



# Analysis of single particle imaging experiments at X-ray Free-Electron Lasers

## Dissertation

zur Erlangung des Doktorgrades  
an der Fakultät für Mathematik, Informatik und  
Naturwissenschaften  
Fachbereich Physik  
der Universität Hamburg

vorgelegt von

DAMELI ASSALAUOVA

Hamburg

2022

Gutachter der Dissertation:

Prof. Dr. Andreas Stierle  
Prof. Dr. Ivan A. Vartaniants

Zusammensetzung der Disputation:

Prof. Dr. Andreas Stierle  
Prof. Dr. Ivan A. Vartaniants  
Prof. Dr. Edgar Weckert  
Prof. Dr. Daniela Pfannkuche  
Dr. Kartik Ayyer

Vorsitzender der Prüfungskommission:

Prof. Dr. Daniela Pfannkuche

Datum der Disputation:

11 Oktober 2022

Vorsitzender des Fach-Promotionsausschusses PHYSIK:

Prof. Dr. Wolfgang Parak

Leiter des Fachbereichs PHYSIK:

Prof. Dr. Günter Sigl

Dekan der Fakultät MIN:

Prof. Dr.-Ing. Norbert Ritter

---

## Kurzfassung

Single Particle Imaging (SPI) ist eine neue Technik in der Röntgenwissenschaft, die darauf abzielt, die dreidimensionale Struktur von Nanopartikeln zu rekonstruieren. Die COVID-19-Pandemie zeigte die Notwendigkeit wissenschaftlicher Entwicklungen vor allem im Bereich der Untersuchung innerer Strukturen biologischer Partikel. Der Hauptvorteil dieses Ansatzes besteht darin, dass atomare Strukturen in ihrer natürlichen nicht-Kristallinen Umgebung aufgelöst werden können.

SPI-Experimente erfordern den Einsatz kurzwelliger elektromagnetischer Strahlung im Sub-Nanometer-Bereich (Röntgenstrahlung), um die innere Struktur des Objekts aufzulösen zu können. Aufgrund der schwachen Wechselwirkung von Röntgenstrahlen mit weicher Materie eine hohe Kohärenz und ein hoher Photonenfluss erforderlich, um die feinsten Strukturen des Objekts aufzulösen. Durch die extreme Strahlendosis werden die biologischen Partikel im Streuprozess zerstört. Um Beugungsmuster der unbeschädigten Struktur zu erhalten, muss die Dauer des Röntgenpulses kürzer sein als die typische Zeitskala des Zerstörungsprozesses. Aus diesem Grund können Synchrotronlichtquellen hoher Brillanz nicht verwendet werden, da der kohärente Fluss in einem einzigen Puls zu gering ist, der für die Aufzeichnung eines ausreichenden Signals erforderlich wäre. Durch die Entwicklung von Röntgenquellen mit hoher Intensität und kurzer Pulsdauer – Freie-Elektronen-Röntgenlaser (XFEL) – kann diese Hürde überwunden werden.

Bei der SPI-Methode werden viele identische Partikel des untersuchten Systems in den Röntgenstrahl injiziert, wodurch Beugungsbilder in zufälligen Orientierungen entstehen. Die dreidimensionale Struktur des Objekts wird durch Anwendung komplexer Algorithmen auf die gesammelten Beugungsmuster ermittelt. Die Größe eines solchen Datensatzes kann Terabytes übersteigen, was die Entwicklung und Implementierung von ausgeklügelten Datenanalysetechniken erfordert, die helfen, wertvolle XFEL-Messzeit zu sparen und die Datenverarbeitung zu beschleunigen.

Die ersten beiden Teile dieser Arbeit basieren auf der methodischen Entwicklung des SPI-Datenanalyse-Workflows. Die experimentellen Daten wurden mit dem Virus PR772 an der Linac Coherent Light Source (LCLS) am SLAC in Stanford, USA, im Rahmen des SPI-Konsortiums gesammelt. Als Ergebnis der entwickelten Methodik, die auch die Klassifizierung von Objekten durch maschinelles Lernen umfasst, konnte eine dreidimensionale Virusstruktur mit einer Auflösung von weniger als 10 Nanometern rekonstruiert werden. Der Vergleich der Ergebnisse mit den kryogenmikroskopischen Untersuchungen zeigte ähnliche Merkmale und eine generelle Übereinstimmung zwischen beiden Techniken. Aufgrund der Komplexität und der Kosten der SPI-Experimente ist die Vorbereitung ein zeit- und arbeitsintensiver Prozess, der eine umfassende Planung erfordert. Der dritte Teil dieser Arbeit befasst sich mit der Optimierung der Aufbauparameter durch die Simulation eines

---

SPI-Experiments mit dem Zeckenzephalitis-Virus. Diese Simulationen trugen zum Erfolg eines tatsächlichen Experiments bei, das am European XFEL in Hamburg durchgeführt wurde.

---

## Abstract

Single particle imaging (SPI) is a novel technique in X-ray science aimed at reconstructing the three-dimensional structure of nanoscale objects. Studying the inner structure of biological particles has become increasingly crucial, as evidenced by the pandemic of coronavirus disease (COVID-19) showing the necessity of scientific development in this field. The main advantage of this approach is that atomic structures can be resolved in their native environment without crystallization.

SPI experiments require using electromagnetic radiation with a sub-nanometer wavelength (such as X-rays) sufficient to resolve the object's internal structure. Because of the weak interaction of X-rays with matter, high coherence and photon flux are required to resolve the finest features in the object. Due to the extreme radiation dose, the biological particles are destroyed in the scattering process. To record a diffraction pattern corresponding to the undamaged structure, the X-ray pulse must have a duration shorter than the typical timescale of the destruction process. Therefore, high-brilliance synchrotron light sources could not be used due to insufficient coherent flux in a single pulse that is required for recording enough signal. The development of X-ray sources that have a high intensity and short pulse duration – X-ray free-electron lasers (XFELs) – overcome this challenge.

In the SPI method, many identical particles of the investigated system are injected into the X-ray beam providing diffraction images in random orientations. The three-dimensional structure of the object is obtained by applying complex algorithms to the collected diffraction patterns. The size of one such dataset could exceed terabytes; this motivates the development and implementation of elaborate data analysis techniques that help to save expensive XFEL time and speed up data processing.

The first two parts of this Thesis are based on the methodological development of the SPI data analysis workflow. The experimental data was collected from the virus PR772 at the Linac Coherent Light Source (LCLS) at SLAC, Stanford, USA in the frame of the SPI consortium. As a result of the developed methodology, which includes machine learning object classification, a three-dimensional virus structure with a resolution below 10 nanometers was reconstructed. The comparison of the result with the cryogenic microscopy studies showed similar features and an overall agreement between both techniques. Due to the complexity and cost of the SPI experiments, the preparation is a time- and effort-consuming process that requires high-level planning. The third part of this Thesis explores the optimization of set-up parameters through the simulation of the SPI experiment with tick-borne encephalitis virus. These simulations contributed to the success of an actual experiment performed at the European XFEL in Hamburg, Germany.

---

## Eidesstattliche Versicherung / Declaration on oath

*Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.*

*I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.*

Hamburg, den 28.11.2022

Dameli Assalauova

A handwritten signature in black ink, appearing to read 'DAmf', written in a cursive style.

*Unterschrift (Signature)*



# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
<b>2</b>	<b>X-ray sources</b>	<b>11</b>
2.1	Synchrotron sources . . . . .	14
2.2	X-ray free-electron lasers . . . . .	20
<b>3</b>	<b>X-ray interaction with matter</b>	<b>27</b>
3.1	X-ray scattering on a single electron . . . . .	28
3.2	X-ray scattering on an atom . . . . .	31
3.3	X-ray scattering on an isolated particle . . . . .	34
3.4	X-ray scattering on a crystal . . . . .	36
3.5	X-ray absorption . . . . .	39
3.6	Coherence . . . . .	41
<b>4</b>	<b>X-ray imaging techniques</b>	<b>45</b>
4.1	Overview . . . . .	45
4.2	Coherent X-ray Diffractive Imaging . . . . .	48
4.3	Iterative phase retrieval techniques . . . . .	54
4.3.1	Gerchberg-Saxton algorithm . . . . .	56
4.3.2	Error-Reduction . . . . .	57
4.3.3	Hybrid-Input-Output . . . . .	59
4.3.4	Continuous Hybrid-Input-Output . . . . .	60
4.3.5	Shrink-Wrap . . . . .	60
4.3.6	Solvent Flipping . . . . .	61
<b>5</b>	<b>Single Particle Imaging with XFELs</b>	<b>63</b>
5.1	SPI experiment . . . . .	64
5.2	Radiation damage . . . . .	66
5.3	Challenges and limitations . . . . .	68
5.4	Data analysis pipeline . . . . .	71
5.5	Data classification in SPI . . . . .	73



---

5.5.1	Principal Component Analysis . . . . .	74
5.5.2	K-means clustering . . . . .	77
5.5.3	Expectation-Maximization algorithm . . . . .	78
5.5.4	Convolutional Neural Networks . . . . .	79
5.6	Orientation determination of diffraction patterns . . . . .	82
5.7	Resolution estimation in CXDI and SPI . . . . .	83
5.7.1	Phase Retrieval Transfer Function . . . . .	83
5.7.2	Fourier Shell Correlation . . . . .	84
<b>6</b>	<b>An advanced workflow for SPI experiment with the limited data at an X-ray free-electron laser</b> . . . . .	<b>87</b>
6.1	Experiment . . . . .	87
6.2	Initial classification steps . . . . .	89
6.2.1	Hit finding . . . . .	89
6.2.2	Analysis of additional instrumental scattering . . . . .	90
6.2.3	Beam center position finding . . . . .	90
6.2.4	Particle size filtering . . . . .	91
6.3	Single hit diffraction patterns classification . . . . .	94
6.4	Orientation determination and background subtraction . . . . .	97
6.5	PR772 virus structure . . . . .	99
6.5.1	Phase retrieval . . . . .	99
6.5.2	Mode decomposition . . . . .	101
6.5.3	PR772 structure analysis . . . . .	102
6.5.4	Resolution estimation . . . . .	106
6.6	Summary . . . . .	106
<b>7</b>	<b>Classification of diffraction patterns using a Convolutional Neural Network in SPI experiments</b> . . . . .	<b>109</b>
7.1	CNN description and architecture . . . . .	111
7.2	Training, validation and test procedure in CNN classification . . . . .	112
7.2.1	Polynomial learning rate (polyLR) policy . . . . .	113
7.2.2	Data augmentation . . . . .	114
7.2.3	K-fold cross-validation . . . . .	116
7.2.4	Ensembling via softmax averaging . . . . .	116
7.2.5	Inference . . . . .	116
7.3	CNN variant: identifying more single hits . . . . .	117
7.4	Particle size determination . . . . .	117
7.5	Results . . . . .	117
7.5.1	CNN performance . . . . .	117

## CONTENTS

---

7.5.2	PSD comparison, EM and particle size filtering . . . . .	119
7.5.3	Intersection over union comparison . . . . .	123
7.5.4	Orientation determination . . . . .	124
7.5.5	Phase retrieval and reconstructions . . . . .	125
7.6	VGG-style network . . . . .	127
7.7	Summary . . . . .	128
<b>8</b>	<b>Simulation of SPI experiment with Tick-borne encephalitis virus</b>	<b>131</b>
8.1	Tick-borne encephalitis virus . . . . .	132
8.2	Data simulations for the SPI experiment . . . . .	133
8.3	Spatial structure of the TBEV from simulated data . . . . .	137
8.4	Summary . . . . .	138
<b>9</b>	<b>Summary</b>	<b>141</b>
	<b>Publications</b>	<b>143</b>
	<b>Bibliography</b>	<b>146</b>



# Chapter 1

## Introduction

The journey to the insight world of things around us began a long time ago with the development of the lenses and first microscopes. Archaeological evidence indicates that the lenses were used in antiquity spanning several millennia. Different references to using lenses in everyday day life also exist. Thus, the fifth Roman emperor Nero was believed to watch the gladiatorial games using an emerald because of his likely nearsightedness which dated to the 5th century BC. Human interest in exploring the possibilities of lenses has eventually led to the invention of optical microscopes which were using visible light's properties. It is hard to name the father of the first optical microscope but the idea of combining lenses in order to get the biggest optical zoom got its effective development in the 17th century. The word "microscope" itself was suggested by Giovanni Faber for the device made by Galileo Galilei and shown at the Accademia dei Lincei in 1625.

The people's thirst to know the world around them led to a boom in the development of science at the end of the 19th and in the 20th centuries. In 1873 Ernst Abbe found that visible light could not satisfy human curiosity to study objects with conventional visible light microscopes – as a resolution limit exists. It was expressed as

$$\Delta = \frac{\lambda}{2n \sin \theta}, \quad (1.1)$$

where  $\lambda$  is the light wavelength,  $n$  is the refractive index of the medium and light is covering the spot with the half-angle  $\theta$ . Abbe diffraction limit states resolution of hundreds of nanometers for visible light and relatively small (with the size of hundreds of nanometers) objects could be studied in such a way. At that time, the desire and need to study even smaller objects, as well as the interactions occurring at smaller distances, already existed. This resulted in searching for other ways to visualize small particles, biological cells, crystals and other materials. It was realized that in order to achieve higher resolution, electromagnetic waves with shorter wavelengths could be exploited.

In 1895 Wilhelm Röntgen discovered X-rays, and this launched an era of studying the interaction of X-rays with matter and their properties. X-ray's short wavelength and its properties of coherence would become the object of research for many decades ahead. The progress in the field opened up the possibility of studying the structure of matter and small objects with high resolution. Moreover, X-ray science became the base of crystallography. Over the years, X-ray techniques, such as X-ray diffraction, X-ray microscopy, coherent X-ray diffractive imaging, X-ray ptychography and many others, have taken their place among the most powerful tools in structural studies and have led to fundamental discoveries in many scientific disciplines.

Special attention to X-ray application has caught the possibility to study biological samples, such as small cells, viruses, proteins, and others. The significance of such studies is hard to overestimate. The pandemic of coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 virion of 60 – 140 nanometers in diameter has already influenced 500 million people all over the world (by the moment of writing this Thesis). Studying and understanding the working principles of viruses and other biological objects can potentially save lives and prevent future pandemics. And X-ray techniques can be used in this research. However, one of the challenges of the X-ray science is the damage of biological specimens from such strong radiation. The energy of X-ray photons is able to ionize atoms of the living tissue and destroy molecular bonds inside. It is harmful to the object and therefore, absorbed dose is an important aspect in biological studies with X-rays.

The radiation damage caused by photo absorption in X-ray imaging of biological particles was overcome with the development of the "diffraction-before-destruction" concept. If one is able to produce ultra-short X-ray pulses, after interaction with it the sample will be destroyed, but if the pulse duration is shorter than the destruction process, X-ray diffraction on such sample can be measured and recorded. This concept brought another challenge – are there devices that can produce femtosecond X-ray pulses? A typical X-ray source – synchrotron – is a good tool for X-ray applications but its pulse duration of dozens of picoseconds does not allow the implementation of the "diffraction-before-destruction" idea with biosamples.

The problem was solved with the development of X-ray free-electron lasers producing intense, femtosecond coherent X-ray pulses and conducting single particle imaging (SPI) experiments. SPI method makes it possible to determine the three-dimensional structure of the biological particles. For that, it requires the use of coherent in phase X-rays and high intensity which are provided by the XFELs. In SPI experiments, identical specimens of the biomolecule are injected into an X-ray beam in random orientations. Samples are destroyed, but due to the ultra-short pulse duration, the diffraction is measured and the pattern is recorded before the subatomic changes of the object become significant.

---

SPI experiment at XFELs is a very difficult project both scientifically and technically and requires careful and thoughtful planning. Throughout the years, the pipeline of experiment preparation and data analysis became well-known and established. After performing a successful SPI experiment that highly depends on a lot of factors, the obtained data has to be analyzed. The large amount of diffraction patterns collected at XFEL is processed and various approaches exist to make every step of the SPI workflow executed. The focus of this Thesis will be on the detailed description of the SPI data analysis pipeline and the implementation of new techniques and approaches which can influence the resolution of the final three-dimensional structure of the biological particle. Today the goal of atomic resolution with SPI has not been reached and the values are still in the range of 5 – 10 nanometers.

One of the possible ways to improve current results in SPI is to use modern machine learning methods which proved to be indispensable during the last decade. They infiltrated everyday life in computer vision, text and voice recognition areas, and are implemented practically everywhere – from smartphone cameras to advertising and custom support. Science, and in particular X-ray techniques, could also highly benefit from utilizing machine learning in data processing and analysis. XFEL provides petabytes of diffraction patterns which have to be classified into certain groups. The data can be evaluated with the help of machine learning algorithms. Examples of such applications, specifically maximum-likelihood methods and convolutional neural networks, will be presented.

SPI experiments based on XFEL operation open up possibilities and provide great potential in studying the structure of biomolecules with the possible subnanometer resolution, observing chemical reactions inside and creating molecular movies. Modern XFELs with megahertz repetition rate and ultimate brightness are the main tools to achieve these goals. Their further development, careful planning of experiments using preliminary simulations, implementation of new approaches in data analysis, and application of new machine learning algorithms – all together can take the progress of SPI experiments with biological particles to a new level.

This Thesis starts with a general description of the X-ray sources and the basics of their operation, starting from the X-ray tube and ending with the 4th generation synchrotron sources and powerful XFELs. Chapter 3 gives the foundation of X-ray interaction with matter emphasizing the scattering process on different objects: single electrons, atoms, isolated particles, and crystals. It includes also the process of X-ray absorption and covers the coherent properties of X-rays. Introduction to the application of modern X-ray imaging techniques is given in Chapter 4 and Chapter 5 where coherent X-ray diffractive imaging (CXDI) and single particle imaging (SPI) are described in detail as well as the accompanying challenges during the experiments and current limitations. In Chapter 4 and Chapter 5, the main features of the data analysis pipeline are given: solving the phase problem with the use of phase retrieval algorithms and the task of diffraction patterns classification with the

use of machine learning techniques. Chapter 6 is focused on thorough data analysis of the SPI experiment performed at Linac Coherent Light Source (LCLS) in Stanford, USA with the bacteriophage PR772. The following Chapter 7 describes the convolutional neural network application in the same SPI experiment. Chapter 8 gives the description of the planned SPI experiment with tick-borne encephalitis virus – the data simulations were made in order to obtain the preferable set-up parameters prior to the experiment. Finally, the Summary analyzes the obtained results and gives an outlook for further progress.

# Chapter 2

## X-ray sources

X-ray is a form of electromagnetic radiation with the wavelength in the region of Åströms ( $10^{-10}$  m) which are atomic scales. Found by Wilhelm Röntgen in 1895, the so-called "X-radiation" was able to penetrate through the books and papers [1]. The fact that X-rays could be the tool to look inside the human body and penetrate into the matter was considered a breakthrough and quickly spread over the world. It was researched and developed over the years and massively affected physics, biology, and technology. Alongside the study of X-ray interaction with matter, the development and improvement of X-ray sources were ongoing.

