximum error of 4 nm.

Moreover, the EI classifier is a preferable classification method for real FXI experiments. A test classification on 578 FXI mimivirus hits from 50,712 raw diffraction patterns was performed, see Table 4.1. We measure the classifiers performances following [29, 74] closely.

This classification procedure can help EMC reducing the input data. Considering the fact that EMC can fit a "good" pattern in only one or a few rotations (due to potential particle symmetry), and smear out a "bad" pattern into many rotations, we can therefore quantify the quality

[1] **J. Liu**, G. van der Schot, and S. Engblom. (2019). Supervised classification methods for flash X-ray single particle diffraction imaging. Optics express, 27(4), 3884-3899. https://doi.org/10.1364/OE.27.003884 [49]

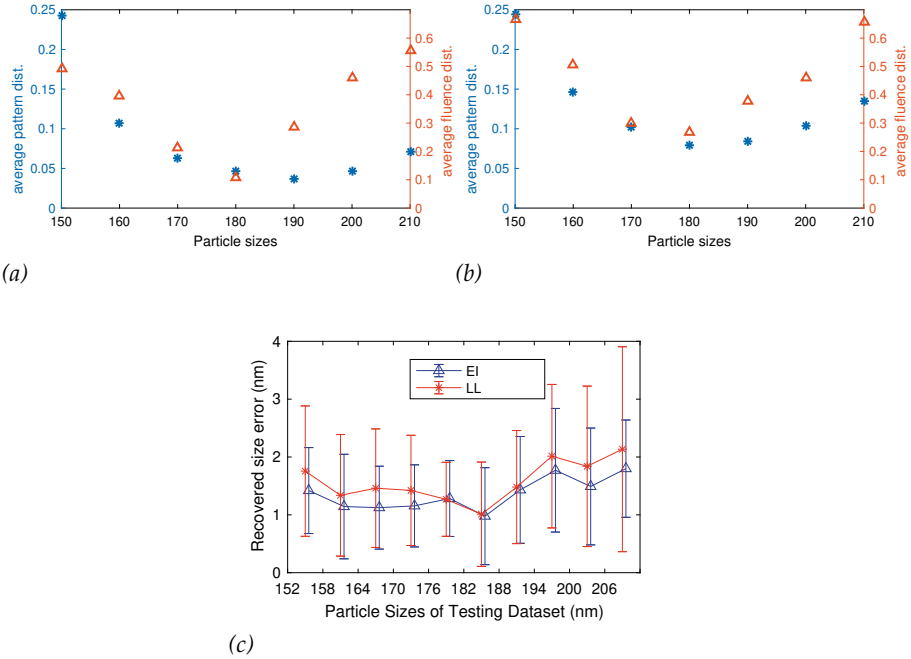
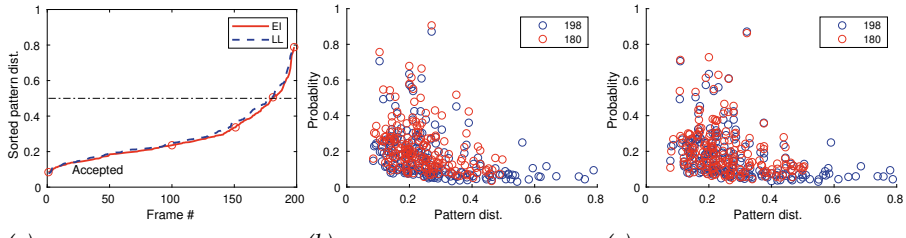


Figure 4.1. Classification of the synthetic dataset. (a): The pattern distances and the fluence distances of the testing dataset from the EI classifier. (b): The corresponding distances from the LL classifier. The testing dataset contained 2000 synthetic icosahedral diffraction patterns of different sizes. The smallest distances were obtained around the template size (180 *nm*) for both classifiers. (c): The absolute errors of the recovered sizes from the EI (*blue triangle*) and the LL (*red star*) classifier. The smallest error was obtained at around 180 *nm* particle size, and the largest error was occurred around the upper boundary of the sizes in our testing dataset. This figure is adopted from ***Paper I***.

Table 4.1. *Classification results from the EI and the LL classifier of the raw mimivirus dataset [26]. In the table, ACC, F1 ,PPV, and TPR are abbreviations of Accuracy, F1 score, Positive Predictive Value and True Positive Rate, respectively.*

	EI		LL	
	Single	Other	Single	Other
Accepted	75	33	71	38
Rejected	14	456	18	451
	ACC=0.92	TPR=0.84	ACC=0.90	TPR=0.80
	PPV =0.69	F1=0.76	PPV=0.65	F1=0.72

of selected patterns by looking at the correlation between the pattern distance and the sum of the largest N ($N = 30$ was used) rotational probabilities of each diffraction pattern, see Figure 4.2.



(a) *(b)* *(c)*

(d) *(e)* *(f)* *(g)* *(h)*

Figure 4.2. (a): The pattern distances from the EI and the LL Classifier of the mimivirus dataset used in the 3D reconstruction [26]. With a pattern-distance threshold of 0.5, both classifier accepted 180 (or 90.9%) patterns [(b) and (c)]: The sum of the largest 0.035% (the largest 30) rotational probabilities of each diffraction pattern (the rotational probabilities at the final iteration of EMC with Gaussian noise model) vs the pattern distances for the EI and LL classifier, respectively. [(d)–(h)]: Combination images at the data points (*red circles*) in (a). We plotted the raw data in the *left* part of each image and the corresponding template scaled by the recovered fluence in the *right* part of each image For the rejected diffraction patterns, (g) was slightly elongated and (h) was a smaller virus comparing to than the template.

5. *Paper II*: Accelerated EMC on GPU clusters [1]

Reconstructing a 3D Fourier intensity from the selected 2D diffraction patterns is a computational and memory intensive task. The single-GPU EMC implementation, which was used in [26], took more than 15 hours for an EMC run with only 198 patterns, and it required massive runs to determine the 3D Fourier intensity as a free parameter was used in the Gaussian model. To speed up the rotation determination procedure, we accelerated the EMC algorithm using GPU clusters in the following ways:

- Distributed EMC: The distributed EMC divides the computations into multiple GPUs by splitting the sampled rotations evenly. For one EMC iteration, each GPU computes a portion of rotation probabilities and the local 3D model from all diffraction patterns. Each processor then updates its local copy of the 3D model by averaging the 3D models from all processors. We attempted to minimize the communications among GPUs in this scheme, and implemented it with the star communication topology (see Figure 3.5).
- Fully Distributed EMC: For a very large diffraction dataset, the Distributed EMC can be slow and problematic. Therefore, our fully distributed EMC distributes the diffraction patterns together with the sampled rotations, and its implementation uses an extended star communication topology, as showed in Figure 3.6.

We have tested our distributed and fully distributed implementations on a 32-nodes homogeneous GPU cluster, where each node contains 24 Intel Xeon E5-2620 CPUs and 4 Nvidia GeForce GTX 680 GPUs. For testing our accelerated EMC implementations, we used the Gaussian noise model Eq. (2.28), which previously successfully determined a 3D electron model for the Mimivirus [26].

The *distributed* EMC implementation got a nearly perfect efficiency, as showed in Figure 5.1. This implementation is favourable for a small number of GPUs and a small number of diffraction patterns.

[1] T. Ekeberg, S. Engblom, and **J. Liu**. (2015). Machine learning for ultrafast X-ray diffraction patterns on large-scale GPU clusters. The international journal of high performance computing applications, 29(2), 233-243. http://doi.org/10.1177/1094342015572030 [25]

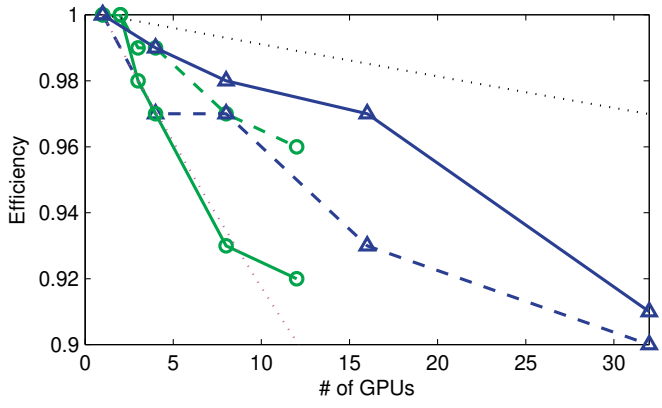


Figure 5.1. Efficiencies of the distributed EMC implementation. *Circles:* the 198 mimivirus diffraction patterns used in [26], *triangles:* 1000 synthetic patterns, *solid:* for 128^3-voxels intensity models, *dashed:* for 64^3-voxels intensity models. *Upper dotted line:* $B = 0.001$ for the Amdahl's efficiency $E = T(1)/nT(n) = 1/nB + (1-B)$, where n is the number of GPUs. *Lower dotted line:* $B = 0.01$.