X-ray radiation differs from other types of radiation by the basic parameters. Its wavelength ranges from picometers to nanometers, the corresponding frequencies range from  $30 \times 10^{15}$  Hz to  $30 \times 10^{18}$  Hz and energies range from 145 eV to 124 keV. Typically, the soft X-ray region is considered between 250 eV to 10 keV, the hard X-ray regime is from 10 keV to 100 keV. Behind the process of generating X-rays, several concepts lie. First, bremsstrahlung radiation: the moving electrons are decelerating when deflected by an atomic nucleus and in this case, the continuous spectrum of X-ray radiation is produced. The second phenomenon is characteristic radiation emitted due to the electron collision induced ionization. The vacancies of the inner shell of the atom from the leaving electrons are filled by the electrons from the higher shells. The energy of the emitted X-ray photon during this transition is equal to the energy difference between the states which is called fluorescent radiation. These two processes are used in X-ray tubes and the spectrum from such devices is shown in Fig. 2.1. In X-ray tubes, the cathode emits electrons and they are then collected on the anode. Between the cathode and anode, the electrons are accelerated by the external power source. The X-ray spectrum depends on the anode material and accelerating voltage; and cooling efficiency determines the limitation of the radiation intensity.

One of the important applications of X-ray tubes of the 20th century was the study of crystalline matter when Max von Laue and his colleagues obtained the first diffraction pattern from the crystal [2]. Later Lawrence Bragg and his father William Henry Bragg studied



a number of crystals [3] and started the field of crystallography which aimed at molecular structure determination.

But since X-ray tubes lack photon intensity and easy tuning for desired wavelengths, the development of other X-ray sources continued. Modern synchrotron sources and X-ray free-electron lasers (XFELs) have driven X-ray science to a whole new level. They are based on the new conceptual approach designed for high efficiency radiation process and produce a very intense and coherent X-ray beam.

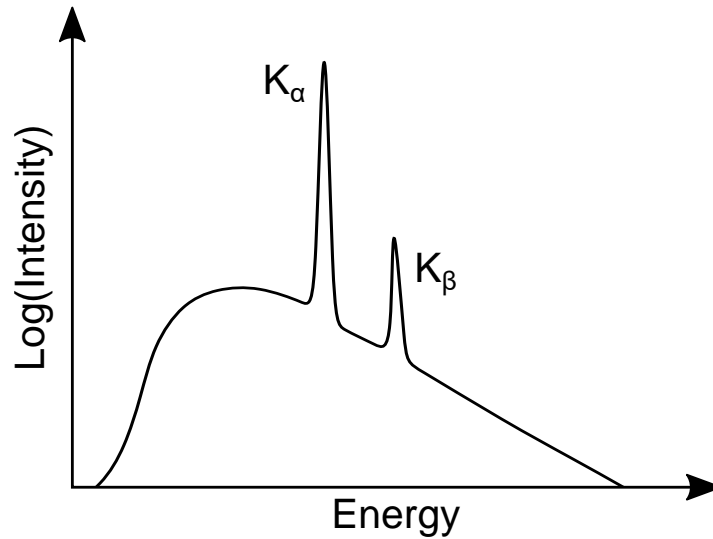


Figure 2.1: Continuous radiation (bremsstrahlung) and characteristic line emission happen in X-ray tubes when cathode electrons hit the anode.  $K_\alpha$  and  $K_\beta$  represent the transitions between an L and K shell, and M and K shell, respectively. The concept was adapted from [4].

One of the characteristics that can describe the intensity of the X-ray radiation from the source is the photon flux which indicates the number of emitted photons per second in 0.1 % of the radiation bandwidth

$$F(\lambda) = \frac{\text{photons/sec}}{0.1\% \text{ bandwidth}} \quad (2.1)$$

where  $\lambda$  is the wavelength of X-ray radiation.

To characterize the X-ray source, spectral brightness is used which is defined as

$$B(\lambda) = \frac{F(\lambda)}{\Delta A \cdot \Delta \Omega} \quad (2.2)$$

which shows the spectral photon flux  $F(\lambda)$  radiated per unit projected area  $\Delta A$  per unit solid angle  $\Delta \Omega$ .

The evolution of different sources of X-ray radiation as a function of time is shown in Fig. 2.2. The brightness of X-ray sources varies a lot: from  $10^7 - 10^{12}$  ph/(s·mm<sup>2</sup>·mrad<sup>2</sup>·0.1%) for X-ray tubes to  $10^{23}$  ph/(s·mm<sup>2</sup>·mrad<sup>2</sup>·0.1%) of the synchrotron radiation of modern storage rings (see Fig 2.3 (a)). XFELs (see Fig. 2.3 (b)) can produce the brightest X-ray radiation of

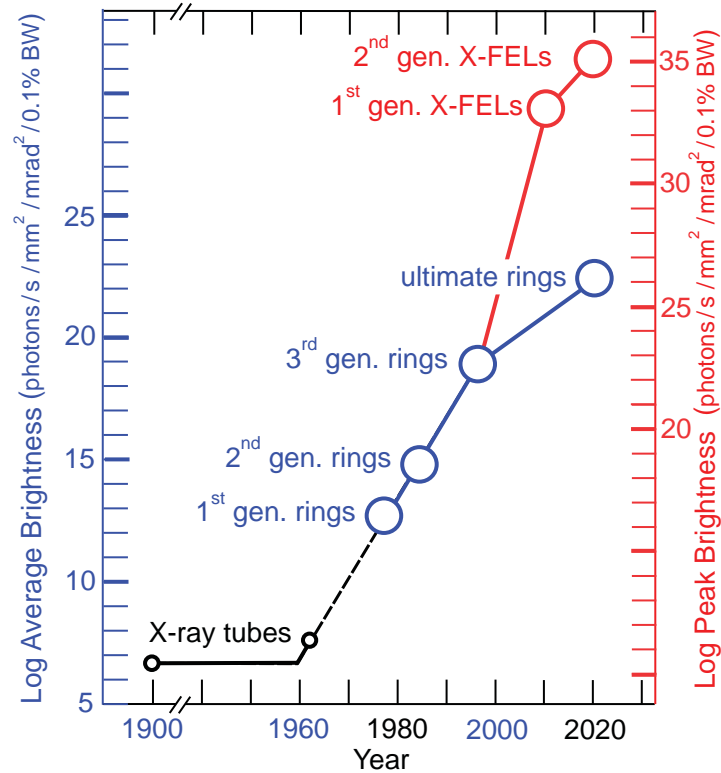


Figure 2.2: The brightness of different X-ray sources as a function of time. The development of ultra-bright XFELs (red) started at the beginning of the 21st century and actively continues nowadays. Adapted from [5].

$10^{35}$  ph/(s·mm<sup>2</sup>·mrad<sup>2</sup>·0.1%) which makes them the current brightest X-ray sources while ultimate storage rings are mostly under construction. Their properties and principles of work will be discussed in the following Sections.

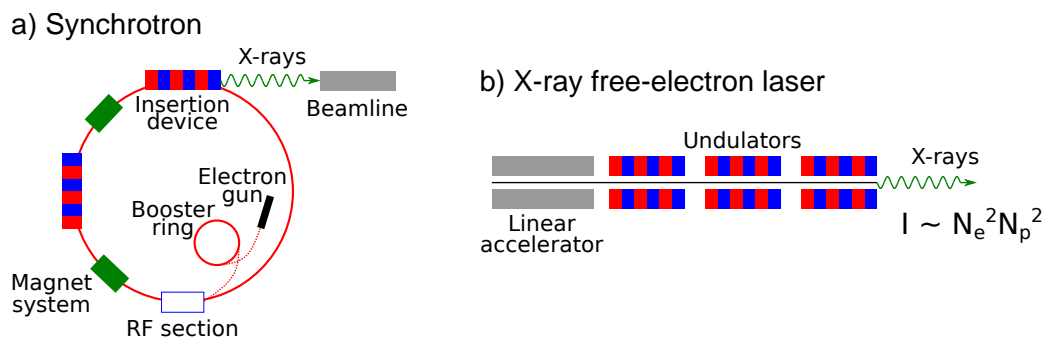


Figure 2.3: X-ray radiation sources: (a) storage ring of the synchrotron and (b) XFEL. (a) Storage rings' main components: an electron gun, radiofrequency (RF) section, magnet systems, insertion devices, and beamlines. (b) XFEL's main components are linear accelerator and undulators.

## 2.1 Synchrotron sources

A typical and well-known source of X-ray radiation is a synchrotron. It includes an electron gun, a booster ring, a storage ring, a radiofrequency section, magnet systems, and beamlines (see Fig. 2.3 (a)). The electrons used in producing synchrotron radiation are coming from the electron gun to the booster ring in order to increase the electron energy. From the booster ring electrons are periodically injected into the storage ring which is a vacuum tube where electrons with velocities close to the speed of light are cycling and producing X-rays when passing special magnetic systems. Special insertion devices (which will be described further) are responsible for the coherent characteristics and brightness of the photon beam. The electrons are losing energy during their path in the storage ring, so the radiofrequency (RF) system's goal is to recover it. In this system, the electric field oscillates with radio frequencies (up to 1 GHz). Electrons in the storage ring travel through magnet systems which usually consist of bending magnets, quadrupole magnets, and sextupole magnets which ensure the stable trajectory of the electrons. Produced X-ray radiation is then used at experimental stations – beamlines.

The development of synchrotron facilities started with the 1st generation synchrotron sources in the 1940s and produced X-ray radiation was considered parasitic, the main aim of such facilities was the study of high-energy and nuclear physics. Synchrotron radiation took place in the bending magnets holding the electrons in the accelerating ring and changing their trajectories (Fig. 2.4 (a)).

Interest in storage rings as a dedicated source of synchrotron radiation followed. The first storage ring which provides X-rays in the GeV range to a community of users was the 2.5 GeV SPEAR ring (Stanford Positron Electron Accelerating Ring) at the Stanford Linear Accelerator Center (SLAC). From 1974 to 1992 another storage ring DORIS (Doppel-Ring-Speicher) at the Deutsches Elektronen-Synchrotron (DESY) laboratory was used in particle physics and research with synchrotron radiation. VEPP-3 electron-positron storage ring at the Institute for Nuclear Physics in Novosibirsk, built in 1967-1971, CESR (Cornell Electron Storage Ring) at Cornell (now known as Cornell High Energy Synchrotron Source (CHESS) facility) completed in 1979, were also early updated with synchrotron radiation capabilities.

In the 1980s the development of 2nd generation synchrotron sources started. These devices were meant to produce stable X-rays and were using storage rings to maintain the kinetic energy of electrons after acceleration. They also contained the key devices which would be used in the next generation sources, such as an electron gun, a booster ring, a storage ring, bending magnets and beamlines. The principle of the operation formed the basis of the work of all modern synchrotron sources and is the following: a linear accelerator (linac) makes the electrons accelerate, further acceleration happens in the booster ring, and relativistic electrons are then injected into the storage ring. In the section of the bend-

ing magnets, X-rays are produced and electrons are kept to travel on a curved trajectory of the storage ring. X-rays end up in the beamline sections where they are focused and enter the experimental hutches. Famous examples of the 2nd generation synchrotron sources are the Stanford Synchrotron Radiation Laboratory at SLAC and HASYLAB (Hamburger Synchrotronstrahlungslabor) at DESY.

X-rays produced in the bending magnets of the 2nd generation sources had quite a broad spectrum which lead to the further development of synchrotron sources aimed to produce brighter radiation and reduce beam size and divergence not at the expense of the photon flux. This demand was made, for example, by crystallography, where it is necessary that the incident beam corresponds to the crystal size and to have an acceptable angular resolution to resolve diffraction peaks. These requirements of the increased coherence and brightness were partially satisfied in 3rd generation synchrotron sources that are used nowadays.

Examples of the modern high energy 3rd generation synchrotron sources are ESRF, Grenoble, France (6 GeV storage ring), APS, Chicago, USA (7 GeV storage ring), SPring-8, Sayo, Japan (8 GeV storage ring) and PETRA III, Hamburg, Germany (6 GeV storage ring). These sources are based on the insertion devices such as undulators and wigglers, and magnetic lattice which are able to reduce electron beam emittance  $\varepsilon_e = \sigma_e \theta_e$ , where  $\sigma_e$  is the beam size and  $\theta_e$  is its divergence. It directly influences the X-ray photon emittance, so the electron beam emittance is considered one of the main properties of the synchrotron sources. Spectral brightness in Eq. (2.2) can be rewritten in terms of photon horizontal  $\varepsilon_x$  and vertical emittance  $\varepsilon_y$  as

$$B(\lambda) = \frac{F(\lambda)}{(2\pi)^2 \varepsilon_x \varepsilon_y}, \quad (2.3)$$

Below we will discuss the principles of producing of X-ray radiation in bending magnets, wigglers, and undulators.

### Bending magnets

Bending magnets' radiation is based on relativistic electron acceleration. The electron is traveling around a circle and has a radial acceleration, in this way it emits radiation through a certain angle (see Fig. 2.4 (a)). Due to the relativistic effects and Lorenz transformation, the angle in the direction of the motion of the electron is

$$\tan(\theta) = \frac{\sin(\theta')}{\gamma(\beta + \cos(\theta'))}, \quad (2.4)$$

where  $\theta'$  is the angle in the frame of a moving electron,  $\theta$  is the angle in the laboratory frame,  $\beta = v/c$ ,  $v$  is the electron velocity,  $c$  is the speed of light,  $\gamma$  is the Lorentz factor defined as

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}. \quad (2.5)$$

Considering relativistic electron, we have  $\beta \approx 1$  and  $\gamma \gg 1$ . So for large  $\theta'$ , the electron emits radiation in a narrow cone tangent to the path of the movement of half-angle

$$\theta \simeq \frac{1}{2\gamma}. \quad (2.6)$$

The spectral distribution of the synchrotron radiation of the bending magnet is shown in Fig. 2.4 (a). It is quite broad and can be characterized by the critical wavelength  $\lambda_c$  – at wavelengths below this value, the intensity of the synchrotron radiation drops sharply. It is defined by the machine parameters

$$\lambda_c[\text{\AA}] = \frac{5.59 \cdot R[m]}{E^3[\text{GeV}]}, \quad (2.7)$$

where  $R$  is the radius of curvature of the electron trajectory and  $E$  is the electron energy. Corresponding critical energy  $\varepsilon_c$  can be expressed as

$$\varepsilon_c[\text{keV}] = \frac{2.218 \cdot E^3[\text{GeV}]}{R[m]}. \quad (2.8)$$

## Wigglers and undulators

As mentioned above, special insertion devices were developed for synchrotron sources to obtain x-ray radiation efficiently. Wigglers and undulators (see Fig. 2.4 (b) and (c)) are the periodic structures of the dipole magnets which can generate the brightest X-rays. Relativistic electrons are going through such a magnet structure and are compelled to the sinusoidal motion caused by alternating magnetic fields.

The wavelength  $\lambda$  of the radiation after the electron travels through the undulator is much smaller than the magnet period  $\lambda_u$  [6]. As said before, an electron is experiencing sinusoidal motions and therefore emits radiation. In this way, we can consider the electron as a radiating dipole in the frame moving with the average speed of the electron. Considering the velocity of the electron close to the speed of light, the relativistic Doppler effect takes place and the emission wavelength is reduced. Using small-angle approximation ( $\theta \neq 0$ )

## 2.1. Synchrotron sources

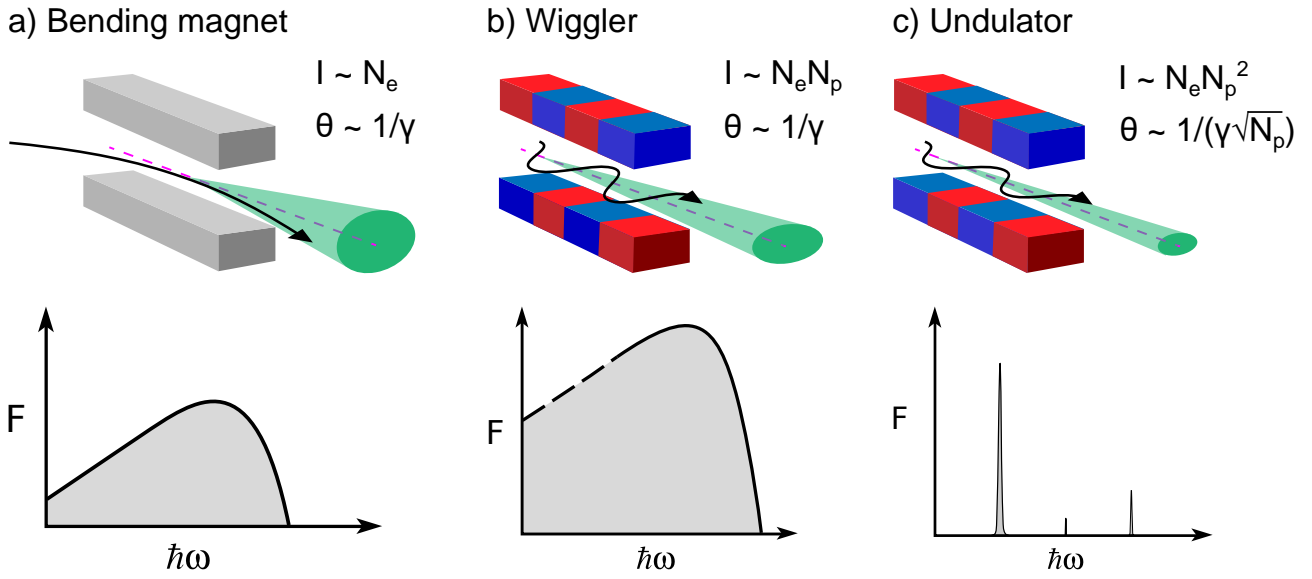


Figure 2.4: The magnetic devices producing the X-rays: (a) bending magnet, (b) wiggler, and (c) undulator. The electron path is shown with a black arrow and an X-ray beam with a green cone. For each device, the intensity  $I$  of X-rays and opening angle  $\theta$  are shown.  $N_p$  is the number of dipole magnets in the device and  $N_e$  is the number of electrons. The radiation spectrum for each device is shown in the bottom row. Adapted from [6].

the fundamental X-ray (observed) wavelength  $\lambda$  is

$$\lambda = \frac{\lambda_u}{2\gamma^2} \left( 1 + \gamma^2 \theta^2 + \frac{K^2}{2} \right), \quad (2.9)$$

where  $\lambda_u$  is the magnet period and  $K$  is the so-called undulator parameter

$$K = \frac{eB_0\lambda_u}{2\pi m_e c} = 0.934 B_0 [T] \lambda_u [cm], \quad (2.10)$$

where  $e$  is the elementary charge,  $B_0$  is the magnetic field of the undulator magnets, and  $m_e$  is the electron rest mass. From Eq. (2.10) it is seen that  $K$  can be changed by changing the magnetic period  $\lambda_u$  or by changing the magnetic field  $B_0$ .

The undulator parameter  $K$  is used to distinguish between a wiggler and an undulator. If the magnetic field  $B_0$  is strong,  $K \gg 1$ , the wiggler radiation is considered. In this type of device, the electron is moving with a high oscillation amplitude (see Fig. 2.4 (b)) and thus a broader spectrum. The brightness of such radiation is proportional to the number of dipole magnets in wiggler  $N_p$  and the number of electrons  $N_e$ :  $I \sim N_p N_e$ . And the emission is produced into the opening angle  $\theta \sim 1/\gamma$ . But due to the broader spectrum, the brightness of such radiation is not very high. In the case of wiggler, the spectral distribution looks similar to the bending magnet's distribution but is characterized by a much larger photon flux and a shift to higher energies (or shorter wavelengths).

If we decrease the magnetic field  $B_0$ ,  $K \leq 1$ , the undulator radiation is considered. In this case, the electron oscillates with a smaller amplitude which leads to a much narrower cone of radiation (see Fig. 2.4 (c)). In addition, the electromagnetic wave emitted from the electron constructively interferes with the wave emitted from the previous turn of the particle. As a result, the undulator radiation intensity is proportional to the square number of dipole magnets in undulator  $N_p$ :  $I \sim N_p^2 N_e$ . The resulting opening angle in the case of undulator can be described as  $\theta \sim 1/(\gamma\sqrt{N_p})$ . Consequently, undulators allow obtaining well-separated and narrow spectral peaks of higher intensity.

In addition, the selection of the highly monochromatic X-rays or narrow spectral bandwidth  $\Delta\lambda/\lambda$  (see Fig. 2.4 (c)) can be obtained by using an undulator. This can be done with the beam-defining slits after undulators. They allow the central part of the beam to go through as the radiation closer to the axis is preferable. At the same time, higher harmonic oscillations defined by the number  $n$  of harmonics occur. In this case, the harmonic wavelength  $\lambda_n$  and the spectral bandwidth decrease

$$\lambda_n = \frac{\lambda}{n}, \quad (2.11)$$

$$\frac{\Delta\lambda}{\lambda} = \frac{1}{N_p n}. \quad (2.12)$$

So if one wants to achieve shorter wavelengths, one may consider using higher harmonics.

Table 2.1: Third generation sources parameters (for the photon energy  $E_{ph} = 500$  eV). Here  $E_e$  is the electron energy,  $I$  is the electron bunch charge,  $\varepsilon_{x,y}$  are the horizontal and vertical emittance, respectively, and  $B$  is the resulting brightness.

Source	APS	ESRF	PETRA III
$E_e$ , GeV	7	6	6
$I$ , mA	100	200	100
$\varepsilon_x$ , nm rad	2.5	1.5	1.2
$\varepsilon_y$ , pm rad	40	5	10
$B$ , ph/(s·mm <sup>2</sup> ·mrad <sup>2</sup> ·0.1%)	$6 \times 10^{18}$	$1.2 \times 10^{20}$	$3.7 \times 10^{19}$

Usage of the insertion devices such as wigglers and undulators made the 3rd generation synchrotron sources useful tools to explore the benefits of more coherent and intense X-rays. The brightness of such sources reached the values of  $10^{21}$  ph/(s·mm<sup>2</sup>·mrad<sup>2</sup>·0.1%) (see Fig. 2.2) and photon horizontal and vertical emittance reaches relatively small values of nm rad and pm rad, respectively. Parameters of some of the 3rd generation synchrotron sources are shown in Table 2.1.

Today's 3rd generation synchrotron sources have a large amount of beamlines covering almost all areas of X-ray applications: Nuclear Resonant Scattering (NRS), Inelastic

## 2.1. Synchrotron sources

X-ray scattering (IXS), powder X-ray diffraction, diffraction experiments at extreme conditions of high pressure and simultaneous high- or low-temperature, small- and wide-angle X-ray scattering (SAXS/WAXS), micro- and nanotomography, X-ray fluorescence (XRF), X-ray absorption spectroscopy (XAS), X-ray diffraction (XRD), X-ray Photon Correlation Spectroscopy (XPCS); coherent diffractive imaging of micro- and nanostructures (CDI); time-resolved SAXS studies of complex liquids (Rheo-SAXS), biological small-angle scattering (BioSAXS), macromolecular X-ray crystallography, hard X-ray photoelectron spectroscopy (HAXPES), grazing incidence diffraction [7]. Chapter 4 will be dedicated to the specifics of imaging X-ray techniques.

Despite the significant improvements of 3rd generation sources, there is still room for enhancement. For example, about 1% of the synchrotron beam is sufficiently coherent which could be not enough in experiments that require a high degree of coherence.

Nowadays the synchrotron sources which are brighter (see Fig. 2.5 for PETRA III and PETRA IV comparison) and produce more coherent X-ray radiation (up to 90%) – the 4th generation synchrotron sources – are constructed. As it is seen from Eq. (2.3), to achieve brighter source one has to lower photon emittance  $\varepsilon_{x,y}$ . The 4th generation synchrotron sources are designed to produce low emittance in both horizontal and vertical directions for a wide range of X-ray wavelengths. It is done by using so-called multi-bend achromat devices [8] for better electron beam focus and control. The concept of a multi-bend achromat system is based on maximizing the number of magnets in the storage ring.

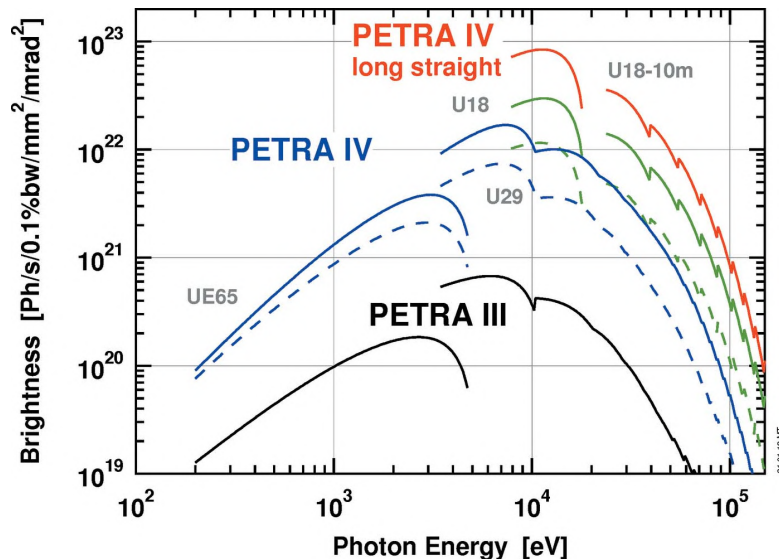


Figure 2.5: The brightness of PETRA III (3rd generation synchrotron) and PETRA IV (4th generation synchrotron) for a ring current of 100 mA depending on the photon energy. Adapted from [9].

The diffraction-limited source is another term which describes the 4th generation synchrotron sources. This principle is based on the photon beam emittance reaching the minimum value. The diffraction-limited beam size and divergence can be calculated from the



Heisenberg's uncertainty principle for the photon  $\Delta x \Delta p_x \geq \hbar/2$  for its root mean square (rms) values [10]. Considering small divergence angles and de Broglie's relation for the photon momentum  $p = \hbar k$ , we have in the transverse direction:  $\Delta x \theta_x \geq 1/(2k)$ . Using rms width  $\sigma_x$ , in two dimensions we have

$$\sigma_x \theta_x \sigma_y \theta_y \geq \left( \frac{\lambda}{4\pi} \right)^2. \quad (2.13)$$

The minimum value of the photon beam emittance is  $\lambda/4\pi$  [11]. A source is called diffraction-limited if the emittance of the electron beam is smaller than the emittance of the photon beam

$$\varepsilon_e^x < \varepsilon_p^x = \frac{\lambda}{4\pi}, \quad (2.14)$$

This limitation in Eq. (2.14) is correct for the Gaussian beam. In the case when X-ray radiation can not be approximated as Gaussian radiation, the lowest value of the photon emittance is reaching  $\lambda/(2\pi)$  [12].

Table 2.2: 4th generation sources parameters, where  $E_e$  is the electron energy,  $I$  is the electron bunch charge,  $\varepsilon_{x,y}$  are the horizontal and vertical electron beam emittance respectively, and  $B$  is the resulting brightness. The parameters of the synchrotron sources were taken from [9, 13–16].

Source	MAX IV	ESRF-EBS	APS-U	PETRA IV
$E_e$ , GeV	3	6	6	6
$I$ , mA	500	200	200	200
$\varepsilon_x$ , pm rad	200 – 330	120 – 30	42 – 32	10 – 30
$\varepsilon_y$ , pm rad	2 – 8	5 – 30	4 – 32	4 – 10
$B$ , ph/(s·mm <sup>2</sup> ·mrad <sup>2</sup> ·0.1%)	$4 \times 10^{21}$	$10^{22}$	$2 \times 10^{22}$	$10^{23}$

Nowadays the 4th generation synchrotron sources are operating and are using multi-bend achromat systems, some of them are still under construction or planned. Basic parameters are shown in Table 2.2. The world's lowest emittance storage ring – PETRA IV (Fig. 2.5) – a future 3D X-ray microscope, will go into operation in 2027. It is planning to extend the X-ray applications to all length scales and will allow research groups to study processes inside a catalyst, batteries, microchips under realistic conditions, and materials made of nanostructures with the highest spatial resolution by focusing the synchrotron radiation on the smallest spot [17].

## 2.2 X-ray free-electron lasers

While most of the 4th generation synchrotron sources are still under construction, other sources, X-ray free-electron lasers (XFELs), with high brightness (see Fig. 2.2) are being run

## 2.2. X-ray free-electron lasers

---

for the last decades. The story of XFELs began in 1971 with Madey showing that the emission radiation could be effectively amplified using periodic magnetic structures [18]. Based on this principle, in 1977 Deacon [19] reported about the first operation of a free-electron laser oscillator which generated coherent radiation in infrared, visible, and ultraviolet regions. In the 1980s the idea of XFEL, a source of highly coherent X-ray radiation, was formulated in Ref. [20–22].

An undulator is the source of x-ray radiation at XFELs. The main constructional difference is that in XFEL, a linear electron accelerator with long undulator paths is used instead of a circular storage ring (see Fig. 2.3 (b)). The typical length of the undulator track in XFEL is 30 – 100 m. Interaction of the undulator radiation and the relativistic electrons results in their “bunching” on the scale of the size of the radiation wavelength. When this occurs, random electron motion within the bunch becomes a well-correlated, electron wave is moving in phase with the X-rays. This process is called self-amplified spontaneous emission (SASE) and it causes an exponential growth of X-ray beam power and production of X-ray pulses with duration down to approximate femtoseconds (in comparison to synchrotron source electrons bunch which are of about 100 picoseconds).

The power gain resulting from electrons traveling in the straight long path of undulators and their interaction with photons is shown in Fig. 2.6 and can be described in three major phases. The first phase – spontaneous radiation – is the result of the chaotic movement of electrons. Then electrons are traveling further in the undulator and start to interact with the emitted radiation, both with almost the same axial velocities  $v_e$  and  $c$  for electrons and radiated wave respectively, where  $v_e/c = 1 - (1 + K^2/2)/2\gamma^2$ . In one period of the magnet system, electrons complete one period of its sinusoidal motion. Meanwhile, the radiated wave moves one wavelength more in comparison to the electron, thus the electron is behind one cycle of phase with respect to the wave. This is called “slip condition” which indicates where constructive interference is still possible. Under this condition, coherent summation of the fields and electron wave modulation occurs which results in modified electron trajectories and “microbunching”. The separation between the microbunches is equal to the X-ray wavelength. As the bunch propagates through a long undulator path, this modulation increases and so constructively summed electric fields. This process results in the exponential growth of the power of the radiated field. The latter is proportional to  $N_e$ , a number of electrons participating in the X-ray radiation emission, and radiated power also grows with  $N_e^2$ . This regime of the operation is called a linear regime and the radiated field can be described by Gaussian statistics.

At the saturation regime, when the microbunches are formed, the radiated power rises up to terawatts with femtosecond X-ray pulses with a high degree of spatial coherence compared to synchrotron sources. After a long path of the undulators, the X-ray beam reaches the beamline experimental station. European XFEL [23], for example, contains three SASE

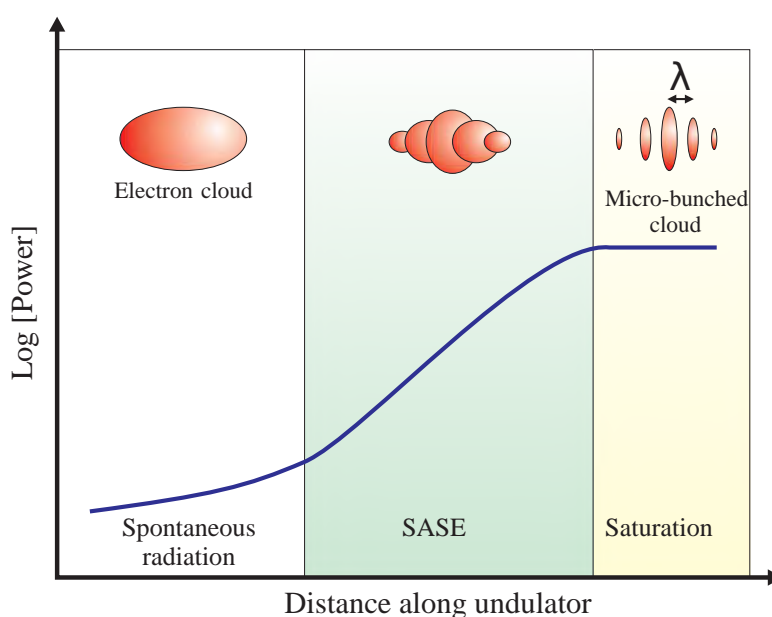


Figure 2.6: Dependence of the radiated power on the distance along the undulators – three major phases of XFEL X-rays generation. Adapted from [4].

undulators: SASE1, SASE2, and SASE3 (see Fig. 2.7). X-ray radiation ends up in six possible beamline stations: Materials Imaging and Dynamics (MID) and High Energy Density Science (HED) at SASE2; Single Particles, Clusters, and Biomolecules and Serial Femtosecond Crystallography (SPB/SFX) and Femtosecond X-ray Experiments (FXE) at SASE1; Small Quantum Systems (SQS) and Spectroscopy & Coherent Scattering (SCS) at SASE3.

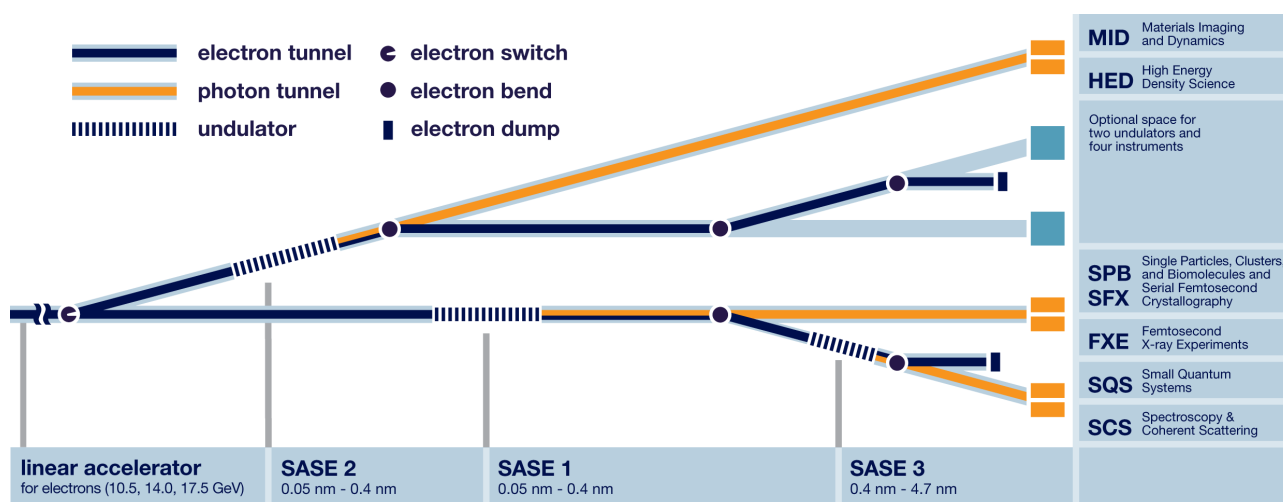


Figure 2.7: Undulators of the European XFEL. Typical wavelengths and photon energies are shown for each SASE. Adapted from [24].

The varying spectral modes can be amplified during electrons passing through the undulator. Spontaneous emission at the beginning of SASE (see Fig. 2.6) can lead to the appearance of different spectral modes, thus the overall spectrum of the final X-ray beam can be quite broad. So methods to narrow the spectrum were developed. One way to do that is

called "seeding" by using an external high-power optical laser. This worked well for EUV wavelengths. For X-ray wavelengths, another technique is used, called "self-seeding" which is a form of spectral filtering.

Seeding implies the external signal before the undulator. It should arrive simultaneously with the electron bunch, they both have to be similar in pulse duration and repetition rate. Besides, the electric field of the seed laser should prevail over spontaneous emission which starts the SASE. The seeding process is usually based on the use of an intense harmonic of a femtosecond atomic laser. As was mentioned before, it highly depends on the wavelength. For the EUV range, it was successfully used in the 2000s [25], where a single EUV harmonic was selected and focused with an electron bunch of FEL.

For soft X-rays, a similar technique is used, called a high gain high harmonic generation (HG HG). In addition to the coherent seeding, further frequency multiplication is used by the generation of high harmonics [26]. The success of this technique was demonstrated at the FERMI (Triest, Italy) with EUV/soft X-ray FEL down to the 4.0 nm wavelength. The process starts with the external coherent laser light (260 nm at FERMI [27]), which modulates the electron beam in the first part of the undulator. The energy modulation grows into a current modulation by a dispersion magnet. Then FEL pulse is tuned to one of the higher harmonics in the radiator section and later the electron bunching occurs. The bunched beam radiates at a harmonic wavelength. It is possible to use the HG HG cascade technique to extend the harmonic process.

For hard X-rays, external lasers' power is not enough, so self-seeding is used [28–31]. It is based on the separation of the generated FEL radiation and the electron beam and using the radiation itself as a seed (see Fig. 2.8). It begins with the SASE radiation generated by the first part of the undulator. Then the radiation pulse and the electron beam are separated: previous electron beam density modulation is destroyed, meanwhile, the radiation is spectrally filtered with preserved properties of the incoming beam and is used as a seed. Then they are rejoining and the SASE process continues. To narrow down the bandwidth of the radiation pulse different techniques are used, in particular, the grating monochromator (which is used for the soft X-rays), four-crystal monochromator, and single-crystal monochromator (which is used for hard X-rays). However, it is hard to realize and align later with the electron beam. Recently it was proposed to use the monochromator of a single crystal in Bragg-transmission geometry [29] which is a simpler scheme. A radiation pulse with a bandwidth much narrower than the initial FEL bandwidth is then used as a seed. The delay between the radiation pulse going through the monochromator device and the electron beam is compensated with the electron beam passing through a dispersive element, such as a magnetic chicane where electron microbunching is destroyed. Then photon pulse of the narrow spectrum rejoins with the electron bunch in the second part of the undulator, and coherent amplification happens. Self-seeding was successfully demonstrated at LCLS

with soft and hard X-rays [30, 32]. It was shown that the SASE signal had a broad spectral bandwidth (around or less than 1%) for a single pulse, and the self-seeded signal had a spectral bandwidth of less than 0.01%.