We have also profiled our fully distributed EMC with 10,000 synthetic diffraction patterns. As showed in Table 5.1, we tested the implementation with up to 100 GPUs, and it achieved a higher floating point performance when compared to the single GPU implementation (32.9 GFLOPS and 39.4 GFLOPS, respectively, for the 64^3- and the 128^3- model).

	64^3		128^3	
# GPUs	Time (s)	GFLOPS/GPU	Time (s)	GFLOPS/GPU
16	164.6	36.3	552.2	43.3
32	83.5	35.8	281.2	42.5
64	42.3	35.3	141.6	42.3
96	28.3	35.2	95.4	41.8
100	27.2	35.2	91.6	41.8

Table 5.1. *Average execution time and floating point performance per GPU and per iteration using the fully distributed EMC for a a testing dataset with* 10,000 *synthetic diffraction patterns.*

Figure 5.2 displays the *scalability* of different configurations. We can interpret the scalability as the execution time per computation unit of configuration C_1 compared to configuration C_2. From the graph, we concluded that as the size of the datasets plays a more prominent role than the size of the grid, and the fully distributed EMC becomes a favourable choice.

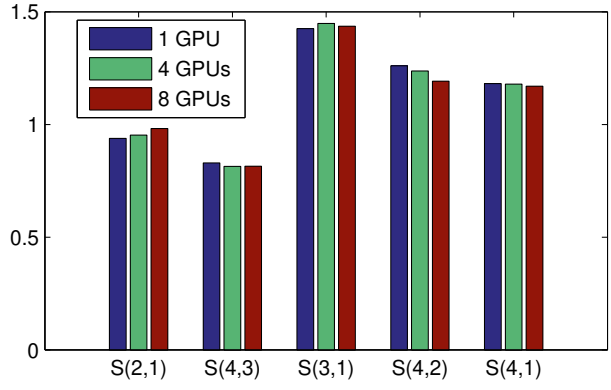


Figure 5.2. Bar plot of the *scalability* $S(C_1, C_2)$ from configuration C_1 and C_2. In this chart, configuration 1 ran the distributed EMC with 198 frames on a 64^3 grid, and configuration 2 ran EMC on the same dataset on a 128^3 grid. Configuration 3 and 4 ran EMC with 1000 synthetic frames on a 64^3 grid and a 128^3 grid, respectively.

5.1 Other Technologies

Other than distributing computations, we can also improve the efficiency by reducing the amount of computations. One adaptive way proposed in ***Paper II*** was the Adaptive EMC. By gradually increasing the sizes of the sampled rotations, instead of using a fine sampling from the beginning, it can reduce the computation time by half. We may also expect that the efficiency gain of a factor of about 2 remains also for larger load cases.

6. ***Paper III***: Uncertainty Quantification [1]

Obtaining high-quality 3D models requires the experimental developments of the FXI technology along with a comprehensive understanding of the uncertainty propagation in the EMC reconstruction procedure. In ***Paper III***, we have identified the sources of uncertainties through the reconstructing process, and measured the EMC-algorithm related errors (algorithmic errors), using a known 3D 'truth' of the Fourier intensity. With a known 3D 'truth', we measured the following algorithmic errors: the smearing, the noise, the rotational and the fluence error.

We have also contributed to bootstrap procedures estimating the reconstruction uncertainty when the 3D 'truth' is unknown. We used both the standard bootstrap and the Expectation-Maximization algorithm with bootstrapping (EMB) estimator. In brief, the standard bootstrap method ran EMC with 100 bootstrap samples, yielding 100 3D Fourier intensities, and it then calculated uncertainties from those 3D intensities. On the other hand, EMB also ran EMC with 100 bootstrap samples, but it calculated the mean and the variance of the rotational probabilities at the last iteration of each EMC run, and then it assembled the mean and the variance into 3D volumes to estimate the 3D uncertainty.

The noise model used in ***Paper III*** was the scaled Poisson model Eq. (2.29). Since the photon fluence ϕ and slices W cannot be negative, we hence borrow ideas from the non-negative matrix factorization (NNMF) and solve the scaled Poisson model as follows:

$$\phi_{jk}^{(n+1)} = \frac{\sum_i K_{ik}}{\sum_i W_{ij}^{(n)}} \frac{\sum_l \mathbb{W}_l^{(n-1)}}{\sum_l \mathbb{W}_l^{(n)}}, \quad W_{ij}^{(n+1)} = \frac{\sum_{k=1}^{M_{\text{data}}} P_{jk}^{(n+1)} K_{ik}}{\sum_{k=1}^{M_{\text{data}}} P_{jk}^{(n+1)} \phi_{jk}^{(n+1)}}, \tag{6.1}$$

where $\sum_l \mathbb{W}_l^{(n-1)} / \sum_l \mathbb{W}_l^{(n)}$ is a normalization term.

Figure 6.1 shows the total algorithmic error and the estimated uncertainties of the standard bootstrap and the EMB estimator, with respect to the voxel-to-center distance r. In the figure, R_{50} is the average error

[1] **J. Liu**, S. Engblom, and C. Nettelblad. (2018). Assessing uncertainties in X-ray single-particle three-dimensional reconstruction. Physical Review E, 98(1), 013303. http://doi.org/10.1103/PhysRevE.98.013303 [50]

for a 3D Fourier intensity, for which 50% patterns are inserted in the correct rotations, and the rest are randomly inserted. Similarly R_{100} is the average error for a 3D Fourier intensity, for which all patterns are randomly inserted. As can be seen from Figure 6.1, our estimated reconstruction uncertainty from bootstrap procedures were accurate for reconstructing larger volumes. Due to the underestimation of the smearing error, we underrated the uncertainty for the 64^3-voxel model.

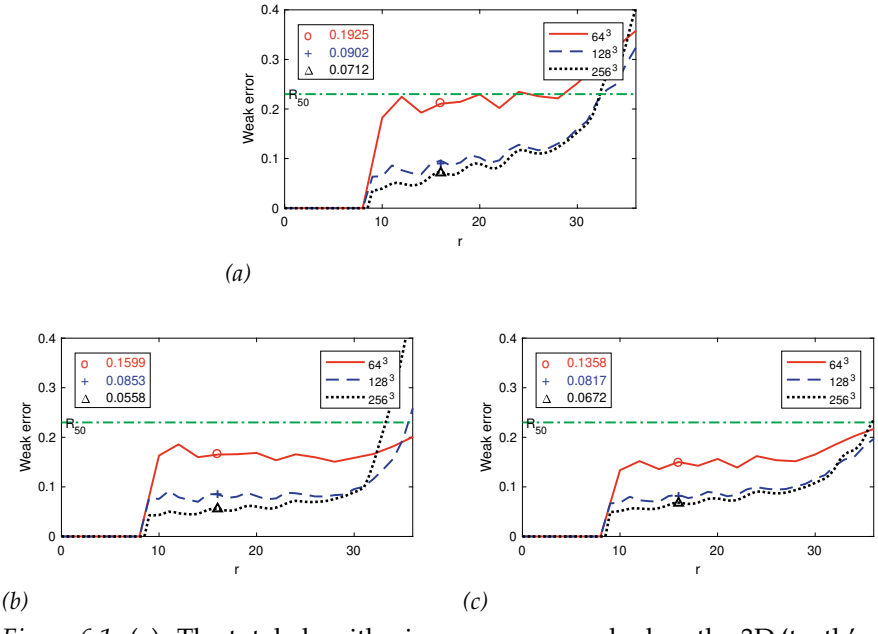
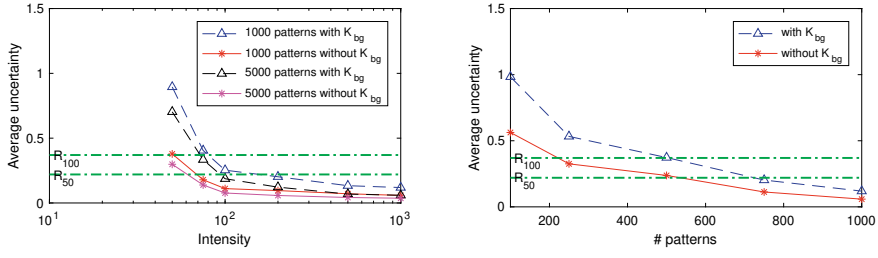


Figure 6.1. (a): The total algorithmic error, measured when the 3D 'truth' was known. [(b) and (c)]: The estimated reconstruction uncertainty from standard bootstrap procedure (b) and the EMB procedure (c). For both estimators, the 3D 'truth' was unknown. R_{50} is the 50% hidden-data error, and we can understand this as only 50% of diffraction patterns are aligned in the correct rotation.