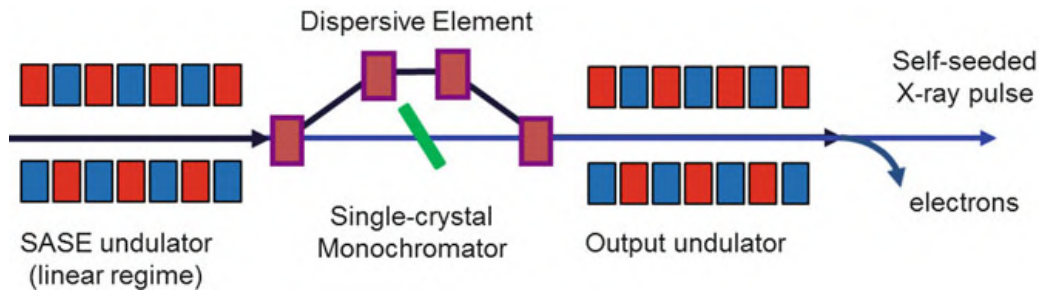


Figure 2.8: Self-seeding scheme done with the single crystal. Courtesy of G. Geloni.

The list of some operating XFELs and their parameters are presented in Table 2.3. European XFEL can be highlighted among all other operating XFEL due to its high repetition rate. It can produce X-ray pulse trains with a repetition rate of 10 Hz with 2700 X-ray pulses in each of them. This gives the ultimate 4.5 MHz total repetition rate, it can be used in a variety of applications which will be discussed further.

Table 2.3: XFEL parameters. Here  $E_e$  is the electron energy,  $E_{ph}$  is the photon energy,  $\Delta\omega/\omega$  is the spectral bandwidth,  $\Delta\tau$  is the pulse length. Adapted from [6].

	FLASH	FERMI	LCLS	PAL XFEL	EuXFEL
Mode	SASE	Seeded	SASE	SASE	SASE
$E_e$ , GeV	1.25	1.5	13.6	10	17.5
$E_{ph}$ , keV	0.03 – 0.3	0.02 – 0.3	0.25 – 10	2 – 12	1 – 25
$\Delta\omega/\omega$	$10^{-2}$	$10^{-3}$	$10^{-2}$	$10^{-2}$	$10^{-3}$
$\Delta\tau$ , fs	50	85	10 – 70	5 – 60	10 – 100
Photons/pulse	$3 \times 10^{12}$	$5 \times 10^{12}$	$2 \times 10^{12}$	$5 \times 10^{11}$	$10^{12}$
Light flashes/sec	10	10	120	60	27,000

The basic parameters which describe XFELs are: the peak radiated power  $\hat{P}$ , FEL parameter  $\rho_{FEL}$  and gain length  $L_G$ . The FEL parameter  $\rho_{FEL}$  [21] describes the energy transfer efficiency between electrons and photons [6]. The estimation of the peak radiated power  $\hat{P}$  from an XFEL is given by an exponential function [6]

$$\hat{P} \sim \exp\left(\frac{z}{L_G}\right), \quad (2.15)$$

where gain length  $L_G$  is the distance the wave should travel to have an exponential power gain

$$L_G = \frac{\lambda_u}{4\sqrt{3}\rho_{FEL}} \quad (2.16)$$

in the idealized one-dimensional case. In these terms, the peak power at saturation can be approximated as

$$\hat{P}_{sat} \simeq \rho_{FEL} \hat{P}_e, \quad (2.17)$$

where  $\hat{P}_e$  is the peak electron beam power. The spectral bandwidth at saturation is  $\Delta\omega/\omega \simeq 2.35\rho_{FEL}$ .

Such properties of the XFELs as spectral, spatial, and temporal coherence of the pulses, their femtosecond characteristics, and the high radiated power, opened new possibilities in X-ray science. The goal of studying femtosecond electron dynamics, biomolecular imaging [33, 34], single particle imaging (SPI) [35–38], pump-probe experiments [39, 40] was achieved with the usage of XFELs.



# Chapter 3

## X-ray interaction with matter

X-rays are the form of electromagnetic waves of shorter wavelengths and higher energy than normal light. They can be described as a form of electromagnetic waves propagating at the speed of light. Here we will consider electric field  $\mathbf{E}$  and magnetic field  $\mathbf{H}$ . This wave can be described with the wavelength  $\lambda$  and the wavenumber  $k = 2\pi/\lambda$ . We consider the electric field as the linearly polarized electromagnetic plane wave. Here we will take into account spatial and temporal propagation and polarization with the unit vector  $\hat{\mathbf{e}}$ , the electric field can be written in the following form

$$\mathbf{E}(\mathbf{r}, t) = \hat{\mathbf{e}}E_0e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \quad (3.1)$$

where  $\mathbf{k}$  is the wavevector along the propagation and  $\hat{\mathbf{e}} \cdot \mathbf{k} = 0$  and  $\mathbf{k} \cdot \mathbf{E} = \mathbf{k} \cdot \mathbf{H} = 0$  which is illustrated in Fig 3.1.

In terms of the particle nature of the wave, it can be quantized into photons with the energy  $\hbar\omega$  and momentum  $\hbar\mathbf{k}$ . The photon energy  $E$  and the wavelength  $\lambda$  are related according to the equation

$$\lambda[\text{\AA}] = \frac{hc}{E} = \frac{12.398}{E[\text{keV}]}. \quad (3.2)$$

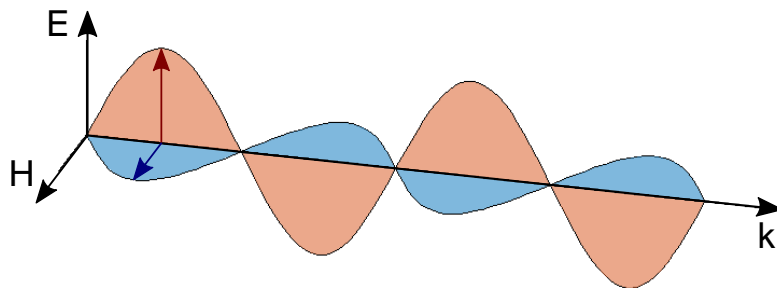


Figure 3.1: An X-ray electromagnetic wave with the electric field  $\mathbf{E}$  perpendicular to the magnetic field  $\mathbf{H}$ . Red arrow shows the polarization unit vector  $\hat{\mathbf{e}}$ . The concept is adapted from [4].



### 3.1 X-ray scattering on a single electron

We will start with the description of the elastic scattering of electromagnetic radiation by charged particles. The electric and magnetic fields of the incident wave accelerate the charged particle and the accelerated charged particle radiates electromagnetic waves. Thus, the energy of the incident wave is converted into the energy of the scattered wave. We will see that the scattering cross-section does not depend on the frequency of the electromagnetic wave and is the same in forward and backward directions. The frequency of the scattered radiation is the same as the frequency of the incident radiation. This type of scattering was explained by the English physicist J. J. Thomson.

In this Section, we will consider the X-ray scattering on a free electron. With the assumption that the dimensions of the system are smaller than the length of emitted wave or in other words, that the velocity of the electron is smaller than the speed of light  $v \ll c$ , we can describe radiation from the electron as dipole radiation. The electric field produced by the oscillating dipole in a vacuum is [41]

$$\mathbf{E} = \frac{3\mathbf{n}(\mathbf{n} \cdot \mathbf{d}) - \mathbf{d}}{R^3} + \frac{3\mathbf{n}(\mathbf{n} \cdot \dot{\mathbf{d}}) - \dot{\mathbf{d}}}{cR^2} + \frac{\mathbf{n}(\mathbf{n} \cdot \ddot{\mathbf{d}}) - \ddot{\mathbf{d}}}{c^2R}, \quad (3.3)$$

where  $R$  is the distance to the observer,  $\mathbf{n}$  is the unit vector in the direction of  $R$ ,  $\mathbf{d}$  is the dipole moment of the charged particle,  $\dot{\mathbf{d}}$  and  $\ddot{\mathbf{d}}$  are first and second time derivatives of the dipole moment, respectively,  $c$  is the speed of light. With the assumptions mentioned above, only the last term containing  $\ddot{\mathbf{d}}$  is present [41]

$$\mathbf{E} = \frac{\mathbf{n}(\mathbf{n} \cdot \ddot{\mathbf{d}}) - \ddot{\mathbf{d}}}{c^2R} = \frac{1}{c^2R} [[\ddot{\mathbf{d}} \times \mathbf{n}] \times \mathbf{n}]. \quad (3.4)$$

The magnetic field  $\mathbf{H}$  of the plane wave is connected to the electric field  $\mathbf{E}$  via  $\mathbf{H} = [\mathbf{n} \times \mathbf{E}]$  and is equal to

$$\mathbf{H} = \frac{1}{c^2R} [\ddot{\mathbf{d}} \times \mathbf{n}]. \quad (3.5)$$

This radiation is called dipole radiation. Since the dipole moment  $\mathbf{d}$  of the charged system is  $\mathbf{d} = \sum e_i \mathbf{r}_i$ , then  $\dot{\mathbf{d}} = \sum e_i \mathbf{v}_i$  and  $\ddot{\mathbf{d}} = \sum e_i \dot{\mathbf{v}}_i$ . Here  $e_i$  is the charge of particle  $i$  in the system,  $\mathbf{r}_i$  is the position of the individual charge,  $\mathbf{v}_i$  is its velocity, and  $\dot{\mathbf{v}}_i$  is its acceleration. So the charged particles can radiate only if they move with acceleration, uniformly moving charges do not radiate. This results from the principles of relativity, as a uniformly moving charge can be observed in an inertial system where the charge is at rest and the charge at rest does not radiate.

### 3.1. X-ray scattering on a single electron

---

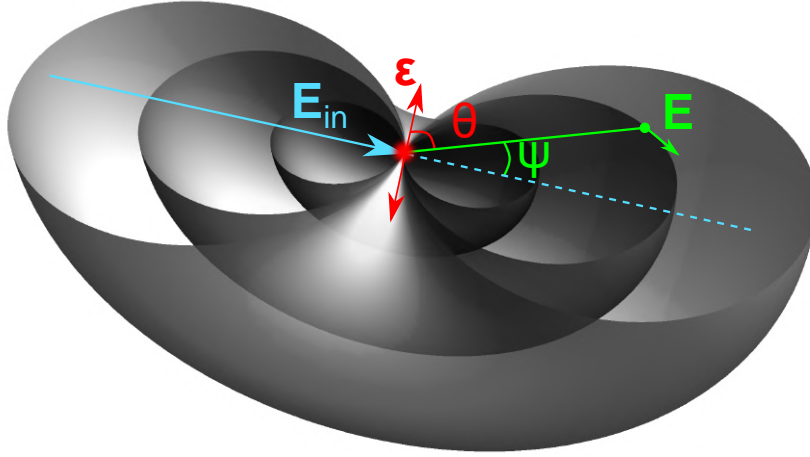


Figure 3.2: X-ray scattering on an electron depending on the angle  $\theta$  between the axis of the electric field polarization and the position of the observation point.  $\mathbf{E}_{in}$  corresponds to the incident wave,  $\mathbf{E}$  corresponds to the scattered field. Grey surfaces show equipotential scattering on the electron.

The electromagnetic waves emitted by the system carry away a certain amount of energy. The energy flux is given by the Poynting vector, equal to

$$\mathbf{S} = c \frac{H^2}{4\pi} \mathbf{n}. \quad (3.6)$$

Then the intensity  $dI$  of the radiation is defined as energy passing through a unit area per second, the unit area is placed at a distance  $R$  and covers the solid angle  $d\Omega$ . It can be written as

$$dI = \mathbf{S} R^2 d\Omega = c \frac{H^2}{4\pi} R^2 d\Omega. \quad (3.7)$$

We will later see that the energy radiated by the system per unit time per solid angle  $d\Omega$  does not depend on the  $R$  and thus, the law of conservation of energy is fulfilled.

Combining Eq. (3.5) and Eq. (3.7), we have the equation for the intensity  $dI$

$$dI = \frac{1}{4\pi c^3} \ddot{\mathbf{d}}^2 \sin^2 \theta d\Omega, \quad (3.8)$$

where  $\theta$  is the angle between  $\ddot{\mathbf{d}}$  and  $\mathbf{n}$  (see Fig. 3.2).

We also can obtain the whole radiation intensity, since  $d\Omega = 2\pi \sin \theta d\theta$  and by integrating  $d\theta$  from 0 to  $\pi$

$$I = \frac{2}{3c^3} \ddot{\mathbf{d}}^2. \quad (3.9)$$

The process of the X-ray scattering can be conveniently described by the ratio of the amount of energy emitted by the system in a given direction per unit of time (which we have obtained above) to the energy flux density of the radiation incident on the system.

This ratio is called differential scattering cross-section ( $d\sigma/d\Omega$ ) and  $d\sigma$  can be written as

$$d\sigma = \frac{dI}{S}, \quad (3.10)$$

where  $dI$  is radiated energy in  $d\Omega$  and  $S$  is the Poynting vector of the incident plane wave, both are time averaged. For the single electron in such conditions, Newton's law works

$$m_e \ddot{\mathbf{x}} = \mathbf{F} = e\mathbf{E}_{\text{in}}, \quad (3.11)$$

where  $m_e$  is electron mass,  $e$  is electron charge, the electron is accelerated by the incident electromagnetic field  $\mathbf{E}_{\text{in}}$  and  $\mathbf{F} = e\mathbf{E}_{\text{in}}$  is the Lorenz force on the electron. For the second time derivative of the dipole moment  $\ddot{\mathbf{d}}$  we then have

$$\ddot{\mathbf{d}} = \frac{e^2}{m_e} \mathbf{E}_{\text{in}}. \quad (3.12)$$

Using it in the Eq. (3.8), we can obtain  $dI$

$$dI = \frac{e^4}{4\pi c^3 m_e^2} E_{\text{in}}^2 \sin^2 \theta d\Omega. \quad (3.13)$$

Poynting vector of incident plane wave is  $S = c/(4\pi)E_{\text{in}}^2$ , so from Eq. (3.10) we now can calculate differential scattering cross-section  $d\sigma/d\Omega$

$$\frac{d\sigma}{d\Omega} = \left( \frac{e^2}{m_e c^2} \right)^2 \sin^2 \theta. \quad (3.14)$$

The term  $r_e = e^2/(m_e c^2) = 2.82 \times 10^{-13}$  cm is called the classical electron radius and it describes the electron interaction with the electromagnetic wave. From Eq. (3.14) the total cross section

$$\sigma = \frac{8\pi}{3} r_e^2 \quad (3.15)$$

and is equal to  $\sigma = 6.65 \times 10^{-25}$  cm<sup>2</sup>. It is known as the Thomson scattering cross-section.

The term  $\sin^2 \theta$  in Eq. (3.14) can be referred to as the polarization factor for scattering  $P$  and plays an important role in different types of X-ray experiments [4]. The experiments aimed to study scattering are mostly performed in the vertical scattering plane when  $P = 1$ . If we want to avoid scattering and study fluorescence (which will be described later), it is better to work in a horizontal scattering plane where  $P = 0$ . In general, polarization factor  $P$  can be defined with the angle  $\psi = 90^\circ - \theta$  as

$$P = \begin{cases} 1 & \text{vertical scattering plane} \\ \cos^2 \psi & \text{horizontal scattering plane} \end{cases} \quad (3.16)$$

**Compton scattering** Above we have discussed the elastic scattering of electromagnetic radiation by the electron. However, the inelastic scattering process can happen, as well meaning that the scattered photon has a lower frequency than the frequency of the incident photon. This process is called the Compton effect and is happening when the energy of the incident photon is quite big, for example, in the range of hard X-ray energies. Part of the energy of the photon is transferred to the electron leading to the decrease in energy which corresponds to an increase in the wavelength of the radiated photon. The effect was discovered by American physicist Arthur Compton in 1923 [42] during his experiments with X-rays: for this discovery, Compton won the 1927 Nobel Prize in physics.

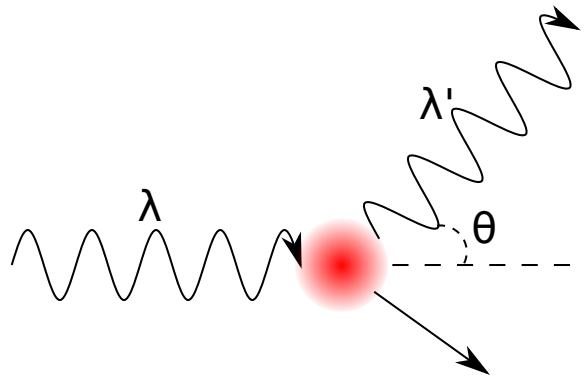


Figure 3.3: Compton scattering. An incident photon with the wavelength  $\lambda$  incoming on the electron. A new photon is scattered with the new wavelength  $\lambda'$  with angle  $\theta$  to the initial direction. As it is an inelastic process, the electron recoils with the energy conservation of the system.

The Compton effect produces incoherent radiation and is usually described by the shift in wavelength of the scattered photon (see Fig. 3.3)

$$\lambda' - \lambda = \frac{h}{m_e c} (1 - \cos \theta), \quad (3.17)$$

where  $\lambda'$  is the wavelength after scattering,  $\lambda$  is the initial wavelength,  $\theta$  is the scattering angle. The constant  $h/m_e c = 2.43 \times 10^{-11}$  cm is called the Compton wavelength of the electron. The Klein–Nishina equation [43] gives the general definition of the differential cross-section of scattering on a single free electron. For the low energy photons  $\lambda/\lambda'$  can be negligible and differential scattering cross-section is expressed with Eq. (3.14). In the Compton scattering  $\lambda/\lambda'$  can not be neglected and the total cross section decreases with increasing the photon energy.

## 3.2 X-ray scattering on an atom

Now we will consider X-ray scattering on the atom. Such a process is mainly determined by the bound electrons of the atom since the other charged particles besides the electrons

are protons which are 1836 times heavier. If we consider protons, the differential scattering cross-section will, according to Eq. (3.14), be much smaller, specifically  $\sim 1/(m^2)$ . That is why we consider X-ray scattering only on the electrons of the atom. To describe the atom we will use electron distribution  $\rho(\mathbf{r})$ , and its integration over volume  $\int \rho(\mathbf{r})d\mathbf{r} = Z$ , where  $Z$  is the number of the electrons in the atom [4]. Another assumption that we make is that electrons only scatter once. It is called the kinematical theory of X-ray diffraction and is also known as the first Born approximation. Multiple scattering events are described by the dynamical theory and equations for scattering amplitudes, in this case, are more complicated.

The total scattering from an atom will be a superposition of the scattering from volume elements containing electrons. As the incident wave interacts with the one volume element and interacts with the other on the position  $\mathbf{r}$ , this causes the phase difference

$$\Delta\phi(\mathbf{r}) = (\mathbf{k}_{in} - \mathbf{k}_f) \cdot \mathbf{r} = \mathbf{q} \cdot \mathbf{r}, \quad (3.18)$$

where  $\mathbf{k}_{in}$  and  $\mathbf{k}_f$  are the wavevectors of the incident and scattered fields respectively, and  $\mathbf{q}$  is known as wavevector transfer or scattering vector (see Fig. 3.4). The elastic nature of scattering gives us  $|\mathbf{k}_{in}| = |\mathbf{k}_f|$ , so wavevector transfer  $\mathbf{q}$  describes the direction of scattering. From the triangle we have  $|\mathbf{q}| = 2|\mathbf{k}| \sin \theta = (4\pi/\lambda) \sin \theta$ .

The volume element then contributes to the resulting scattering amplitude as  $-r_e\rho(\mathbf{r})d\mathbf{r}$  with a phase factor described in Eq. (3.18):  $e^{i\mathbf{q}\cdot\mathbf{r}}$ . The resulting scattering amplitude from electrons in the whole volume  $\mathbf{r}$

$$A(\mathbf{q}) = -r_e f_0(\mathbf{q}) = -r_e \int \rho(\mathbf{r})e^{i\mathbf{q}\cdot\mathbf{r}}d\mathbf{r}, \quad (3.19)$$

where  $r_e$  is the classical electron radius (or Thomson scattering length) and  $f_0(\mathbf{q})$  is called atomic form factor. Eq. (3.19) is known as the total scattering length of the atom or its ability to scatter an X-ray. As we consider the atom as the nucleus with the charge cloud of the electrons, after integration we have

$$f_0(\mathbf{q}) = \int \rho(\mathbf{r})e^{i\mathbf{q}\cdot\mathbf{r}}d\mathbf{r} = \begin{cases} Z & \text{for } \mathbf{q} \rightarrow 0 \\ 0 & \text{for } \mathbf{q} \rightarrow \infty \end{cases} \quad (3.20)$$

Here for the large scattering vectors  $\mathbf{q} \rightarrow \infty$ , the electrons of the atom are scattering with random phases which gives  $f_0(\mathbf{q}) \rightarrow 0$ .

It can be noted that the right part of the Eq. (3.19) is a Fourier transform of the electron density distribution. The intensity of the scattering is then

$$I(\mathbf{q}) = |A(\mathbf{q})|^2. \quad (3.21)$$

### 3.2. X-ray scattering on an atom

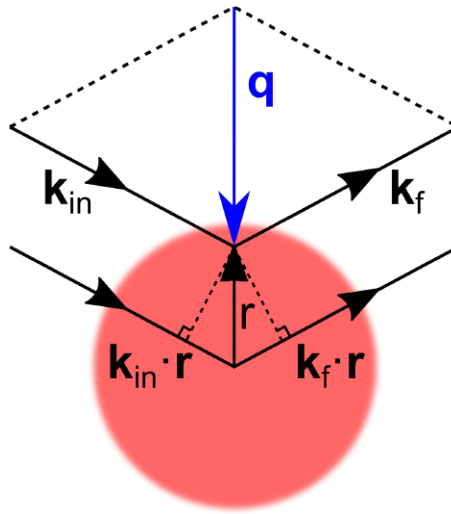


Figure 3.4: X-ray scattering on the atom. The incident wave is shown with  $\mathbf{k}_{in}$ , and the scattered wave in forward direction is shown with  $\mathbf{k}_f$ . The phase difference is shown with  $\mathbf{k}_{in} \cdot \mathbf{r}$  and  $\mathbf{k}_f \cdot \mathbf{r}$ .

Scattering from the molecule can be considered as the scattering from the group of atoms each denoted by  $j$

$$F^{mol}(\mathbf{q}) = \sum_j f_j(\mathbf{q}) e^{i\mathbf{q} \cdot \mathbf{r}_j}, \quad (3.22)$$

All above was discussed with the assumption that electrons are free, however, in reality, electrons are bound to the atoms. The energy of highly bound electrons (of K and L shells) is much bigger than the one of the electrons in less bound shells (M shell). When the incident energy of X-rays is higher than all binding energies of the atom, all electrons can be considered free which gives the total scattering length shown in Eq. (3.19). At lower X-ray energies, the corrections to the atomic form factor in Eq. (3.20) have to be introduced

$$f(\mathbf{q}, \omega) = f_0(\mathbf{q}) + f'(\omega) + if''(\omega), \quad (3.23)$$

where  $f'(\omega)$  and  $f''(\omega)$  are the real and imaginary parts of the so-called dispersion correction. When X-ray energy is close to the energy of the absorption edge (when the frequency of the X-rays is close to the frequencies of the bound electrons), the resonant effect can lead to the stronger scattering displayed by  $f'$ . That's why  $f_0(\mathbf{q})$  is called non-resonant scattering term and  $f'(\omega)$  and  $f''(\omega)$  are resonant terms. Previously it was called "anomalous dispersion". Besides the real part of the dispersion correction, we expect the phase difference between the electron and the incident field. It is shown with the term  $f''(\omega)$  and it is related to the absorption and is proportional to the absorption cross-section (see Chapter 3.5). Both the real and imaginary parts of the dispersion correction are specific for different chemical elements and highly depend on  $\omega$ .

### 3.3 X-ray scattering on an isolated particle

Here we will consider scattering from an isolated particle with the assumption that it is uniformly charged. Also, we will be looking at small scattering angles. This approach is called small angle X-ray scattering (SAXS) and allows revealing information on the size and morphology of the isolated particles, polymers, and crystals. In this case, scattering intensity can be written as

$$I(\mathbf{q}) = (\rho_{sl})^2 \left| \int_V e^{i\mathbf{q}\cdot\mathbf{r}} dV \right|^2, \quad (3.24)$$

where  $\rho_{sl} = \rho_{at}f(\mathbf{q})$  is the scattering length density (the term  $-r_e$  is implied),  $f(\mathbf{q})$  is the atomic form factor,  $\rho_{at}$  is the atom density of the particle,  $V$  is volume of the particle. By introducing a single particle form factor

$$F(\mathbf{q}) = \frac{1}{V} \int_V e^{i\mathbf{q}\cdot\mathbf{r}} dV, \quad (3.25)$$

we can rewrite Eq. (3.24) as

$$I(\mathbf{q}) = (\rho_{sl})^2 V^2 |F(\mathbf{q})|^2. \quad (3.26)$$

The form factor of different shapes and sizes usually has to be calculated according to Eq. (3.25). For simple shapes, it is well known and, for example, for the sphere with the radius  $R$  it is calculated as

$$F(q) = \frac{1}{V} \int_0^R \int_0^{2\pi} \int_0^\pi e^{iqr \cos \theta} r^2 \sin \theta d\phi d\theta dr = \frac{4\pi}{V} \int_0^R \frac{\sin(qr)}{qr} r^2 dr = 3 \frac{\sin(qR) - qR \cos(qR)}{(qR)^3} = \frac{3J_1(qR)}{qR}, \quad (3.27)$$

where  $J_1(qR)$  is the Bessel function of the first kind. For  $\mathbf{q} \rightarrow 0$  we have  $|F(\mathbf{q})|^2 = 1$  and  $I(0) = (\rho_{sl})^2 V^2$ . So as expected, the scattered intensity is proportional to the number of electrons in the particle squared. Examples of the scattered intensity from the spherical particles of the different radii are shown in Fig. 3.5.

If we want to take into account that the scattering length density is not uniform, the electron density of the particle  $\rho_p(\mathbf{r})$  can be written as

$$\rho_p(\mathbf{r}) = \rho(\mathbf{r}) \otimes [\rho_{at}(\mathbf{r}) \cdot s_p(\mathbf{r})], \quad (3.28)$$

where  $\rho(\mathbf{r})$  is the electron density of the atom,  $s_p(\mathbf{r})$  is the shape function (which is '1' inside the particle and '0' outside), and  $\otimes$  denotes convolution which is by the definition  $f(t) \otimes g(t) = \int f(\tau)g(t - \tau)d\tau$ . The Convolutional theorem [44] states that the Fourier transform of a convolution of two functions is the product of their Fourier transforms and another way

### 3.3. X-ray scattering on an isolated particle

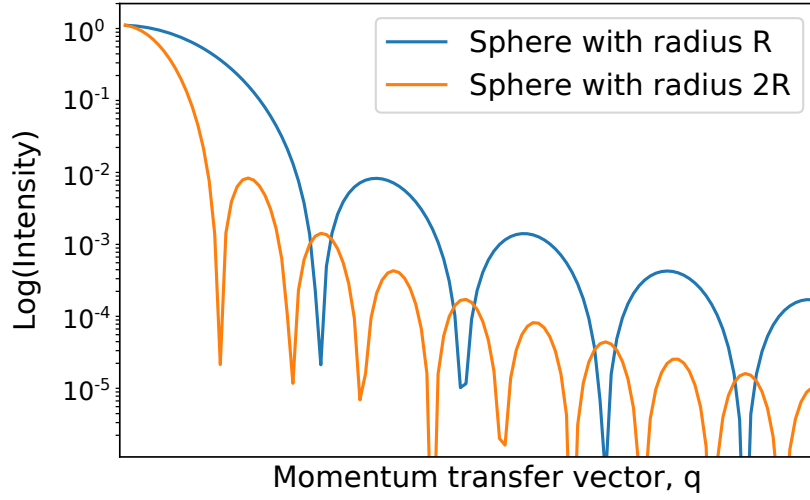


Figure 3.5: Scattering from a spherical particle of different radii. The blue line shows the intensity recorded by the detector from the scattering of the spherical particle with the radius  $R$ , orange line - from the spherical particle with the radius  $2R$ .

around

$$\mathcal{F}[f(t) \otimes g(t)] = \mathcal{F}[f(t)] \mathcal{F}[g(t)]. \quad (3.29)$$

The general form of the scattering intensity can be expressed as

$$I(\mathbf{q}) = \left| \int_V \rho_{sl} e^{i\mathbf{q} \cdot \mathbf{r}} dV \right|^2. \quad (3.30)$$

Taking the total electron density given by Eq. (3.28) and introducing particle form factor as integral over the volume of the shape function  $s(\mathbf{r})$

$$F_p(\mathbf{q}) = \frac{1}{V_p} \int_V s_p(\mathbf{r}) e^{i\mathbf{q} \cdot \mathbf{r}} dV, \quad (3.31)$$

we have the scattering intensity

$$I(\mathbf{q}) = V^2 |f(\mathbf{q})|^2 \left| \int_V \rho_{at}(\mathbf{r}) e^{i\mathbf{q} \cdot \mathbf{r}} dV \otimes F(\mathbf{q}) \right|^2. \quad (3.32)$$

The described method is widely used in X-ray science applications [45]. The structural determination of nanoparticles, macromolecules, and viruses can be performed in the so-called single particle imaging (SPI) experiments which will be discussed in Chapter 5.



### 3.4 X-ray scattering on a crystal

X-ray scattering is a powerful non-destructive experimental technique where the sample is illuminated by an x-ray beam. In this way, diffraction patterns can be recorded which contain information about the intensities and angles of scattering. X-ray diffraction on a crystal was studied by Max von Laue [2, 46] and by William Bragg [3] at the beginning of the 20th century. Since then it has become one of the most fascinating areas of science to study crystalline structures.

A crystal is a solid material containing the elements (such as atoms, molecules, ions etc.) formed into an ordered microscopic structure. It can be described with the translational vector  $\mathbf{R}_n$  (see Fig. 3.6)

$$\mathbf{R}_n = n_1 \mathbf{a}_1 + n_2 \mathbf{a}_2 + n_3 \mathbf{a}_3, \quad (3.33)$$

where  $\mathbf{a}_1$ ,  $\mathbf{a}_2$  and  $\mathbf{a}_3$  are vectors forming a unit cell which is a basic structural element of the crystal lattice and the lattice is formed by the translational symmetry of the unit cell. In Eq. (3.33),  $n_1$ ,  $n_2$  and  $n_3$  are integer numbers. While speaking of the crystal and its unit cell, usually the primitive unit cell is discussed. It means that the unit cell is described by primitive lattice vectors, and it forms the smallest possible volume.

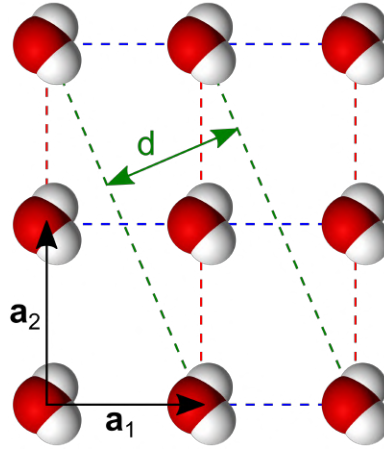


Figure 3.6: Crystal representation with the translational vector  $\mathbf{R}_n$ ,  $\mathbf{a}_1$ ,  $\mathbf{a}_2$  are forming the unit cell of the rectangular lattice. Red dashed lines show (01) planes, blue dashed lines show (10) planes, and green dashed lines show (42) planes, where  $d$  is the spacing between the planes.

The scattering amplitude of the crystal can be written as

$$F(\mathbf{q}) = -r_e \sum_l^{\text{All atoms}} f_l(\mathbf{q}) e^{i\mathbf{q} \cdot \mathbf{r}_l}, \quad (3.34)$$

where  $f_l(\mathbf{q})$  is the atomic form factor of the atom  $l$  in the position  $\mathbf{r}_l = \mathbf{R}_n + \mathbf{r}_j$ ,  $\mathbf{r}_j$  is the position of the atom in the unit cell. In this case, the Eq. (3.34) for the infinite crystal can be

### 3.4. X-ray scattering on a crystal

---

rewritten with the two terms

$$F(\mathbf{q}) = -r_e \sum_n e^{i\mathbf{q} \cdot \mathbf{R}_n} \sum_j f_j(\mathbf{q}) e^{i\mathbf{q} \cdot \mathbf{r}_j}, \quad (3.35)$$

where the first term is responsible for the lattice scattering and the second term for the unit cell scattering. The latter is called the unit cell structure factor

$$F_{uc}(\mathbf{q}) = -r_e \sum_j f_j(\mathbf{q}) e^{i\mathbf{q} \cdot \mathbf{r}_j}. \quad (3.36)$$

The scattering from the crystal can also be described via the convolution theorem. If we consider the crystal as the convolution of the lattice and basis (unit cell) part, then the scattering amplitude will be the Fourier transform of the whole crystal structure. Which according to the convolution theorem is the product of Fourier transforms of the lattice and basis parts. In calculations, Discrete Fourier Transform (DFT) is used.

**Miller indices** When we talk about X-ray diffraction on the crystal, another useful term is Miller indices. The Miller indices describe the X-ray diffraction from certain plains of the crystal. They are denoted as  $(h, k, l)$  are characterized as the plane with the intersections  $a_1/h, a_2/k, a_3/l$  on the  $(a_1, a_2, a_3)$  respectively (see Fig. 3.6). For the one plane family, all plains are equally spaced and, for example, for the cubic lattice the lattice spacing is

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}}. \quad (3.37)$$

where  $a$  is the lattice parameter.

**Laue condition** Laue determined the conditions when we can observe interference maxima as a result of the X-ray diffraction on the crystal. They are seen when the waves scattered by the crystal atoms coincide in phase or differ by an integer number of periods  $n$  which allows constructive interference. This is shown by the first sum in Eq. (3.35)

$$\mathbf{q} \cdot \mathbf{R}_n = 2\pi \cdot n. \quad (3.38)$$

As  $R_n$  is described by Eq. (3.33), the condition in Eq. (3.38) can be satisfied, when

$$\mathbf{q} = h\mathbf{a}_1^* + k\mathbf{a}_2^* + l\mathbf{a}_3^*, \quad (3.39)$$

where  $h, k, l$  are Miller indices, and  $(\mathbf{a}_1^*, \mathbf{a}_2^*, \mathbf{a}_3^*)$  are the basis vectors in so-called reciprocal space. In this terms, right part of the Eq. (3.39) is describing reciprocal lattice with the vector  $\mathbf{G}$ , so that  $\mathbf{G} \cdot \mathbf{R}_n = 2\pi(hn_1 + kn_2 + ln_3)$ . The right part gives us an integer number of  $2\pi$  and

so we have the Laue condition for the constructive interference of scattering on the crystal

$$\mathbf{q} = \mathbf{G}. \quad (3.40)$$

This is also known as the Laue condition of X-ray diffraction: we can observe intensity from the elements of the crystal if the momentum transfer vector  $\mathbf{q}$  is equal to the reciprocal lattice vector  $\mathbf{G}$ . These intensity spots are called Bragg peaks. The intensity of the Bragg peaks is described by the absolute square of the unit cell structure factor.

The reciprocal lattice vectors ( $\mathbf{a}_1^*$ ,  $\mathbf{a}_2^*$ ,  $\mathbf{a}_3^*$ ) are constructed so  $\mathbf{a}_i \cdot \mathbf{a}_j^* = 2\pi\delta^{ij}$ , where  $\delta^{ij}$  is the Kronecker symbol. In three dimensions

$$\mathbf{a}_1^* = 2\pi \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}; \quad \mathbf{a}_2^* = 2\pi \frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_2 \cdot (\mathbf{a}_3 \times \mathbf{a}_1)}; \quad \mathbf{a}_3^* = 2\pi \frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_3 \cdot (\mathbf{a}_1 \times \mathbf{a}_2)}. \quad (3.41)$$

**Bragg's Law** The Laue condition of X-ray diffraction on the crystal is equivalent to Bragg's Law which states

$$m\lambda = 2d \sin \theta, \quad (3.42)$$

where  $m$  is an integer number,  $\lambda$  is the X-ray wavelength,  $d$  is the lattice plane spacing,  $\theta$  is the incident angle (or Bragg angle). Bragg diffraction happens when the X-rays with the wavelength  $\lambda$  are hitting the crystal surface at a certain angle and experiencing constructive interference.

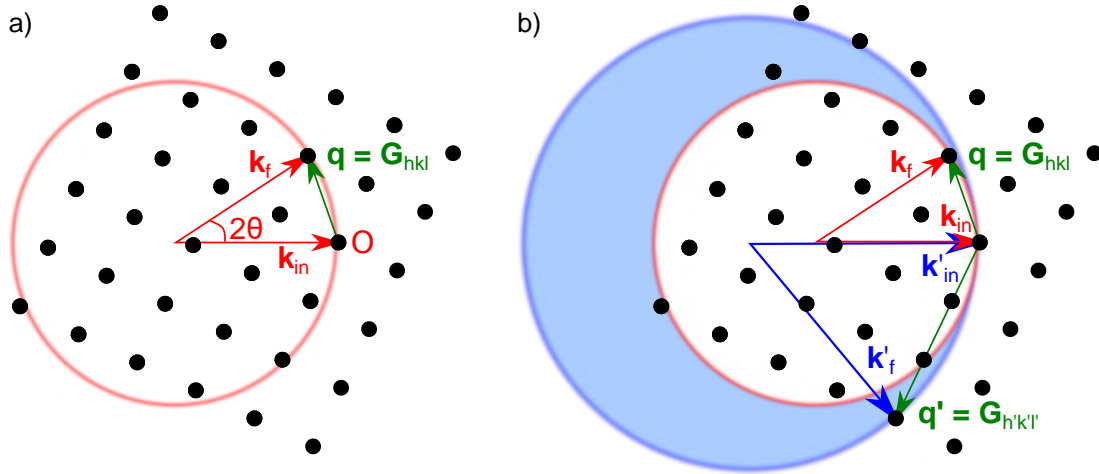


Figure 3.7: (a) The Ewald sphere for 2D crystal and the scattering triangle. The reciprocal lattice structure is shown with the dots. The incident radiation is  $\mathbf{k}_{in}$  and the scattered wavevector is  $\mathbf{k}_f$ . If another point of the reciprocal lattice lies on the Ewald sphere (red circle), then multiple scattering happens. (b) For the polychromatic X-ray beam with different X-ray wavelengths, Ewald spheres cross different reciprocal lattice points. Their difference represents the monochromaticity of the incident beam.