We have also studied the influence of background noise, pattern intensity, and data volumes, see Figure 6.2. We concluded that the quality of reconstruction increases (the reconstruction uncertainty decreases) with increasing number of diffraction patterns and photons counts. On the other hand, we can also obtained better reconstructions if we could remove the background noise. With the scaled Poisson model, we needed at least 500 noiseless data frames or 750 noisy frames to pass the R_{50} threshold for a dataset similar to the Mimivirus [26]. For the R_{100} threshold, we need at least 250 diffraction patterns without background noise, or 500 frames with background noise. We therefore recommend to use at least $500-1000$ fairly high quality frames to obtain

a minimally accurate reconstruction for a dataset similar to mimivirus [26]. We also recommend to increase the number of diffraction patterns to compensate the noisy and low intensity.



(a) *(b)*

Figure 6.2. (a): The relationship among the average reconstruction uncertainty, the diffraction pattern intensity, and the number of diffraction patterns using the standard bootstrap estimator. R_{50} and R_{100} are the average 50% and 100% hidden-data error, and the later one meant all patterns were assembled into a 3D volume randomly.

7. *Paper IV*: FXI data analysis pipeline illustration [1]

Paper IV recalled the data analysis idea illustrated in Figure 3.7, and developed a proposed FXI data analysis pipeline, see Figure 7.1. Our pipeline aims to provide a fast and robust way to reconstruct 3D biomolecules from FXI diffraction patterns, and it has the potential to obtain a 3D structure during the FXI experiment. Further, we put more efforts in the post analysis step, handling multiple issues in phase retrieval, together with analysis of uncertainties and shapes.

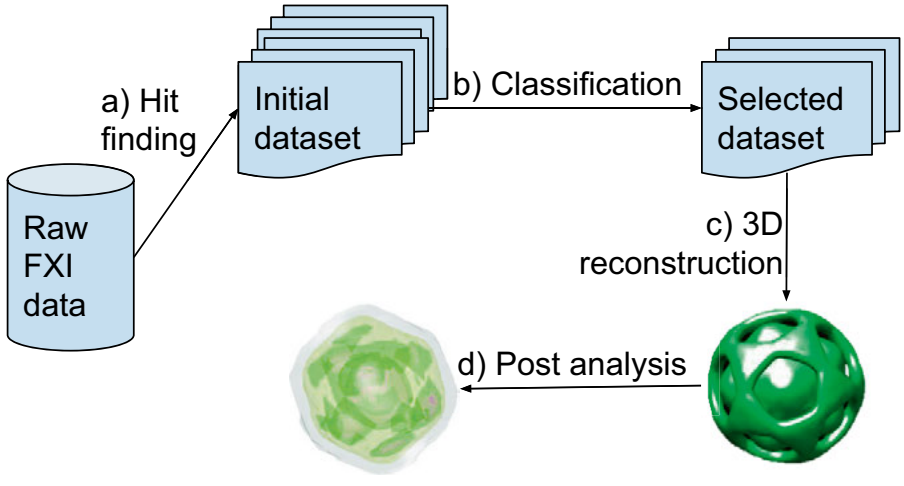


Figure 7.1. The FXI data-analysis pipeline. a) The hit finding procedure selects an initial dataset from the raw FXI data. b) From the initial dataset, the classification procedure selects high-quality homogeneous single-particle diffraction patterns with the designed features. We use our EI classifier (presented in ***Paper I***) for classification. c) The 3D reconstruction procedure assembles 3D Fourier intensity from the selected diffraction patterns via our efficient implementation of EMC (as in ***Paper II***), with the scaled-Poisson model (see ***Paper III***). d) In the Post analysis step, we retrieve the 3D real-space structure, validate the results, analyse uncertainties (see ***Paper III***), etc.

As an illustration of our pipeline, we reconstructed 3D intensities from PR772 [75] experiment. The initial dataset [75, 93] was downloaded from [57], and it contained N_{14k} = 14,772 single-hit patterns.

[1] (manuscript) **J. Liu**, S. Engblom, and C. Nettelblad. Flash X-ray Imaging in 3D: A Proposed data analysis pipeline

With our EI classifier and the thresholds of particle sizes and fluences, we have selected two datasets N_{1k} and N_{3k} that contained 1,084 and 3,140 diffraction patterns, respectively. We then pushed both datasets into the distributed EMC implementation described in ***Paper II*** with the NNMF solution of the scaled Poisson model. The algorithm converged quickly and stopped within 41 EMC iteration, i.e., we have reconstructed 3D Fourier intensities within 45 minutes using 3 Nvidia GTX 680 GPUs. Both datasets fitted well into the scaled Poisson model, we obtained large values of the probability to the most-likely rotation of each diffraction patterns, see Table 7.1. For a pattern composed of random numbers, the expectation of the probability to the most-likely rotation is approximately $1/M_{\text{rot}} \approx 2\times10^{-5}$. The minimal value we obtained for N_{3k} was 0.177 ($\gg 2\times10^{-5}$), indicating that the selected patterns fitted with the scaled Poisson model.

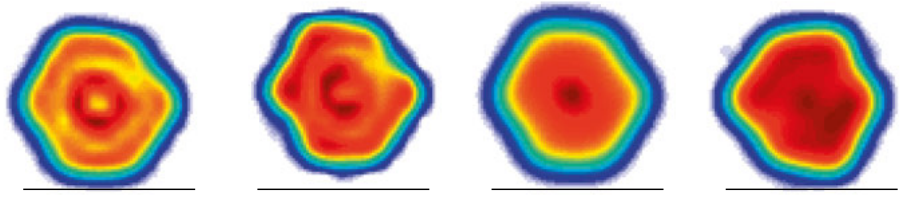
Table 7.1. *The statistics of the probability to the most-likely rotation of diffraction patterns in datasets N_{1k} and N_{3k}.*

	Max	Mean	Median	Min	Peak
N_{1k}	1	0.924	0.989	0.235	0.98
N_{3k}	1	0.846	0.940	0.177	1.0

In the post-analysis step, we retrieved real space intensities using a combination of algorithms — 10000 iterations of the relaxed averaged alternating reflections (RAAR) followed by 2000 iterations of Error Reduction (ER) algorithm. For robustness, we averaged 100 phased objects, and calculated the the phase retrieval transfer function (PRTF) to determine resolutions.

The recovered real-space intensity directly obtained from the N_{3k} dataset, see Figure 7.2(a), had pseudo-icosahedral capsids with asymmetric interior structures at the resolution of around 10.7 nm. We also observed three concentric layers and the central rings had high intensity values. The uneven intensity distribution and layers might be due to noise and aliasing effects. By subtracting the background noise per frequency bin from the 3D Fourier intensity, we obtained less concentrated interior structure in Figure 7.2(b). To handle the aliasing effects, we applied 3D Hann window over the Fourier intensities with/without background noises before phasing them, and the windowed intensities gave smoother objects in the real-space without layers, see Figure 7.2(c) and Figure 7.2(d). Further, the background subtraction improved the resolution from 10.7 nm to 8.7 / 8.4 nm with / without Hann windowing.

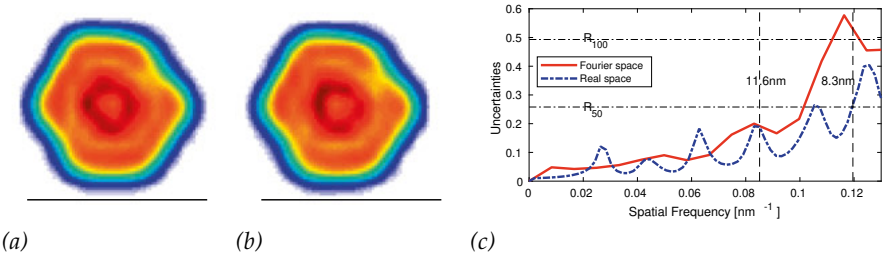
In ***Paper III***, we have brought up the bootstrap idea for quantifying the uncertainties of 3D Fourier intensities, and in ***Paper IV*** we further extend the idea to the Real-space. Briefly, the bootstrap idea estimated the total uncertainty from the estimation of bias, and standard error,



(a) (b) (c) (d)

Figure 7.2. Cross-section images of the average recovered real-space intensities. (a): the average intensity directly phased from EMC reconstruction of N_{3k}. (b): with the 3D background subtraction from Fourier intensity of (a). [(c) and (d)]: applied Hann window before phasing for the Fourier intensities of (a) and (b), respectively. The resolutions of [(a) – (d)] calculated from PRTF were 10.7 nm, 8.4 nm, 10 nm and 8.7 nm, respectively. And their estimated vertex-to-vertex distance were 69.4 nm, 68.9 nm, 69.2 nm and 69 nm.

either in the Fourier or in the Real space. Further, with the threshold of the 50% error R_{50}, we can use the uncertainty measurements to judge the resolution, obtaining a resolution of around 10 nm for the dataset N_{3k}.



(a) (b) (c)

Figure 7.3. The bootstrap results and the uncertainty analysis of the dataset N_{3k}. (a): a cross-section image of the real-space intensity that was averaged from the phased objects of the 100 bootstrap Fourier intensities. (b): the average phased object from the bootstrap mean of the 100 Fourier intensities. (c): the uncertainty measurements in the Fourier and the real space. Both analyses gave a resolution of around 10 nm, with the threshold of R_{50}.

8. Summary and Outlook

Visualizing small biology objects has been an interesting topic over many years. The Flash X-ray single-particle diffraction imaging (FXI), which relies on the theory of "diffract and destroy ", is a modern way to illuminate and reconstruct single-particle structures using XFEL pulses. We foresee that FXI experiments will be able to determine sub-nanometer structures in the future.

The incredible high repetition rate of XFELs allows capturing an enormous amount of diffraction images, and hence a robust automatic or semi-automatic pipeline with uncertainty analysis is necessary. To be more specific, this Ph.D. thesis tackled the following problems in FXI data analysis:

- **Pattern classification (in *Paper I*):** FXI experiments run in XFELs with high repetition rates, and therefore the volume of raw images is too huge to analyze manually. To select high-quality diffraction patterns with desired features, we have developed two supervised machine learning methods to classify the raw 2D diffraction data in ***Paper I***. Both methods are highly parallelizable and can speed up to match the XFELs repetition rates.
- **Efficient 3D reconstruction in Fourier space (in *Paper II*):** The captured diffraction patterns contain only intensity values, and the information of particle rotations are unobservable. The computations to determine the rotations and form a 3D Fourier intensity are enormous. To compensate for the low signal-to-noise signals from smaller objects, we need more patterns for determining rotations, and in consequence, the computations needed are enlarged. Our implementations of 3D alignment algorithms are efficient and highly scalable, which run on GPU clusters. We have tested our implementation with up to 100 GPUs, achieving up to 43.3 GFLOPS per GPU, and up to 4.2 teraFLOPS for 100 GPUs with limited efficiency loss.
- **Scaled Poissonian model(in *Paper III*):** Upon the idea adopted from Non-negative matrix factorization (NNMF), we solved the Scaled Poissonian model in the 3D alignment algorithm.
- **Uncertainty analysis (in *Paper III* and *Paper IV*):** We also introduced a practically applicable computational methodology in the form of bootstrap procedures for assessing reconstruction uncertainty for the 3D Fourier intensity and the Real-space objects. The radial plots of the 3D uncertainties can be considered as a new way to determine the resolution.

- **FXI data analysis pipeline (*Paper IV*):** With the methods proposed in [***Paper I*** – ***Paper III***], we proposed a multi-steps pipeline to handle FXI data efficiently and robustly. We also suggested ways to handle background noise and signal leakage of the 3D Fourier intensity.
- **PR772 virus structure (*Paper IV*):** As a demonstration of our data analysis pipeline, we produced the electron densities from an FXI experiment of the PR772 virus [75]. The results show the PR772 structure derivates from an ideal icosahedral symmetry, and the obtained resolution was above the detector-edge resolution (11.6 nm). However, we argue that higher resolution diffraction frames are needed for studying the internal structures of the PR772, as 11.6 nm is far away from an atomic resolution.

To achieve 5-Å resolution or better, there are still many challenges lay ahead for FXI. Researchers and technicians have made great efforts in technical issues, such as detectors, sample delivery, pulse duration, XFEL repetition rate, etc, and they are continuously improving them. On the other hand, the improvements in data analysis methods such as classification, hit-finding, data sharing, 3D orientation determination, and uncertainty analysis, etc, are also in progress. With our work, we improved the efficiency, robustness, and accuracy of the FXI data analysis, and we hope that we can obtain the 3D electron density during the FXI experiment along with the appropriate uncertainty analysis in the near future.

Summary in Swedish

Hur små saker man kan observera begränsas fundamentalt av våglängden på det ljus man använder. Därför är också undersökningar med kortvågig röntgenstrålning så populär i studier av biomolekylstrukturer. Men röntgenstrålning interagerar svagt med materien och således kan klassiska röntgeninstrument i vanliga laboratorium inte bestämma strukturen av små enstaka biomolekyler, såsom proteiner, DNA, virus, och liknande.

Med den moderna röntgenlasertekniken ("X-ray free-electron laser, XFEL") [55, 60] är det teoretiskt möjligt att avbilda enstaka molekyler. Strategin kallas "diffraktion och förstör" och använder XFEL-röntgenpulser för att skapa diffraktionssignaler innan proverna förstörs [66]. Strategin har fått stor uppmärksamhet inom strukturell biologi [39, 13, 9, 45, 26].

Den senaste metodologin baserad på idén med "diffraktion och förstör" kallas Flash X-ray single-particle diffraction imaging (FXI), eller ibland X-ray Single-Particle Imaging (SPI) [3]. I ett FXI-experiment injicerar man en ström av partiklar i röntgenstrålen, och sedan samverkar provmolekylerna med de extremt intensiva röntgenpulserna, vilket ger tvådimensionella diffraktionsmönster som visar de upplysta föremålen under olika slumpmässiga orienteringar. På grund av den höga repetitionsfrekvensen för XFEL och den slumpmässiga egenskapen hos FXI har avläsningarna från de digitala detektorerna olika kvalitéer, och en stor del av utdata består helt enkelt av tomma mönster utan någon spridning från provpartiklarna. Vi får också en betydande mängd spridningar från föroreningar och från prov som innehåller flera provpartiklar samtidigt. De mest intressanta avläsningarna är tydliga diffraktionsmönster från en ensam partikel, men tyvärr tillhör inte de flesta av avläsningarna denna klass. Vidare är avläsningarna från digitala detektorer intensiteter som inte innehåller någon fasinformation och dessutom beror på den strålintensitet provet utsattes för. Båda dessa storheter måste skattas innan partikeln kan återskapas.

Eftersom FXI studerar relativt små provpartiklar och använder diffraktionsintensiteter från fjärrfältet, kan diffraktionsmönstret från detektorerna betraktas som kontinuerliga signaler från Fourier-domänen. Översampling används som en teknik för att återskapa fasinformationen [62, 31, 78, 61]. Eftersom många biologiska partiklar finns i praktiskt taget

identiska kopior vid de relevanta upplösningsskalorna kan de tvådimensionella diffraktionsmönstren behandlas som olika orienterade exponeringar av samma partikel. Följdaktligen kan den tredimensionella strukturen erhållas genom medelvärden av 2D-diffraktionsmönster givet att partikelorienteringarna kan skattas. För att sedan få tredimensionella bilder av provpartikeln kan vi utföra en tvåstegsprocedur — rekonstruera 3D Fourier-intensiteten först och bestäm sedan information om 3D-fasen [53, 16, 7, 76]. Som ett alternativ är det också möjligt att kombinera fasalgoritmerna med rotationsbestämningen [22, 48].

Den otroliga höga repetitionsfrekvensen för XFEL gör det möjligt att ta en enorm mängd diffraktionsbilder, och därför är en robust automatisk eller halvautomatisk dataanalys med osäkerhetsskattning nödvändig.