To prove that the Laue condition of X-ray diffraction on the crystal is equivalent to Bragg's Law, it is useful to look at the Ewald's sphere shown in Fig. 3.7. Since we consider

### 3.5. X-ray absorption

---

elastic scattering and  $|\mathbf{k}_f| = |\mathbf{k}_{in}| = |\mathbf{k}|$ , we can represent X-ray diffraction on the crystal with the sphere (which is called Ewald sphere) with the radius of  $|\mathbf{k}| = 2\pi/\lambda$  and fulfilling the Laue condition. In other words, if we want to observe another point of the reciprocal lattice, we should either rotate it or adjust the energy of the incident beam.

By the definition, the momentum transfer vector is  $|\mathbf{q}| = \mathbf{k}_{in} - \mathbf{k}_f$ . Since the Ewald sphere intersects another element of the reciprocal lattice, the Laue condition is valid and

$$\mathbf{q} = |\mathbf{G}_{hkl}| = 2|\mathbf{k}| \sin \theta. \quad (3.43)$$

Reciprocal vector  $\mathbf{G}_{hkl}$  can be also described in the terms of the Miller indices and the distance between certain plains which contributes to the diffraction. The more general way of Eq. (3.37) is

$$|\mathbf{G}_{hkl}| = \frac{2\pi}{d_{hkl}} = |h\mathbf{a}_1^* + k\mathbf{a}_2^* + l\mathbf{a}_3^*|. \quad (3.44)$$

which in combination with the previous equation gives us Bragg's law in Eq. (3.42).

The thickness of the Ewald sphere  $\Delta\mathbf{k}_{in}$  is determined by the monochromaticity of the incident beam. It should be taken into account, because under certain conditions it can make a significant contribution, for example, while calculating  $\mathbf{q}$  values on the detector.

## 3.5 X-ray absorption

Above the scattering process was considered as a result of X-ray interaction with an atom or a molecule. Another process that can happen is photoelectric absorption – X-ray photon can be absorbed by the atom and according to the law of conservation of energy, the atom is releasing the electron. It is also described with the X-ray intensity  $I(z)$  depending on the penetration position  $z$  of the sample

$$-dI(z) = I(z)\mu dz, \quad (3.45)$$

where  $\mu$  is the linear absorption coefficient. The solution of this differential equation is known

$$I(z) = I_0 e^{-\mu z}, \quad (3.46)$$

where  $I_0$  is the X-ray intensity of the incident beam at the position  $z = 0$ .

As the X-ray scattering, X-ray absorption is also characterized by the absorption cross-section. It depends on the atomic number  $Z^4$  which helps to distinguish different materials. It is also proportional to the photon energy  $1/E^3$ , tuning this parameter helps to get the desired penetration depth into the material. The linear absorption coefficient  $\mu$  is related to

the absorption cross-section  $\sigma_a$  via

$$\mu = \rho_a \sigma_a = \frac{\rho_m N_a}{M} \sigma_a, \quad (3.47)$$

where  $\rho_a$  is the atomic density,  $\rho_m$  is the mass density,  $N_a$  is the Avogadro's number,  $M$  is the molar mass.

For the composite material, the linear absorption coefficient  $\mu$  becomes the sum of different kinds of atoms  $i$

$$\mu = \sum_i \rho_{at,i} \sigma_{a,i}. \quad (3.48)$$

As was said above, photoelectric absorption is the process when an X-ray photon is absorbed by the atom, the energy is passed to the electron of the inner shell and the atom is releasing the electron. The appeared hole can be filled in two different ways. The first is called fluorescent X-ray emission. In this case, the hole is filled with the electron from the outer shell, emitting a photon with the energy corresponding to the binding energy between the shells. Another case is the so-called Auger electron emission. In this case, the inner shell vacancy is filled with another electron of the atom. This is happening with the emission of the electron from the same atom. The second emitted electron is called an Auger electron. Both processes are schematically shown in Fig 3.8. The Auger emission process was first discovered in 1922 by Lise Meitner [47]. Independently, french scientist Pierre Victor Auger discovered it in 1923 [48].

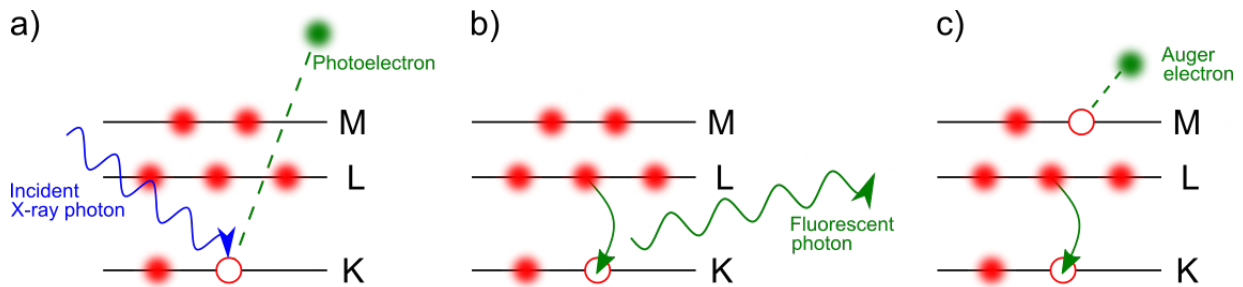


Figure 3.8: (a) The photoelectric absorption process – the atom is releasing the ionized electron. Then two following options can happen: (b) Fluorescent X-ray emission, accompanied by emitting the photon and (c) Auger electron emission, accompanied by emission of the second electron.

The X-ray absorption is studied in different kinds of experiments, and a lot of methods were developed based on the X-ray absorption features. As the absorption cross-section is following the  $1/E^3$  dependence, at characteristic energy (K-edge energy) there is a significant rise in the cross-section value due to the ionization of the K-electron. Then the value is following the mentioned energy dependence. This is studied in Extended X-ray Absorption Fine Structure (EXAFS) and X-ray Absorption Near Edge Structure (XANES) experiments. As the absorption cross-section is following the  $Z^4$  dependence, this allows observing good quality contrast between different types of matter.

## 3.6 Coherence

Coherence is an important parameter of X-rays used in many applications which are described in detail in Chapter 4 and Chapter 5. Coherence is an idealization concept which determines the possibility to observe temporal and spatial interference which is, for example, very crucial for x-ray imaging experiments. In general, coherence characterizes the correlation between wave fields, so it is usually described with the first order correlation function [10]

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, t_1, t_2) = \langle E^*(\mathbf{r}_1, t_1)E(\mathbf{r}_2, t_2) \rangle, \quad (3.49)$$

where we are considering two complex valued fields  $E(\mathbf{r}_1, t_1)$  and  $E(\mathbf{r}_2, t_2)$  in two different points  $\mathbf{r}_1$  and  $\mathbf{r}_2$  and at different times  $t_1$  and  $t_2$ . The brackets stand for the ensemble average and are defined as

$$\langle f(\mathbf{r}, t) \rangle = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{\tau=1}^N f^{(\tau)}(\mathbf{r}, t), \quad (3.50)$$

where the statistical function  $f(\mathbf{r}, t)$  has  $N$  different realizations and  $f^{(\tau)}(\mathbf{r}, t)$  is one realization  $\tau$  from the whole ensemble. If we are considering stationary fields where all ensemble averages are independent of the origin of time, then

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, t, t + \tau) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau). \quad (3.51)$$

Plus, the fields are considered to be ergodic which means that the sum (statistical average) in Eq. (3.50) can be replaced by the time average

$$\langle f(\mathbf{r}, t) \rangle_t = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(\mathbf{r}, t) dt. \quad (3.52)$$

At the same point  $\mathbf{r}$ ,  $\Gamma(\mathbf{r}, \mathbf{r}, \tau = 0)$  denotes the time average of instantaneous intensity  $I(\mathbf{r}, t)$

$$\langle I(\mathbf{r}, t) \rangle_t = \Gamma(\mathbf{r}, \mathbf{r}, \tau = 0) = \langle |E(\mathbf{r}, t)|^2 \rangle_t. \quad (3.53)$$

The cross-correlation function  $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$  is known as the mutual coherence function, and it can be normalized

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)}{\sqrt{\langle I(\mathbf{r}_1, t) \rangle} \sqrt{\langle I(\mathbf{r}_2, t) \rangle}}, \quad (3.54)$$

which is called the complex degree of coherence and can have values from 0 to 1.

If  $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = 0$ , the waves are fully incoherent and  $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = 1$  corresponds to the fully coherent case. The modulus of the degree of coherence corresponds to the contrast of the interference fringes.

The volume where the waves of the electromagnetic field are highly correlated is called the coherence volume (see Fig. 3.9 (a)). The characteristic dimensions of this volume in the spatial and temporal domains are called the transverse and longitudinal coherence lengths.

The longitudinal coherence length is responsible for the monochromaticity of the beam and shows the distance in the propagation direction when two waves are out of phase. Considering we have two waves with the wavelengths  $\lambda$  and  $\lambda + \Delta\lambda$ , they will be out of phase at the distance equal to longitudinal coherence length  $L_l$ . Then at the distance  $2L_l$ , they will be in the phase again (see Fig. 3.9 (b)), so

$$2L_l = N\lambda = (N + 1)(\lambda + \Delta\lambda), \quad (3.55)$$

where  $N$  is the number of wavelengths in the distance  $2L_l$ . Considering  $N \approx \lambda/\Delta\lambda$ , we have for the longitudinal coherence length  $2L_l$

$$L_l = \frac{1}{2} \frac{\lambda^2}{\Delta\lambda}. \quad (3.56)$$

As one can see from Eq. (3.56), the more narrow the spectral bandwidth of the beam is, the better the longitudinal coherence is. It can be estimated for the 3rd generation synchrotrons. There the longitudinal coherence is defined by the energy resolution of the monochromator used, usually, it is a crystal. At PETRA III P10 beamline, Si(111) double crystal monochromator is used. Considering the photon energy  $\lambda = 8$  keV and energy resolution of  $\Delta\lambda/\lambda \approx 10^{-4}$ , the longitudinal coherence length  $L_l$  is approximately  $1 \mu\text{m}$ . So we can assume that we obtain high contrast diffraction pattern from the desired object if its size does not exceed the longitudinal coherence length of  $1 \mu\text{m}$ .

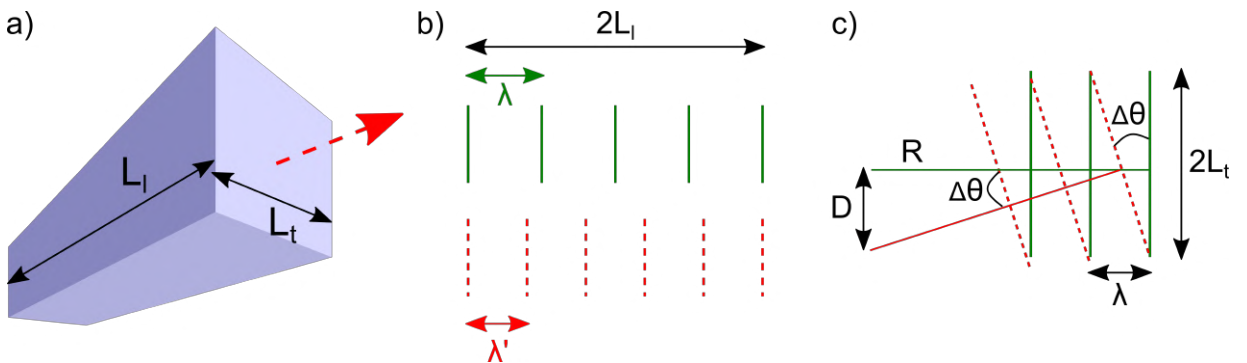


Figure 3.9: (a) The coherence volume. The red arrow shows the direction of propagation. The dimensions of the coherence volume are defined by the longitudinal coherence length  $L_l$  and transverse coherence length  $L_t$ . (b) Schematic representation of the longitudinal coherence length  $L_l$ . Two waves (green and red) with the wavelengths  $\lambda$  and  $\lambda' = \lambda + \Delta\lambda$ . (c) Schematic representation of the transverse coherence length  $L_t$ . The difference between two waves is  $\Delta\theta$  and the distance between two sources is  $D$ . The concept of (b) and (c) was adapted from [4].

The transverse coherence length is responsible for the difference in the propagation directions of the waves. For example, we have two waves which are propagating with a difference of angle  $\Delta\theta$  (see Fig. 3.9 (c)). In this case, the transverse coherence length shows where the two waves will be out of phase. Solving the problem where two waves will be in phase again, we have  $2L_t\Delta\theta = \lambda$  and  $L_t = \lambda/(2\Delta\theta)$ . If two waves are produced from two sources at the distance  $D$  and the distance to the observation point is  $R$ , then  $\Delta\theta = D/R$  and the final transverse coherence length  $L_t$  is

$$L_t = \frac{1}{2} \frac{\lambda}{(D/R)} = \frac{\lambda}{2} \frac{R}{D}. \quad (3.57)$$

As one can see from Eq. (3.57), the fully incoherent source of radiation is allowed to have non-zero transverse coherence length at large distances  $R$ . Eq. (3.57) is derived geometrically, the transverse coherence is accurately determined by the Van Citter-Zernike theorem [49] and had additional  $\pi$  in the denominator. It can also be estimated for the PETRA III synchrotron: taking the photon energy  $\lambda = 8$  keV,  $R = 90$  m for the P10 beamline station, the size of the photon beam is  $6 \mu\text{m}$  and  $36 \mu\text{m}$  in vertical and horizontal directions respectively. It gives  $360 \mu\text{m}$  and  $60 \mu\text{m}$  of transverse coherence length in both directions.

Coherent properties of the X-ray sources are considered to be of great interest in order to make them applicable for different studies and X-ray experiments. In Chapter 2 we have already mentioned that 4th generation synchrotron sources and XFELs were built to fully exploit coherent X-ray radiation.





# Chapter 4

## X-ray imaging techniques

### 4.1 Overview

Microscopic studies of living organisms and their basic smallest units – cells – are of great interest to biology, chemistry, and physics. Biological cells vary in size from 1  $\mu\text{m}$  to 100  $\mu\text{m}$ , while the proteins are from 1 to 100 nm in size. For almost four centuries, visible light microscopy, based on lenses, has been the main technology for studying such systems. It is possible to obtain images of living cells with a resolution defined by the Abbe diffraction limit for a microscope (Eq. (1.1)) of hundreds of nm but it is not enough to determine the atomic structure of the object.

Scientific interest in the structure and evolution of the smallest living objects motivated for development of advanced microscopic techniques. To probe the sample structure with higher resolution, radiation of a much smaller wavelength than visible light has to be used. In the 20th century after the discovery of X-ray radiation, X-ray microscopy started developing. X-rays with the wavelengths of Åströms, make it possible to determine the positions of individual atoms. The energy of X-ray photons ranges from several hundred eV to tens of keV. This range covers the values of binding energy in atoms for all chemical elements. X-ray microscopy is widely used to characterize the structure of various systems; one of the main directions is the imaging of biological objects and materials. Another advantage is the weak interaction with the matter which enables non-destructive probing of the bulk samples. However, radiation damage to the sample can become an issue when X-ray microscopy techniques are used.

**Tomographic Imaging** It is well known that the first x-ray radiographic image was taken by Wilhelm Röntgen in 1895. The technique evolved through time and is now widely used in medical imaging and non-destructive imaging in many industries. In modern medicine, Computer Tomography (CT) [50] is routinely used to obtain x-ray images of the 3D data by its slicing. The physicist Allan M. Cormack and engineer Godfrey N. Hounsfield were

awarded the Nobel Prize in Physiology or Medicine in 1979 [51] for the development of CT. The observed parameter is the X-ray absorption coefficient  $\mu$  (see Chapter 3.5) determined from X-ray attenuation by the body tissues. The technique is based on collecting 2D X-ray images taken from different angles. Each image is then Fourier transformed to reciprocal space where it is treated as the slice of the 3D Fourier transform of the whole object. To obtain its real space volume, the Fourier slice theorem is applied. In the 2D case, it states the following

$$F_1 P_1 = S_1 F_2, \quad (4.1)$$

where  $F$  is the Fourier transform operator,  $P$  is the projection operator, and  $S$  is the central slice operator, 1 and 2 denote the dimensions. In other words, the Fourier transform of the projected 2D function is equal to the central slice of the Fourier transform of the same 2D function. During CT, 2D x-ray images are collected and then their Fourier transform is interpolated into the one 3D Fourier transform of the sample. For that purpose, computer tomographic reconstruction algorithms are used, such as the filtered back projection method. It allows to enhance each slice and then modify them into the real space by inverse Fourier transform. Real space slices are then combined into real space 3D image of the studied object. From the reconstructed 3D density of the object, 2D slices of interest could be obtained.

**Scanning transmission X-ray microscopy** Scanning transmission X-ray microscopy [52] (STXM) is based on measuring absorption by using the zone plate lenses (see Fig. 4.1 (a)). The Fresnel zone plates (FZP) [53] allow producing images of high spatial resolution. They can be thought of as diffraction gratings of the circular form with the concentric grating lines called zones. The zone width is decreasing with the position from the center. Due to constructive interference, the zone plates can focus x-rays into a small confined spot to probe material in STXM. The resolution achieved with such a method is in the order of 10 nm [54]. In such experiments, x-ray radiation is focused on the sample, transmitted through it, and recorded with the detector as a function of the position where the sample was scanned. STXM is often used with other modalities. In combination with X-ray absorption spectroscopy, transmission images are measured at a series of wavelengths near one of the absorption edges thus providing information about the elemental composition and electronic structure [55]. In addition, fluorescent and photoelectric photons can be recorded as a function of the scanned position.