Denna avhandling behandlar följande problem inom FXI-dataanalys:

- **Klassificering (i *Paper I*):** FXI-experiment körs i XFEL med höga repetitionshastigheter, och därför är volymen av råbilder för stor för att manuellt kunna analysera. För att välja högkvalitativa diffraktionsmönster har vi utvecklat två övervakade maskininlärningsmetoder som kan klassificera de råa tvådimensionella diffraktionsbilderna. Båda metoderna har goda paralleliseringsegenskaper och klarar därför av att matcha repetitionshastigheten hos XFEL.
- **Effektiv 3D-rekonstruktion i Fourierrymden (i *Paper II*):** De insamlade diffraktionsmönstren innehåller endast intensitetsvärden, och informationen om partikelrotationen är inte observerbar. Beräkningarna för att bestämma rotationerna och bilda en 3D Fourier-intensitet är mycket stora. För att kompensera för de brusiga signalerna från mindre objekt behöver vi behandla ett stort antal diffraktionsmönster för att skatta rotationerna, och följaktligen blir de beräkningar som behövs mycket stora. Våra implementeringar är såväl effektiva som skalbara, och kan köras effektivt på GPU-kluster. Vi har testat vår implementering med upp till 100 GPU:er, under upp till 43,3 GFLOPS per GPU, eller upp till 4,2 TFLOPS totalt med en mycket begränsad effektivitetsförlust.
- **Skalad Poisson-modell (i *Paper III*):** Baserat på en idé från algoritmer för icke-negativa matrisfaktoriseringar, lyckades vi ta fram en algoritm för att skatta en skalad Poisson-modell, en stokastisk spridningsmodell baserad på första principer. Tidigare har enklare heuristiska modeller baserade på normalfördelningar varit att föredra.
- **Osäkerhetsanalys (i *Paper III* och *Paper IV*):** Vi introducerade en praktiskt tillämpbar beräkningsmetodik i form av bootstrap-procedurer för att bedöma rekonstruktionsosäkerhet för 3D Fourier-intensiteten och för det slutligt rekonstruerade objektet. Denna metodologi är ett helt nytt sätt att skatta den erhållna upplösningen.

- **PR772-virusstruktur (*Paper IV*):** Som en praktisk demonstration av vår metodologi för dataanalys, producerade vi elektronstätheten från ett FXI-experiment med PR772-viruset [75]. Resultaten visar att PR772-strukturen kan härledas från en idealisk ikosahedral symmetri, och den erhållna upplösningen var bättre än upplösningen dikterad av detektor-kantavståndet (11.6 nm). Vi hävdar dock att högre upplösning av data behövs för att studera de interna strukturerna i PR772, eftersom 11.6 nm är långt ifrån en atomär upplösning.

För att uppnå 5 Ångströms upplösning eller bättre finns det fortfarande många utmaningar för FXI. Forskare och tekniker har gjort stora ansträngningar i tekniska frågor, såsom detektorer, provleverans, pulsvaraktighet, XFEL-repetitionsfrekvens osv. Och de förbättrar dem kontinuerligt. Å andra sidan pågår även förbättringarna i dataanalysmetoder som klassificering, delning av data, bestämning av 3D-orientering och osäkerhetsanalys osv. Arbetet i den här avhandlingen har lett till förbättringar av effektiviteten, robustheten och noggrannheten för FXI-dataanalysen. En förhoppning är att vi i framtiden kan använda dessa resultat till att snabbt och säkert bestämma noggranna tredimensionella elektrondensiteter baserat på data från FXI-experiment.

Acknowledgments

Firstly, I would like to thank all for the helps, discussions and happiness we had together. All through my PhD studies, I had issues here and there, it is you, my dear supervisors, colleagues, co-authers, and my friends and families, helped out and cheered me up.

I would like to express my special appreciation and thanks to my advisor Professor Stefan Engblom, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. You have been so patient with me, answering all my questions and sharing your experiences and knowledges with me. Your advices on both research as well as on my career have always been great.

I would also like to thank my supervisors: Filipe Maia and Janos Hajdu. You are so brilliant and supportive. I benefited a lot with the discussion and knowledges you shared with me.

I would like to thank my dear (former) colleagues in BMC: Tomas Ekerberg, Gijs van der Schot, Carl Nettelblad, Alberto Pietrini, Max Hantke, and Benedikt Daurer for being supportive and helpful. I enjoyed every seconds I had with you.

Thanks to my (former) co-workers at ITC for sharing your experiences, great ideas, excitements, etc. with me. You made my life in Sweden much more fruitful, and make my research life happier, and also made me a better teacher.

Special thanks to my co-authors: Stefan Engblom, Tomas Ekerberg, Carl Nettelblad and Gijs van der Schot for the inspiring discussions and efforts made for publications.

In the end, I also want to thank my family and friends for the help and support in my life. Many thanks to my dear mummy and daddy for encouraging my work and supporting my life. Thanks my darling Gang Xu and my baby girl Elin for cheering me up and bringing me happiness and joy, and also stopping me from working overtime. Thanks also to my dear friends Yuan Tian for being my friend and sharing your PhD life with me.

Thank you all, and hope you will have a bright future, a great life, a good health. With my best wishes!

Jing Liu
Feb. 2020
in Stochlom, Sweden

Summary in Chinese

写在最前面：这是写给老爸老妈，以及不懂英文的亲戚朋友。如需要了解更具体、更精确的信息，请阅读英文部分。

人眼可观测到的最小的物体的大小是由照射此物体的光线的波长决定的。所以，可 见光只允许人眼观测到大于200纳米的物体。为了观测更小的物体,我们需要使用 波长更短、能量更强的光线。对于生物结构学来说，波长为1Å的X光射线是一种非常好 的探测光，这是由于1Å几乎等同于一个原子的大小。 然而X光射线与物 质的交互非常弱，这意味着单个粒子在实验室X光的照射下不能产生足够强的衍射信号。 传统的结构生物学，通过结晶体来增强信号，这种方法通常被称为X射线结晶学。 但非常不幸的是，并非所有的粒子都能结晶，这也就是说我们不能使用X射线结晶学来研究它们。

高速发展的X射线电子激光器技术（X-Ray Free Electron Lasers，下简称 XFELs）使得使用X光研究单粒子结构变为可能。最新一代的XFELs技术可以产生极为明亮 的飞秒极激光束。这些激光束可在每平方微米提供10^12个光子，并且其波长 可以短至1Å。这些前提能让高速电子探测器在样品的原子 核产生显著变化之前捕获到信号，这种方法被称为“在销毁之前的衍 射”（“diffract before destroy”）。

对于瞬时X射线单粒子衍射成像（Flash X-ray single particle diffraction imaging,下简称FXI）技术而言，XFELs为其提供了波长极短,光子极密集的X射 线束。FXI使用气体或流体注射器将单粒子注 入X射线束中，从而产生二维的衍射图。这些图可使用迭代相位检索(iterative phase retrieval)的方法转化成为二维样本结构图。如若考虑 许多生物学样本都存在相同的样本，那么从这些样本获取的2维衍射图则可以用以重构样本的三维结构图。

史上第一个FXI实验是于2009年在德国汉堡的FLASH实验中心 （soft Free electon Laser in Hamber, FLASH） 完成，使用的是人造样品。而第一个非 人造样品是巨病毒（mimivius)，实验于2011年在美国的LCLS (LINAC Coherent Light Source）实验中心完成。尽管其二维重构图的分辨率仅为32 纳米，该研究克服了许多FXI实验上的技术难题，并成功进行了相位恢复，为后续的研究带来了可借鉴的经验。随后，其三维重构图也被发布了。虽然其分辨率极差， 但其作者表示若有足够多足够清晰的衍射图，巨病毒的三维重构图也将变得更为 清晰。近年来，更多的学者开始参与到FXI实验中。比巨病毒更小更同构的病毒， 如水稻矮缩病毒(Rice Dwarf Virus，RDV)、大肠杆菌噬菌体PR772(Coli phage PR772)等，也被做为实验样本送入X射线电子束中，并成功被重绘为3维结构。

为得到更多更清晰的三维重构图，除却高速发展的XFELs技术，我们需要考虑如何 更好、更有效地处理这些二维衍射图。最新一代的XFEL设备(EUXFE)每秒能产生 27000张二维衍射图，每小时可存储不小于一千万张二维衍射图。如此巨大的数 据量使得数据的储存、转移、分析等变得十分困难。其次，当前的三维重构算法 需要非常大的计算量，这是由于算法需要将单张二维衍射图置配到一 个巨大的三维离散空间，而提高分辨率则需要提供更多的二维衍射图。 再次，为了更好地理解 样本三维重构图，我们需要量化地分析三维重构的不确定性。最后，当前的数据 分析全部基于已被存储于硬盘的衍射数据，而大量的无效或低质量数据充斥于这 些储存数据中。因此，使用图像处理方法直接从探测器获得的信号中挑选优质数 据，并用于三维重构是一项极有意议的工作。

此博士论文将分篇论述以上XFI数据分析面临的挑战。

1. 从探测信号中挑选优质数据（***Paper I***)

正如前文所提，高速发展的自由电子X射线技术将为每小时产生数以万计的数据 提供可能性。为了从这些数据中提取高质量的单粒子的衍射图，我们在 ***Paper I***中提供了两种不同的分类器：本征分类器与可能性分类器。这两种方 法都是基于模板的分类方法。本征分类器使用了特征分解(Eigendecomposition） 的方式对模板进行训练，后使用其训练结果(模板的特征值与特征向量)对新的衍 射图进行分类。可能性分类器则通过计算衍射图属于某一模板的可能性程度进行 分类。这两种分类方法都极其有效与准确。前者在模板不变更的情况下平均分类 速度更快，而后者则更加灵活，模板变更并不影响其训别速率。

2. 分布式三维重构（***Paper II***)使用三维重构算法重绘粒子三维结构需要极大 的计算量，使用单显卡进行计算已无法满足我们日益增长的计算量。为此我们 通过使用信息传递介面（Message passing Interface, MPI）将数据平均地分 布到不同服务器的多个显卡（GPU）中,并将运算结果同步回主服务器,这种主 从结构能够最大限度地保证主服务器对其从属的控制并最可能地减少数据传输.在 ***Paper II***中,我们使用了主从结构的分布求计算。值得一提是,我们在最多100 个显卡(GPUs）中进行计算,其效率基本上达到了线性增长.

3. 三维重构的不确定性量化分析（***Paper III***)

在追求更精确的三维重构的过程中，不确定性的理化分析是重要且不容忽视的一 部分。这是因为通过分析三维重构过程的不确定性，可以有效地发现算法漏洞并 提升算法的准确性。

在***Paper III***中，我们分离算法中使用的变量并对它们对算法结果地影响逐个进行分析，通过模拟数据确定了算法精确度瓶颈。

而对于实际数据，我们应用自助法(Bootstrap methods)于三维重构算法，通过 对衍射图进行自助抽样与重构。我们可以研究重构图的均值与方差，从而达到定 量分析重构图不确定性的目的。 在***Paper III***中，我们提出了延展式泊松分布(scaled Poisson model)，并使用 非负矩阵分解(Non-negative matrix factorization)。

4. 即时重构（***Paper IV***)

以上三篇论文为那时重构提供的理论与实践基础.那时重构意味着粒子被X射线照 射瞬间产生的衍射图将立即被重构,也就是说，我们有望输入粒子到X射线机中的 同时得到其三维重构图。这个过程分为在衍射图被探测器捕获之后分为以下步骤:

i.由分类算法对捕获信息进行分类。

ii.将分类后的衍射图应用于三维重构算法。

iii.将三维重构从倒易（傅立叶）空间(recipical/Fourier domain)转换到正格 空间(real domain)。

iv).控制及测量重构的不确定性。

使用上述步骤，我们成功地重构PR772病毒的三维结构。

在获取5埃格斯特朗（5Å）或者更好的单粒子结构的路上，我们依然面临着诸多挑战。研究人员及工程师在FXI技术领域的诸多方面做出了杰出的贡献，例如检测器，样品输送，脉冲持续时间，XFEL重复率等。 另一方面，数据科学家和数据工程师也在不断地改进FXI数据分析算法的各个方面，例如：分类，数据共享，三维结构重构和不确定性分析等数据分析方法等。 通过我们的工作，我们提高了FXI数据分析的效率和准确性。我们希望在未来我们的工作可以在FXI实验期间获得三维样品电子密度。

Acknowledgments in Chinese

首先，感谢我所在的学院的各位老师，你们严谨的治学风范深深影响了我。最需 要感谢的是我的导师们。你们和蔼 可亲、幽默风趣。在这几年的学习中，从实验设计思路到具体调研，老师都给予 了耐心的指导。让我深深地学习到了：理论源于实践，又指导实践。我同时要感谢我的科研小伙伴们。 谢谢你们为我解答各种疑问，与我讨论解决问题 的不同方法。

再次，要感谢我的所有家人和亲人，感谢家人的帮助和关心，感谢家人在工作和 学习中的关注和帮助。经过一番努力，得以顺利完成学业，对自己的专业从开始 的慢慢熟悉、进步，再到现在的得心应手，老师和同学们的关心与帮助，家人的 教育和培养，父母的支持与鼓励，无疑是我努力至今的力量源泉。感谢父亲母亲 这么多年来的操劳付出，感谢爸爸妈妈在我学业遇到挫折时，及时鼓励我。

最后，感谢所有这么多年走过来的好朋友，尤其要感谢我的爱人，在繁忙的工作 之余，还要帮我解决一些技术问题，感谢你陪我走过的风风雨雨，感谢你对我无 微不至的关怀和帮助。感谢在我学业任务繁忙时，为我做各种美食、咖啡，教我 放松。感谢我的小宝贝，他让我体会到了母亲的伟大，让我快乐，更让我明白了 如何做一个合格的家长以及如何更加合理高效地使用时间。

写在最后的最后：2020年开春似乎就是一个不平凡的开春。我们的国家正在遭受 着自03年以来最大的传染性病毒袭击。身在国外的我们虽不能与国人并肩作战， 却也感同身受。愿新型冠状病毒感染人数拐点早点到来，愿大家身体健康，不受 病毒的侵扰。

祝所有人安好！幸福！快乐！

刘静

2020年2月

于斯德哥尔摩，瑞典

References

[1] A. Allahgholi, J. Becker, L. Bianco, et al. AGIPD, a high dynamic range fast detector for the european XFEL. *Journal of Instrumentation*, 10(01):C01023, 2015. doi:10.1088/1748-0221/10/01/c01023.

[2] Massimo Altarelli, R Brinkmann, M Chergui, et al. The european X-ray free-electron laser. *Technical Design Report, DESY*, 97:1–26, 2006.

[3] A Aquila, A Barty, C Bostedt, et al. The linac coherent light source single particle imaging road map. *Structural Dynamics*, 2(4):041701, 2015. doi:10.1063/1.4918726.

[4] Anton Barty, Richard A. Kirian, Filipe R. N. C. Maia, et al. *Cheetah*: software for high-throughput reduction and analysis of serial femtosecond x-ray diffraction data. *Journal of Applied Crystallography*, 47(3):1118–1131, 2014. doi:10.1107/S1600576714007626.

[5] Anton Barty, Regina Soufli, Tom McCarville, et al. Predicting the coherent X-ray wavefront focal properties at the Linac Coherent Light Source (LCLS) X-ray free electron laser. *Optics Express*, 17(18):15508–15519, 2009. doi:10.1364/OE.17.015508.

[6] Heinz H Bauschke, Patrick L Combettes, and D Russell Luke. Hybrid projection–reflection method for phase retrieval. *Journal of the Optical Society of America A*, 20(6):1025–1034, 2003. doi:10.1364/JOSAA.20.001025.

[7] Christopher M. Bishop, Markus Svensén, and Christopher K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998. doi:10.1162/089976698300017953.

[8] Gabriel Blaj, Pietro Caragiulo, Gabriella Carini, et al. X-ray detectors at the Linac Coherent Light Source. *Journal of Synchrotron Radiation*, 22(3):577–583, 2015. doi:10.1107/S1600577515005317.

[9] Michael J Bogan, W Henry Benner, Sébastien Boutet, et al. Single particle x-ray diffractive imaging. *Nano letters*, 8(1):310–316, 2008. doi:10.1021/nl072728k.

[10] Michael J Bogan, Dmitri Starodub, Christina Y Hampton, and Raymond G Sierra. Single-particle coherent diffractive imaging with a soft X-ray free electron laser: towards soot aerosol morphology. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 43(19):194013, 2010. doi:10.1088/0953-4075/43/19/194013.

[11] Gábor Bortel and Miklós Tegze. Common arc method for diffraction pattern orientation. *Acta Crystallographica Section A*, 67(6):533–543, 2011. doi:10.1107/S0108767311036269.

[12] J. D. Bozek. AMO instrumentation for the LCLS X-ray FEL. *The European Physical Journal Special Topics*, 169(1):129–132, 2009. doi:10.1140/epjst/e2009-00982-y.