**Full-field transmission X-ray microscopy** Full-field transmission X-ray microscopy [56] (TXM) is also based on using the zone plates and they give the full 2D image on the detector (see Fig. 4.1 (b)). It is often used to illuminate the nanoscale structure of different biological samples. In full-field TXM, objective zone plates are used, they are collecting the transmitted

## 4.1. Overview

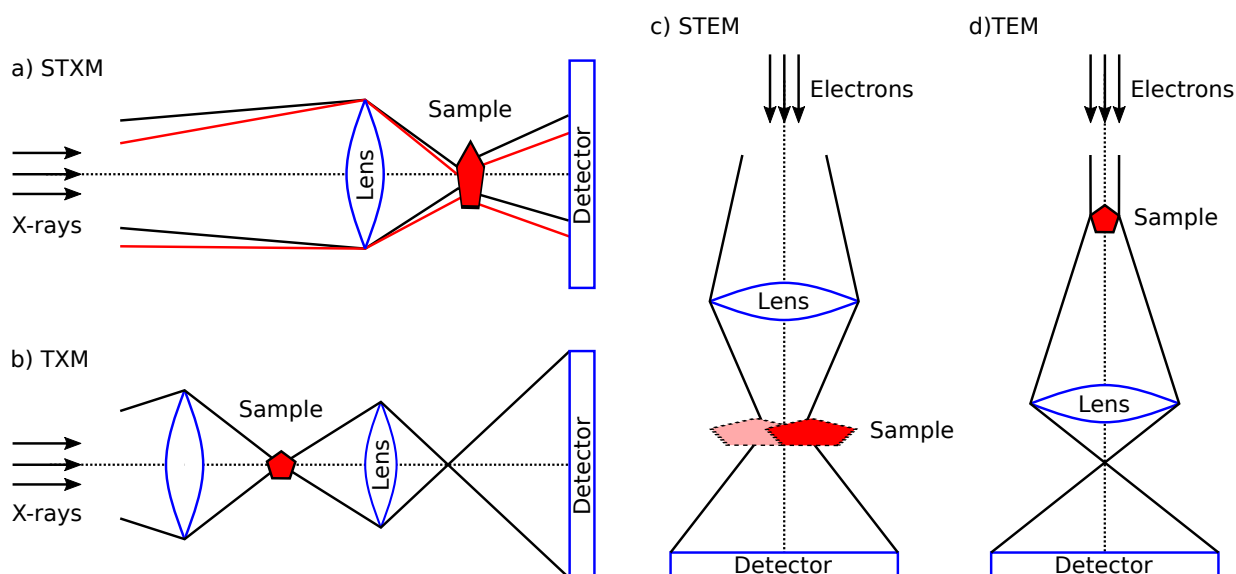


Figure 4.1: (a) Scanning transmission X-ray microscopy. The 2D images are obtained by scanning the sample. (b) Full-field transmission X-ray microscopy. The whole 2D image of the spatial absorption distribution is recorded by the detector. (c) Scanning transmission electron microscopy and (d) transmission electron microscopy are using electrons for structural determination of biological objects and allow to achieve angstrom resolution.

and diffracted radiation and construct the image on the detector. This technique can also be updated to the soft X-ray tomography (SXT) [57] which involves many projection images. This implies high total exposure, therefore it is important to use dose-efficient measurement strategies when radiation-sensitive objects are studied.

**Transmission electron microscopy** A widely used complementary alternative to X-ray microscopy in the study of structures with high resolution is electron microscopy (see Fig. 4.1 (d)). The method uses accelerated electrons and allows to study materials with better resolution and of smaller sizes. With appropriate sample preparation, the transmission electron microscope (TEM) allows determining the structure of non-crystalline biological objects with a resolution of up to several nanometers [58] and even smaller [59]. For crystalline objects, the limit of the obtained resolution is several Å, which allows imaging individual atoms [60, 61]. Electron microscopes are using electron optical lens systems to focus the electron beam. TEM has become the standard and well-known tool for structural studies in biology. It is based on electron diffraction and benefits over, for example, X-ray crystallography because there is no need to solve phase problem (missing phase information while intensity is recorded by the detector).

The main limitation of electron microscopy is related to a strong interaction of electrons and matter. The mean free path of electrons in a substance is less than 500 nm which limits the maximum thickness of the sample. When studying the structure by electron microscopy, the sample is frozen and cut into layers with a thickness of no more than a micron, and

damage occurs in the sample structure. Compared to electrons, the X-ray beam has a higher penetration depth and allows non-destructive studies of material.

**Scanning transmission electron microscopy** Scanning transmission electron microscopy (STEM) is based on scanning the sample with the focused electron beam and recording transmitted intensities (see Fig. 4.1 (c)). The position of the sample where the signal was generated and the recorded signal on the camera are then matched. Typically, the STEM technique has lower resolution compared to TEM thus allowing to study thicker objects [62]. While TEM and STEM techniques provide a resolution of about 1 Å, electron ptychography imaging recently achieved a resolution of 0.2 Å [63].

**Cryogenic electron microscopy** Another powerful technique to study biological materials with high resolution is cryogenic electron microscopy (cryo-EM) [64–66] which currently outperforms other methods of analysis of single particles, including viruses. Cryo-EM enables the imaging of biological particles with a resolution up to interatomic distances and is actively used in structural biology and material science. In this method, studied samples are frozen to cryogenic temperatures. Preparing the sample for cryo-EM is a difficult process, however, it has some benefits over, for example, X-ray crystallography. Samples do not have to be crystallized and the resulting resolution does not depend on the crystal quality. Typical resolution achieved in cryo-EM is less than 4 Å and the improvement of the technique is still ongoing [67]. The Nobel Prize in Chemistry was awarded to J. Dubochet, J. Frank, and R. Henderson “for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution” in 2017.

## 4.2 Coherent X-ray Diffractive Imaging

Coherent X-ray Diffractive Imaging (CXDI) [68–73] is a lens-less x-ray imaging technique which uses constructive interference to recover structural information of the sample in a dose-efficient way. This method allows achieving a resolution of up to several nm when determining the structure of the samples. In the study of biological samples, the CXDI technique complements electron microscopy in terms of spatial resolution, contrast in the structure of objects, and general possibilities. At the same time, the CXDI has several advantages. This method is highly sensitive to changes in the density of biological objects which allows reconstruction of the sample with the internal volume.

Optical elements (such as Fresnel Zone Plates (FZP), Kirkpatrick-Baez (KB) [74] mirrors or compound refractive lenses (CRL) [75]) are used for the focusing of the X-ray beam. At the same time, there are no optical elements between the sample and the detector in CXDI experiments which simplifies the measurement at the cost of data analysis complexity. The

intensity of the diffraction patterns measured in CXDI experiments decreases as  $I(|\mathbf{q}|) \propto |\mathbf{q}|^{-k}$ , where  $k$  ranges from 3 to 4 depending on the geometry [76, 77]. The resolution in CXDI method is theoretically limited by the highest momentum transfer  $|\mathbf{q}|$  at which the measured intensity exceeds the noise level.

The calculation of the wave field propagation in CXDI can be simplified within certain approximations. The near-field (Fresnel) and far-field (Fraunhofer) approximations are described with the Fresnel number

$$N_f = \frac{D^2}{\lambda z}, \quad (4.2)$$

where  $D$  is the size of the aperture,  $\lambda$  is the wavelength of the incident wave, and  $z$  is the distance from the aperture. If  $N_f \sim 1$ , Fresnel diffraction or the near-field is applied. If  $N_f \ll 1$ , implying that the size of the aperture is much smaller than the wavelength and the propagation distance, Fraunhofer diffraction or the far-field is applied. In CXDI, far-field diffraction of a coherently illuminated object is recorded.

There are two main ways to perform CXDI experiments. The first one is based on the recording forward scattering signal in the transmission geometry (Fig. 4.2). It is typically used for imaging non-crystalline objects. The second one is based on recording diffracted x-rays from the sample oriented to fulfill the Bragg condition (see Fig. 4.3). Therefore this method is called Bragg Coherent X-ray Diffractive Imaging (BCDI). It is used for imaging crystalline objects. Later in this Section, we will discuss this method in more detail.

**Coherent X-ray Diffractive Imaging in forward direction** The typical scheme of the CXDI experiment in the forward scattering geometry is shown in Fig. 4.2. In this case, the electron density of the object is reconstructed. The assumptions which are made here are that the scattering is weak (single scattering events) and it is neglecting diffraction within the scattering volume (projection approximation [69]). With these assumptions, the transmitted and scattered waves coexist and can be comparable in terms of intensities.

After exiting the object, the further propagation of the wave field can be described as

$$A(\mathbf{q}) = |A(\mathbf{q})| e^{i\varphi(\mathbf{q})} \propto \mathcal{F}\{P(\mathbf{r})O(\mathbf{r})\}, \quad (4.3)$$

where  $\varphi(\mathbf{q})$  is the phase,  $P(\mathbf{r})$  is the probe function,  $T(\mathbf{r})$  is the object transmission amplitude, and  $O(\mathbf{r})$  is the object function so that  $O(\mathbf{r}) = T(\mathbf{r}) e^{i\varphi(\mathbf{r})}$ .

Eq. (4.3) uses a simple Fourier transform as we consider that detector is located in the far-field region. The wave propagated through the object in Eq. (4.3) is called the exit surface wave. If one can reconstruct the exit surface wave and the incident wave (or it is known), the projected electron density of the object can be obtained. This is used in ptychography [78] which is a scanning CXDI method using divergent waves for studying extended objects. Different set-ups are used in ptychography, such as classical pinhole set-up where a chosen

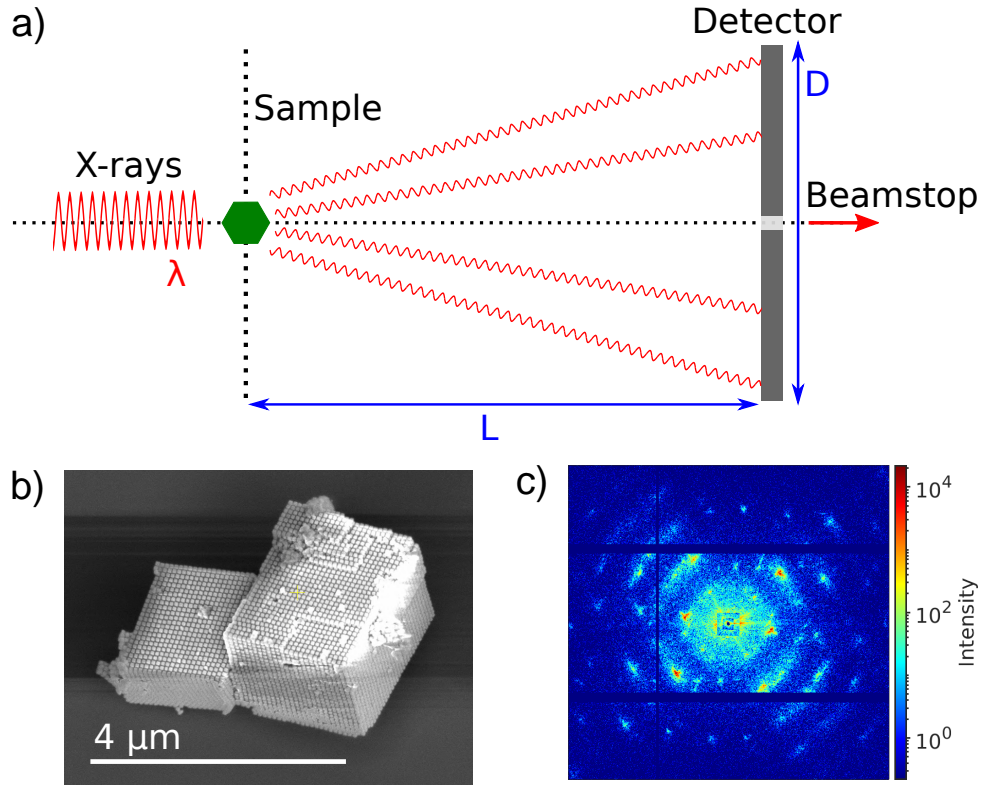


Figure 4.2: (a) Experimental set-up for CXDI experiment in forward direction. X-rays with the wavelength  $\lambda$  are illuminating the sample, the scattered intensity is recorded by the detector in the far-field. (b) Example of the structures which could be analyzed in CXDI. This is SEM of the gold mesocrystal consisting of the 60 nm particles. (c) An example of the diffraction pattern from the CXDI experiment with this mesocrystal at PETRA III synchrotron.

pinhole sets the size and the shape of the incident X-ray beam. In the framework of this Thesis, a plane incident wave is used on an isolated particle, so the term of  $P(\mathbf{r})$  will be omitted.

The detector in CXDI records 2D diffraction patterns in the far-field, each of them corresponding to the cross-section in reciprocal space and is described by the Ewald sphere. Each diffraction measurement recorded by the detector contains information only about the amplitude of the generally complex-valued wave field

$$I(\mathbf{q}) = |A(\mathbf{q})|^2, \quad (4.4)$$

where  $I(\mathbf{q})$  is the measured intensity and  $A(\mathbf{q})$  is the scattering amplitude of the object. It is important to note that the phase information is missing. In CXDI the sample is rotated and by collecting diffraction patterns at different angles, the whole reciprocal space of the object can be constructed

$$A(\mathbf{q}) \propto \int \rho(\mathbf{r}) e^{i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r}. \quad (4.5)$$

The procedure to transform the reciprocal space volume to the real space electron density of the object will be described in Sec. 4.3.

**Bragg Coherent X-ray Diffractive Imaging** As mentioned at the beginning of Section 4.2, BCDI is the experiment when the Bragg condition (see Eq. (3.42)) is satisfied. It is used for imaging crystalline objects and the measurements of the diffracted signal are performed in the vicinity of the selected Bragg peak. It allows studying material properties of the semiconducting nanowires [79], strain fields in nanocrystals [73, 80–82]. For example, a recent BCDI experiment [83] showed structural changes in a platinum nanoparticle of 160 nm diameter under different reaction conditions.

The typical scheme of the BCDI experiment is shown in Fig. 4.3. The X-ray diffraction patterns are the result of the constructive interference on the crystallographic planes of the sample and are recorded by an area detector. Bragg geometry allows seeing defects in the crystal lattice with the broadening of the corresponding Bragg peak. In the BCDI experiments, not only strain fields and atomic structure can be studied, but it is also possible to obtain the shape of the crystal as the facets of the crystal bring additional scattering contribution to the diffraction pattern. The real-space reconstruction in the BCDI method is computationally obtained from the measured diffraction intensities using the phase retrieval techniques. (see Sec. 4.3).

The scattering amplitude of an infinite crystal was expressed by Eq. (3.35). Here we will consider the crystalline object to be finite [84] and to have a shape denoted with the shape function  $s(\mathbf{r})$  that is '1' inside the volume of the crystal and '0' outside the volume. The total electron density  $\rho(\mathbf{r})$  is now combined not only of unit cell factor  $\rho_{uc}(\mathbf{r})$  and lattice factor  $\rho_{\infty}(\mathbf{r})$  but also the shape function  $s(\mathbf{r})$

$$\rho(\mathbf{r}) = \rho_{uc}(\mathbf{r}) \otimes [\rho_{\infty}(\mathbf{r}) \cdot s(\mathbf{r})]. \quad (4.6)$$

Applying the convolution theorem (Eq. (3.29)), we have for the scattering amplitude  $A(\mathbf{q})$

$$A(\mathbf{q}) = F_{uc}(\mathbf{q}) \cdot \hat{\rho}_{\infty}(\mathbf{q}) \otimes \hat{s}(\mathbf{q}), \quad (4.7)$$

where  $F_{uc}(\mathbf{q})$  is the unit cell structure factor from Eq. (3.36) and  $\hat{\rho}_{\infty}(\mathbf{q})$  is the lattice factor and can be written as

$$\hat{\rho}_{\infty}(\mathbf{q}) = \frac{(2\pi)^3}{V_{uc}} \sum_{hkl} \delta(\mathbf{q} - \mathbf{G}_{hkl}). \quad (4.8)$$

Here  $V_{uc}$  is the volume of the unit cell,  $\mathbf{G}_{hkl}$  is the reciprocal lattice vector, and  $\delta(\mathbf{q} - \mathbf{G}_{hkl})$  denotes the Dirac delta function.

With the presence of the defect in the crystal, the structure is deformed which can be described with the displacement  $\mathbf{u}(\mathbf{r})$ . In this case, the electron density of the crystal has an



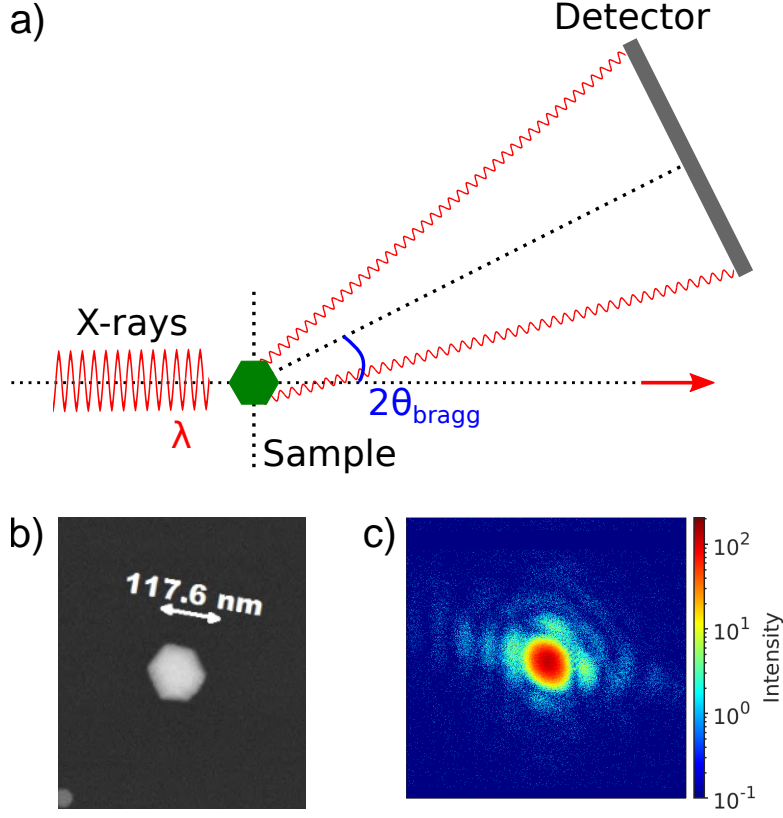


Figure 4.3: (a) Experimental set-up for BCDI experiment. X-rays with the wavelength  $\lambda$  are illuminating the sample, the scattered intensity is recorded by the detector in the Bragg geometry. (b) Example of the structures which could be analyzed in BCDI. This is an SEM of the platinum nanoparticle with a size of 117.6 nm. (c) One example of the diffraction patterns from the BCDI experiment with this nanoparticle at PETRA III synchrotron.

additional displacement term in the sum of all atoms

$$\rho(\mathbf{r}) = \sum_{n=1}^N \sum_{j=1}^M \rho_{nj}(\mathbf{r} - \mathbf{R}_{nj} - \mathbf{u}(\mathbf{R}_{nj})), \quad (4.9)$$

where  $\mathbf{R}_{nj} = \mathbf{R}_n + \mathbf{r}_j$  is the position of the atom  $j$  in the unit cell  $n$ .

The scattering amplitude  $A(\mathbf{q})$  then also has this additional displacement term

$$A(\mathbf{q}) = \sum_{n=1}^N F_{uc}^n(\mathbf{q}) e^{-i\mathbf{q} \cdot \mathbf{u}(\mathbf{R}_n)} e^{-i\mathbf{q} \cdot \mathbf{R}_n}, \quad (4.10)$$

where  $F_{uc}^n(\mathbf{q})$  is a structure factor of the unit cell  $n$ . Having finite crystal with the shape function  $s(\mathbf{r})$ , the scattering amplitude around the Bragg peak is then

$$A(\mathbf{Q}) = \int s(\mathbf{r}) e^{-i\mathbf{G}_{hkl} \cdot \mathbf{u}(\mathbf{r})} e^{-i\mathbf{Q} \cdot \mathbf{r}} d(\mathbf{r}), \quad (4.11)$$

where  $\mathbf{Q} = \mathbf{q} - \mathbf{G}_{hkl}$ .