[13] J. Chalupský, L. Juha, J. Kuba, et al. Characteristics of focused soft X-ray free-electron laser beam determined by ablation of organic molecular solids. *Optics Express*, 15(10):6036–6043, 2007. doi:10.1364/OE.15.006036.
[14] Henry N. Chapman, Anton Barty, Michael J. Bogan, et al. Femtosecond diffractive imaging with a soft-X-ray free-electron laser. *Nature Physics*, 2(12):839–843, 2006.
[15] Henry N. Chapman, Carl Caleman, and Nicusor Timneanu. Diffraction before destruction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1647), 2014. doi:10.1098/rstb.2013.0313.
[16] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006. doi:10.1016/j.acha.2006.04.006.
[17] Leonardo Dagum and Ramesh Menon. OpenMP: an industry standard API for shared-memory programming. *IEEE computational science and engineering*, 5(1):46–55, 1998. doi:10.1109/99.660313.
[18] D. Damiani, M. Dubrovin, I. Gaponenko, et al. Linac Coherent Light Source data analysis using it psana. *Journal of Applied Crystallography*, 49(2):672–679, 2016. doi:10.1107/S1600576716004349.
[19] B. J. Daurer, M. F. Hantke, C. Nettelblad, and F. R. N. C. Maia. Hummingbird : monitoring and analyzing flash X-ray imaging experiments in real time. *Journal of applied crystallography*, 49:1042–1047, 2016.
[20] C David, S Gorelick, S Rutishauser, et al. Nanofocusing of hard X-ray free electron laser pulses using diamond based Fresnel zone plates. *Scientific reports*, 1:57, 2011. doi:10.1038/srep00057.
[21] D P DePonte, U Weierstall, K Schmidt, J Warner, et al. Gas dynamic virtual nozzle for generation of microscopic droplet streams. *Journal of Physics D: Applied Physics*, 41(19):195505 – 19512, 2008. doi:10.1088/0022-3727/41/19/195505.
[22] Jeffrey J Donatelli, Peter H Zwart, and James A Sethian. Iterative phasing for fluctuation x-ray scattering. *Proceedings of the National Academy of Sciences*, 112(33):10286–10291, 2015. doi:10.1073/pnas.1513738112.
[23] Richard J Dudley. *Microsoft Azure: Enterprise Application Development*. Packt Publishing Ltd, 2010.
[24] Tomas Ekeberg. CXIDB ID 30, 2015. doi:10.11577/1236752.
[25] Tomas Ekeberg, Stefan Engblom, and Jing Liu. Machine learning for ultrafast X-ray diffraction patterns on large-scale gpu clusters. *International Journal of High Performance Computing Applications*, pages 233–243, 2015. doi:10.1177/1094342015572030.
[26] Tomas Ekeberg, Martin Svenda, Chantal Abergel, et al. Three-dimensional reconstruction of the giant mimivirus particle with an X-ray free-electron laser. *Physical Review Letters*, 114(9), 2015. doi:10.1103/PhysRevLett.114.098102.
[27] Veit Elser. Phase retrieval by iterated projections. *Journal of the Optical Society of America A*, 20(1):40–55, 2003. doi:10.1364/JOSAA.20.000040.
[28] Stefan Engblom and Jing Liu. X-ray laser imaging of biomolecules using multiple GPUs. In *Parallel Processing and Applied Mathematics*, Lecture

Notes in Computer Science. Springer, 2014. doi:10.1007/978-3-642-55224-3_45.
[29] Tom Fawcett. Introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006. doi:10.1016/j.patrec.2005.10.010.
[30] J. R. Fienup. Reconstruction of an object from the modulus of its fourier transform. *Optics Letters*, 3(1):27–29, 1978. doi:10.1364/OL.3.000027.
[31] James R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982. doi:10.1364/AO.21.002758.
[32] Mark Fox. *Quantum optics: an introduction*, volume 15, chapter 5, pages 75–105. OUP Oxford, 2006. doi:10.1063/1.2784691.
[33] R.W. Gerchberg and W.O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–250, 1971.
[34] Dimitrios Giannakis, Peter Schwander, and Abbas Ourmazd. The symmetries of image formation by scattering. I. Theoretical framework. *Optics Express*, 20(12):12799–12826, 2012. doi:10.1364/OE.20.012799.
[35] D Greiffenberg. The AGIPD detector for the european XFEL. *Journal of Instrumentation*, 7(01):C01103–C01103, 2012. doi:10.1088/1748-0221/7/01/c01103.
[36] Max F. Hantke, Tomas Ekeberg, and Filipe R. N. C. Maia. *Condor*: a simulation tool for flash X-ray imaging. *Journal of Applied Crystallography*, 49(4):1356–1362, 2016. doi:10.1107/S1600576716009213.
[37] Max F. Hantke, Dirk Hasse, Filipe R.N.C. Maia, et al. High-throughput imaging of heterogeneous cell organelles with an X-ray laser. *Nature Photonics*, 8(12), 2014. doi:10.1038/nphoton.2014.270.
[38] Max Felix Hantke. *Coherent Diffractive Imaging with X-ray Lasers*. PhD thesis, Uppsala University, Molecular biophysics, 2016.
[39] Stefan P. Hau-Riege, Richard A. London, Henry N. Chapman, et al. Encapsulation and diffraction-pattern-correction methods to reduce the effect of damage in X-ray diffraction imaging of single biological molecules. *Physical Review Letters*, 98(19):198302, 2007. doi:10.1103/PhysRevLett.98.198302.
[40] Stefan P. Hau-Riege and Tom Pardini. The effect of electron transport on the characterization of X-ray Free-Electron Laser pulses via ablation. *Applied Physics Letters*, 111(14):144102, 2017. doi:10.1063/1.4996190.
[41] A Hosseinizadeh, P Schwander, A Dashti, R Fung, RM D'Souza, and A Ourmazd. High-resolution structure of viruses from random diffraction snapshots. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1647):20130326, 2014. doi:10.1098/rstb.2013.0326.
[42] Ahmad Hosseinizadeh, Ghoncheh Mashayekhi, Jeremy Copperman, et al. Conformational landscape of a virus by single-particle X-ray scattering. *Nature Methods*, 14:877, 2017. doi:10.1038/nmeth.4395.
[43] J. H. Jungmann-Smith, A. Bergamaschi, M. Brückner, et al. Towards hybrid pixel detectors for energy-dispersive or soft X-ray photon science. *Journal of Synchrotron Radiation*, 23(2):385–394, 2016. doi:10.1107/S1600577515023541.
[44] Takashi Kameshima, Shun Ono, Togo Kudo, et al. Development of an

X-ray pixel detector with multi-port charge-coupled device for X-ray free-electron laser experiments. *Review of Scientific Instruments*, 85(3):033110, 2014. doi:10.1063/1.4867668.

[45] Stephan Kassemeyer, Jan Steinbrener, Lukas Lomb, et al. Femtosecond free-electron laser X-ray diffraction data sets for algorithm development. *Optics Express*, 20(4):4149–4158, 2012. doi:10.1364/OE.20.004149.

[46] R. A. Kirian, S. Awel, N. Eckerskorn, et al. Simple convergent-nozzle aerosol injector for single-particle diffractive imaging with X-ray free-electron lasers. *Structural Dynamics*, 2(4):041717, 2015. doi:10.1063/1.4922648.

[47] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990. doi:10.1109/5.58325.

[48] Ruslan P. Kurta, Jeffrey J. Donatelli, Chun Hong Yoon, et al. Correlations in scattered X-ray laser pulses reveal nanoscale structural features of viruses. *Physical Review Letters*, 119:158102, 2017. doi:10.1103/PhysRevLett.119.158102.

[49] J. Liu, G. van der Schot, and S. Engblom. Supervised Classification Methods for Flash X-ray single particle diffraction Imaging. *Optics Express*, 5:3884–3899, 2019. doi:10.1364/OE.27.003884.

[50] Jing Liu, Stefan Engblom, and Carl Nettelblad. Assessing uncertainties in X-ray single-particle three-dimensional reconstruction. *Physical Review E*, 98(1):013303, 2018. doi:10.1103/PhysRevE.98.013303.

[51] Miron Livny, Jim Basney, Rajesh Raman, and Todd Tannenbaum. Mechanisms for high throughput computing. *SPEEDUP journal*, 11(1):36–40, 1997.

[52] NeTe Duane Loh, Michael J. Bogan, Veit Elser, et al. Cryptotomography: Reconstructing 3D Fourier intensities from randomly oriented single-shot diffraction patterns. *Physical Review Letters*, 104:225501, 2010. doi:10.1103/PhysRevLett.104.225501.

[53] NeTe Duane Loh and Veit Elser. Reconstruction algorithm for single-particle diffraction imaging experiments. *Physical Review E*, 80(2):026705, 2009. doi:10.1103/PhysRevE.80.026705.

[54] D Russell Luke. Relaxed averaged alternating reflections for diffraction imaging. *Inverse Problems*, 21(1):37–50, 2005. doi:10.1088/0266-5611/21/1/004.

[55] John M. J. Madey. Stimulated emission of bremsstrahlung in a periodic magnetic field. *Journal of Applied Physics*, 42(5):1906–1913, 1971. doi:10.1063/1.1660466.

[56] F. R. N. C. Maia, T. Ekeberg, D. van der Spoel, and J. Hajdu. Hawk : the image reconstruction package for coherent X-ray diffractive imaging. *Journal of applied crystallography*, 43(6):1535–1539, 2010. doi:10.1107/S0021889810036083.

[57] Filipe R. N. C. Maia. The Coherent X-ray Imaging Data Bank. *Nature methods*, 9:854–855, 2012. doi:10.1038/nmeth.2110.

[58] Stefano Marchesini. Phase retrieval and saddle-point optimization. *Journal of the Optical Society of America A*, 24(10):3289–3296, 2007. doi:10.1364/JOSAA.24.003289.

[59] Stefano Marchesini, H He, Henry Chapman, et al. X-ray image reconstruction from a diffraction pattern alone. *Physical Review B*, 68:1401011–1401014, 2003. doi:10.1103/PhysRevB.68.140101.
[60] Brian Mcneil and Neil R. Thompson. X-ray free-electron lasers. *Nature Photonics*, 4:814, 2010. doi:10.1038/nphoton.2010.239.
[61] Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400:342–344, 1999. doi:10.1038/22498.
[62] Teague Michael Reed. Deterministic phase retrieval: a Green's function solution. *Journal of the Optical Society of America*, 73(11):1434–1441, 1983. doi:10.1364/JOSA.73.001434.
[63] SV Milton, E Gluskin, ND Arnold, et al. Exponential gain and saturation of a self-amplified spontaneous emission free-electron laser. *Science*, 292(5524):2037–2041, 2001. doi:10.1126/science.1059955.
[64] A. Munke, G. van der Schot, F. Maia, et al. Coherent diffraction of single rice dwarf virus particles using soft X-rays at the Linac Coherent Light Source. *Nature Scientific Data*, 3:160064, 2018. doi:10.1038/sdata.2016.64.
[65] Aaftab Munshi, Benedict Gaster, Timothy G. Mattson, and Dan Ginsburg. *OpenCL programming guide*. Pearson Education, 2011.
[66] Richard Neutze, Remco Wouts, David van der Spoel, et al. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, 406(6797):752–757, 2000. doi:10.1038/35021099.
[67] Bradford Nichols, Dick Buttlar, and Jacqueline Farrell. *Pthreads programming: A POSIX standard for better multiprocessing*. O'Reilly Media Inc., 1996.
[68] CUDA Nvidia. *CUBLAS LIBRARY*. NVIDIA Corporation, 8th edition, 2017.
[69] CUDA Nvidia. *CUSPARSE LIBRARY*. NVIDIA Corporation, 8th edition, 2017.
[70] CUDA Nvidia. *DEVELOPING A LINUX KERNEL MODULE USING RDMA FOR GPUDIRECT*. NVIDIA Corporation, 8th edition, 2017.
[71] David Paganin. *Coherent X-ray optics*. Number 6. Oxford University Press on Demand, 2006. doi:10.1093/acprof:oso/9780198567288.001.0001.
[72] Kanupriya Pande, Jeffrey J. Donatelli, Erik Malmerberg, et al. Ab initio structure determination from experimental fluctuation X-ray scattering data. *Proceedings of the National Academy of Sciences*, 115(46):11772–11777, 2018. doi:10.1073/pnas.1812064115.
[73] H. J. Park, N. D. Loh, R. G. Sierra, et al. Toward unsupervised single-shot diffractive imaging of heterogeneous particles using X-ray free-electron lasers. *Optics Express*, 21(23):28729–28742, 2013. doi:10.1364/OE.21.028729.
[74] David Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2:37–63, 2008.
[75] Hemanth K N Reddy, Chun Hong Yoon, Andrew Aquila, et al. Coherent soft X-ray diffraction imaging of coliphage PR772 at the Linac coherent

light source. *Scientific Data*, 4:170079, 2017. doi:/10.1038/sdata.2017.79.
[76] Max Rose, Sergey Bobkov, Kartik Ayyer, et al. Single-particle imaging without symmetry constraints at an X-ray free-electron laser. *IUCrJ*, 5, 2018. doi:10.1107/S205225251801120X.
[77] D. K. Saldin, H.-C. Poon, P. Schwander, M. Uddin, and M. Schmidt. Reconstructing an icosahedral virus from single-particle diffraction experiments. *Optics Express*, 19(18):17318–17335, 2011. doi:10.1364/OE.19.017318.
[78] D Sayre, HN Chapman, and J Miao. On the extendibility of X-ray crystallography to noncrystals. *Acta Crystallographica Section A: Foundations of Crystallography*, 54(2):232–239, 1998. doi:10.1107/S0108767397015572.
[79] J. Schulz, J. Bielecki, R. B. Doak, et al. A versatile liquid-jet setup for the European XFEL. *Journal of Synchrotron Radiation*, 26:339–345, 2019. doi:10.1107/S1600577519000894.
[80] Peter Schwander, Dimitrios Giannakis, Chun Hong Yoon, and Abbas Ourmazd. The symmetries of image formation by scattering. II. Applications. *Optics Express*, 20(12):12827–12849, 2012. doi:10.1364/OE.20.012827.
[81] M. Marvin Seibert, Tomas Ekeberg, Filipe R. N. C. Maia, et al. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*, 470(7332):78–81, 2011. doi:10.1038/nature09748.
[82] Raymond G. Sierra, Hartawan Laksmono, Jan Kern, et al. Nanoflow electrospinning serial femtosecond crystallography. *Acta Crystallographica Section D*, 68(11):1584–1587, 2012. doi:10.1107/S0907444912038152.
[83] Craig A. Stewart, Richard Knepper, Matthew R. Link, Marlon Pierce, Eric Wernert, and Nancy Wilkins-Diehr. *Cyberinfrastructure, Cloud Computing, Science Gateways, Visualization, and Cyberinfrastructure Ease of Use*, chapter 92. IGI Global, 2017. doi:10.4018/978-1-5225-2255-3.ch092.
[84] Lothar Strüder, Sascha Epp, Daniel Rolles, et al. Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 614(3):483 – 496, 2010. doi:10.1016/j.nima.2009.12.053.
[85] Zhibin Sun, Jiadong Fan, Haoyuan Li, and Huaidong Jiang. Current status of single particle imaging with X-ray lasers. *Applied Sciences*, 8(1):132, 2018. doi:10.3390/app8010132.
[86] Ryotaro Tanaka, T Fukui, Y Furukawa, et al. Inauguration of the XFEL Facility, SACLA, in SPring-8. *Proc. of ICALEPCS2011*, pages 585–588, 2011.
[87] Miklós Tegze and Gábor Bortel. Atomic structure of a single large biomolecule from diffraction patterns of random orientations. *Journal of structural biology*, 179(1):41–45, 2012. doi:10.1016/j.jsb.2012.04.014.
[88] J. Thayer, D. Damiani, C. Ford, and others. Data systems for the Linac Coherent Light Source. *Journal of Applied Crystallography*,

49(4):1363–1369, 2016. doi:10.1107/S1600576716011055.
[89] Marin van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of Structural Biology*, 151(3):250–262, 2005. doi:10.1016/j.jsb.2005.05.009.
[90] U. Weierstall, J. C. H. Spence, and R. B. Doak. Injector for scattering measurements on fully solvated biospecies. *Review of Scientific Instruments*, 83(3):035108, 2012. doi:10.1063/1.3693040.
[91] Uwe Weierstall, Daniel James, Chong Wang, et al. Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nature communications*, 5:3309, 2014. doi:10.1038/ncomms4309.
[92] Tom White. *Hadoop: The definitive guide*. O'Reilly Media, Inc., 2012.
[93] Chun Hong Yoon. CXIDB ID 58, 2017. doi:10.11577/1349664.
[94] Chun Hong Yoon, Mikhail V Yurkov, Evgeny A Schneidmiller, et al. A comprehensive simulation framework for imaging single particles and biomolecules at the European X-ray Free-Electron Laser. *Scientific reports*, 6:24791, 2016. doi:10.1038/srep2479.
[95] Lawrence S. Young, Elliot P. Kanter, Bertold Krässig, et al. Femtosecond electronic response of atoms to ultra-intense X-rays. *Nature*, 466(7302):56–61, 2010. doi:10.1038/nature09177.
[96] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. Technical Report UCB/EECS-2010-53, EECS Department, University of California, Berkeley, 2010.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1905

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-403878

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2020