From Eq. (4.11), it is seen that in the case of the perfect crystal, we observe symmetric intensity distribution of the Bragg peak. However, the strain in the crystalline object leads to asymmetry of the Bragg peak, the higher the strain, the more asymmetric the Bragg peak is because of the displacement contribution to the phase of the object function. BCDI experiments are performed by rotation around the Bragg peak position so that one can collect several cross-sections through reciprocal space. Then the recorded diffraction patterns are merged into the single 3D distribution of intensities in reciprocal space. This 3D distribution is used for computing the missing phase distribution and reconstruction of the object function in real space by means of phase retrieval algorithms. From Eq. (4.11), we see that the amplitude of the reconstruction result denotes the shape of the sample and the phase corresponds to the projection of the displacement field on the reciprocal lattice vector. It plays an important part in the operando studies, for example, BCDI is actively used [85] for high-resolution structural characterization of the nanoparticles in coin cell batteries.

**Phase problem** In CXDI experiments only the amplitude of the complex-valued wave field is measured

$$\sqrt{I(\mathbf{q})} = |A(\mathbf{q})|. \quad (4.12)$$

Without the phase information, the inverse Fourier transform will not give the correct real space image of the sample which is referred to as the phase problem in optics [86]. Throughout the years, different techniques have been developed using additional constraints for recovering the phases from the measured intensities. Various iterative phase retrieval techniques [87, 88] have been developed to solve this complex problem efficiently. The following Section is dedicated to different iterative phase retrieval algorithms used in this Thesis.

**Sampling** In X-ray imaging techniques, the scattering signal is recorded by 2D detector with the finite number of pixels that sample the continuous diffraction pattern. The frequency of such a discrete representation of a continuous function is generally known as the sampling rate. The minimum sampling rate for the highest measured frequency is defined by the Nyquist-Shannon theorem [89, 90] (or Kotelnikov theorem [91]). Shannon's interpretation is the following: if spectrum of the signal contains frequencies less or equal to  $n$  Hz, the signal can be completely resolved by a set of values spaced  $1/(2n)$  seconds between them. Kotelnikov theorem states basically the same: any function which contains frequencies from 0 to  $f_c$ , can be transferred continuously with any precision using numbers following one after another with the spacing of  $1/(2f_c)$  seconds.

In X-ray imaging, the concept above is also applied. As in Eq (4.12), if we take the Inverse Fourier transform of the measured intensity, we obtain the auto-correlation of the wave field

$$\mathcal{F}^{-1}\{I(\mathbf{q})\} = A^*(-\mathbf{r}) \otimes A(\mathbf{r}), \quad (4.13)$$

while

$$I(\mathbf{q}) = |\mathcal{F}\{\rho(\mathbf{r})\}|^2, \quad (4.14)$$

where  $\rho(\mathbf{r})$  is the electron density of the object. If we implement Kotelnikov or Nyquist-Shannon theorem to the measured diffraction intensity (Eq. (4.14)), the feature of the size  $\Delta l$  is recovered when the corresponding fringes in reciprocal space have at least two pixels per fringe. In other words, since we measure the auto-correlation function of the object (Eq. (4.13)), it extends twice the object size. Consequently, recovering the feature  $\Delta l$  in reciprocal space requires the frequency  $2\pi/\Delta l$  to fulfill the condition

$$\frac{2\pi}{\Delta l} \geq 2 p, \quad (4.15)$$

where  $p$  is the pixel size of the detector.

The real and reciprocal space samplings ( $\Delta x$  and  $\Delta q$ , respectively) are related by the Discrete Fourier transform (DFT)

$$\Delta x \Delta q = \frac{2\pi}{N}, \quad (4.16)$$

where  $N$  is the number of sampling points. If we neglect the curvature of the Ewald sphere, the sampling in reciprocal space is

$$\Delta q = \frac{2\pi p}{\lambda L}. \quad (4.17)$$

From Eq. (4.16) and Eq. (4.17) we obtain a formula for the sampling in real space

$$\Delta x = \frac{\lambda L}{N p}. \quad (4.18)$$

Eq (4.18) means that the resolution in real space is limited to  $(\lambda L)/(Np)$ . In most real cases, the resolution is, in fact, worse than that and is rather determined by the biggest scattering angle where the measured signal is still meaningful. This highly depends on how much the sample scatters and how much of x-ray radiation it can withstand. The scattering intensity quickly drops with the scattering angle, therefore high-resolution measurements require higher radiation dose. The radiation dose and its effect on biological samples will be discussed in more detail in Section 5.2.

### 4.3 Iterative phase retrieval techniques

Throughout the years, different techniques have been developed to solve the phase problem from measured intensities. The invention of the iterative algorithms was demonstrated by Gerchberg and Saxton [92]. An essential contribution to the development of the iterative phase retrieval algorithms was made by James R. Fienup [93] and Stefano Marchesini [88].

### 4.3. Iterative phase retrieval techniques

In this Section, we will give a description of the number of algorithms, their principles, and convergence conditions.

The basis of all iterative phase retrieval algorithms is iterative applying Direct and Inverse Fourier transformations, so that at each iteration, the output amplitudes are replaced by the measured ones while the phases are left to evolve. The questions about the uniqueness of the solution were actively studied [94–96]. It was shown that the solutions to the phase problem in more than one dimension are almost always unique for the localized objects. For the two-dimensional real positive sample it was proven [97] that its form is uniquely related to the intensity of its Fourier transform and with the appropriate sampling (see Section 4.2), the auto-correlation of the image can be reconstructed.

The iterative phase retrieval algorithms rely on fulfilling the constraints. From the measurements, we have the first constraint – we know the diffraction intensity which gives the amplitude of the complex values wave field (Eq. (4.12)). This is the modulus constraint and it is operating in reciprocal space. The second constraint is based on *a priori* information about the object. The sample is isolated and has a finite size which can be shown with the binary mask

$$S(\mathbf{r}) = \begin{cases} 1 & \text{if } \mathbf{r} \in S \\ 0 & \text{else} \end{cases}, \quad (4.19)$$

where  $S(\mathbf{r})$  is the volume in real space and is commonly referred to as support. In some cases, additional real space constraints, such as non-negative amplitudes, can be applied.

Solving the phase problem is equivalent to finding the intersection of the set of all objects with the measured amplitudes in the experiment and the set of all objects within the predefined support volume (see Fig. 4.4 (a)).

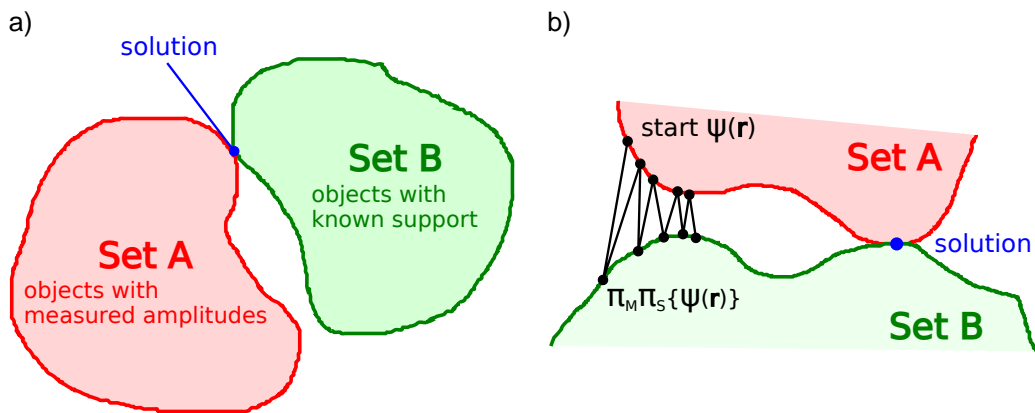


Figure 4.4: (a) Schematic interpretation of the phase problem. The task of the iterative phase retrieval is to solve the phase problem with the measured diffraction intensities and *a priori* information about the sample – usually, it is the finite dimensions of the object. (b) Convergence of the phase retrieval algorithm aiming to minimize the error metric.  $\pi_S$  and  $\pi_M$  denotes support and modulus projectors from Eq. (4.21) and Eq. (4.25).

As was stated, the phase retrieval process consists of the iterations between real and reciprocal space. Each iteration starts with the known information about the amplitude in reciprocal space (modulus constraint). For the phases, random values are taken as an initial approximation. The obtained values of the complex amplitude are converted into real space by the Inverse Fourier transform and the initial approximation of the object is calculated. Based on the *a priori* information about the structure of the object, support constraint is imposed on the obtained structure. The resulting electron density distribution is translated into reciprocal space by the Fourier transform. The output phases are preserved while amplitudes are replaced by those measured in the experiment and the next iteration starts (see Fig. 4.4 (b)).

In the following Sections, an overview of the most commonly used algorithms is given.

### 4.3.1 Gerchberg-Saxton algorithm

The first iterative phase retrieval algorithm was developed in 1972 by Gerchberg and Saxton [92]. It utilized diffraction data to reconstruct the phase of the object in the case of two intensity measurements. Originally, the first measurements were considered as the amplitude of the object in real space and the amplitude of its diffraction pattern. The aim was to recover phases

$$A(\mathbf{q}) = |A(\mathbf{q})|e^{i\phi(\mathbf{q})}, \quad (4.20)$$

using the measured intensity  $I(\mathbf{q}) = |A(\mathbf{q})|^2$  and the amplitudes of the image in real space. The first iteration starts by combining the amplitudes of the image and the first random estimation of the phases. Its Fourier transform is calculated and the resulting obtained phases are then combined with the respective amplitudes of the diffraction pattern of the image. The new field is then again Fourier transformed and the phases of the sample are again combined with the amplitudes of the image. Thus, a new estimation of the complex wave field  $A(\mathbf{q})$  is created and the process is starting all over again until both constraints in real and reciprocal space are satisfied. The schematic representation of the algorithm is shown in Fig 4.5.

In general, iterative algorithms could be expressed mathematically with the projector operator  $\pi$  on the constraint set of the Fourier amplitudes  $M$  which were measured in the experiment  $\sqrt{I(\mathbf{q})}$

$$A'(\mathbf{r}) = \pi_M\{A(\mathbf{r})\} = \mathcal{F}^{-1} \left\{ \frac{\mathcal{F}\{A(\mathbf{r})\}}{|\mathcal{F}\{A(\mathbf{r})\}|} \sqrt{I(\mathbf{q})} \right\}, \quad (4.21)$$

where  $A'(\mathbf{r})$  is an updated image in real space.

The important attribute of all iterative phase retrieval algorithms is the performance metric. If the algorithm works ideally, the calculated phases in combination with the measured

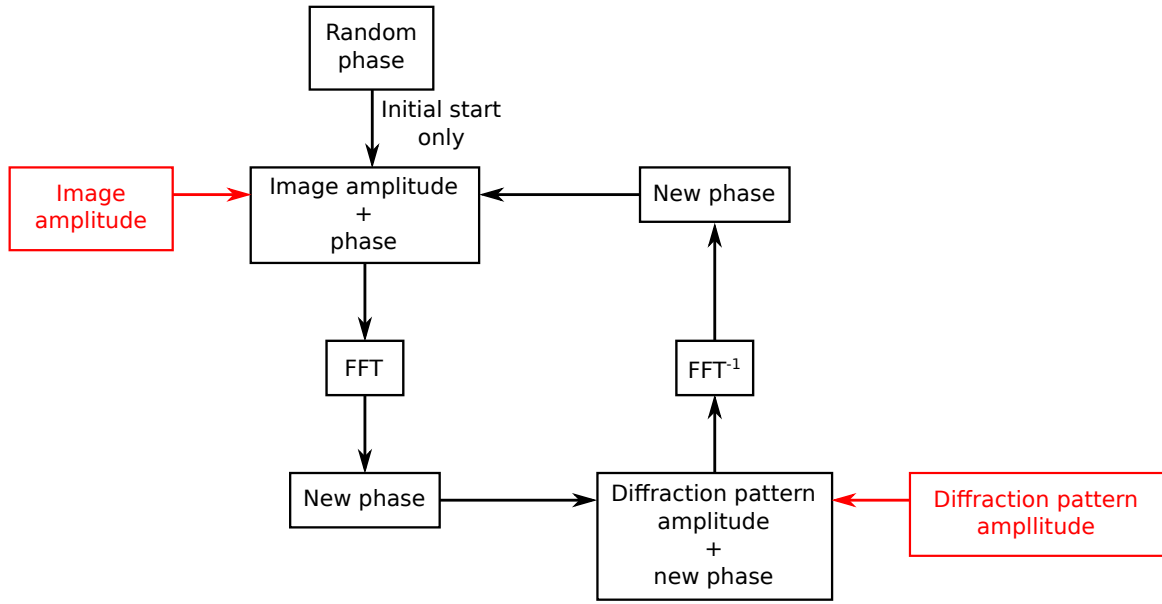


Figure 4.5: Schematic representation of the Gerchberg-Saxton algorithm. It starts with the estimation of the phase of the image, at the first estimation, it is a random phase. During the iteration of the algorithm two constraints are applied: we know the image amplitude from the measurements and we know diffraction pattern amplitude from the measurements as well. Known parameters of the iteration are marked with red.

amplitudes correspond to the object of a certain form (match the support constraint) and the intensities of its diffraction pattern corresponding to the measured data (match the modulus constraint). In practice, however, the data are not ideal, we have detector gaps where data is missing, and experimental data are not noise-free. Because of all these data issues, algorithms often are not able to find the ideal solution. That is why as a measure of convergence, the normalized error metric in reciprocal space is used

$$\varepsilon_M = \frac{\sum_{\mathbf{q}} \left| |A'(\mathbf{q})| - \sqrt{I(\mathbf{q})} \right|^2}{\sum_{\mathbf{q}} |\sqrt{I(\mathbf{q})}|^2}. \quad (4.22)$$

The iteration process is repeated until the error value  $\varepsilon_M$  at each iteration is less than a predetermined threshold value.

#### 4.3.2 Error-Reduction

The Error-Reduction (ER) algorithm is the improved version of the Gerchberg-Saxton algorithm proposed by J. Fienup [98]. The difference between these two algorithms is that the real space constraint in ER is based on using the localized support region of the object (Eq. (4.19)). It is not an exact image of the sample as it was in the Gerchberg-Saxton algorithm but the area where the electron density function of the object is non-zero. The schematic representation of ER is shown in Fig 4.6.

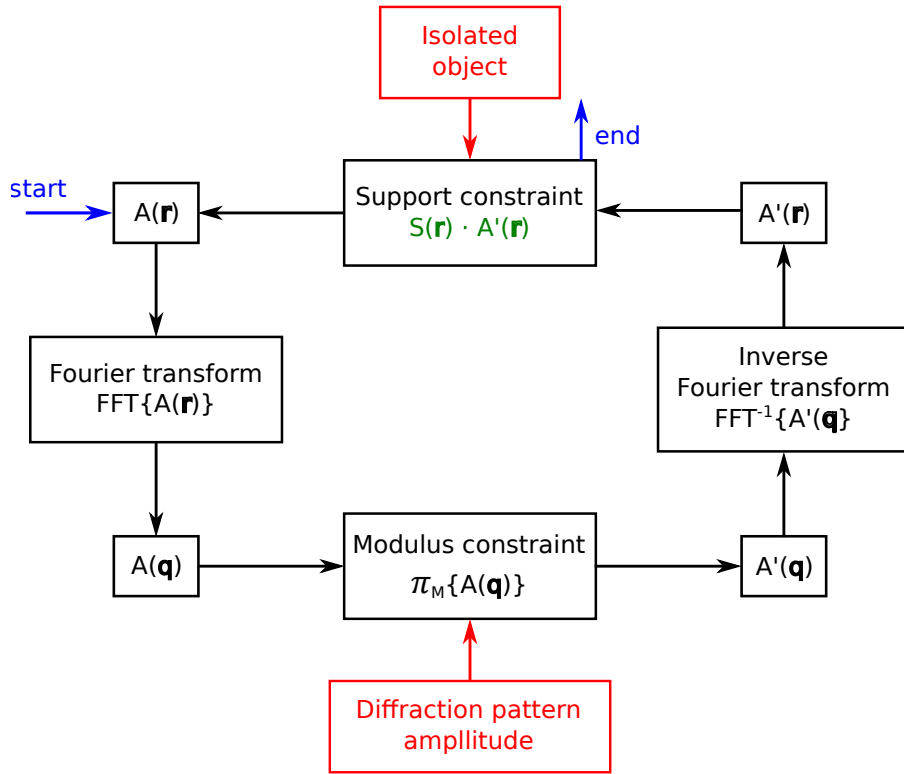


Figure 4.6: Schematic representation of the Error-Reduction algorithm. The support constraint during the algorithm is a binary mask as we consider the object to be isolated. The modulus constraint is available from the experimental measurements. The final result of the iteration is marked with green.

The first iteration (which will be denoted with the subscript '1') starts with the combining of the measured amplitudes  $\sqrt{I(\mathbf{q})}$  and randomly assigned phases  $A_1(\mathbf{q})$

$$A_1(\mathbf{q}) = \sqrt{I(\mathbf{q})} e^{i\phi_1(\mathbf{q})}. \quad (4.23)$$

Inverse Fourier transform of obtained  $A_1(\mathbf{q})$  will give us the image in real space

$$A_1(\mathbf{r}) = \mathcal{F}^{-1}\{A_1(\mathbf{q})\}. \quad (4.24)$$

The output is then combined with the support constraint as our object is isolated and has a certain shape which gives us an updated estimate of the wave field (denoted with ') at this step

$$A'_1(\mathbf{r}) = A_1(\mathbf{r}) \cdot S(\mathbf{r}), \quad (4.25)$$

which can be also called as support projection  $\pi_S\{A(\mathbf{r})\}$ . The updated field of the object  $A'_1(\mathbf{r})$  after applied Fourier transform will give new reciprocal space amplitudes  $A'_1(\mathbf{q})$  from the measured ones and phases  $A'_1(\mathbf{q})$ . Here modulus constraint is applied which yields the

### 4.3. Iterative phase retrieval techniques

---

use of measured amplitudes  $\sqrt{I(\mathbf{q})}$  instead of the new ones

$$A'_1(\mathbf{q}) = \sqrt{I(\mathbf{q})}e^{i\phi'_1(\mathbf{q})}, \quad (4.26)$$

which is the application of the modulus projection in Eq. (4.21). Then the second iteration starts from the previously computed wave field

$$A_2(\mathbf{q}) = A'_1(\mathbf{q}) \quad (4.27)$$

and this process is repeated through iterations.

Usually, the first guess of the support  $S(\mathbf{r})$  is estimated as the auto-correlation function. This function according to the Eq. (4.13) is an Inverse Fourier transform of the measured intensity

$$f_{ac}(\mathbf{r}) = |\mathcal{F}^{-1}\{I(\mathbf{q})\}|, \quad (4.28)$$

which denotes the first assumption of the support. The support can be also updated during the iterative phase retrieval and certain algorithm exist to target the support. It is called Shrink-Wrap (SH) algorithm and it will be discussed further.

As it was mentioned before, the ultimate goal is obtaining the ratio between measured data and predicted wave amplitudes equal to one

$$\frac{\sqrt{I(\mathbf{q})}}{|A_n(\mathbf{q})|} = 1, \quad (4.29)$$

where  $A_n(\mathbf{q})$  is the result of the  $n$ -th iteration of the algorithm. The error metric (Eq. (4.22)) in the case of ER always decays monotonically [93, 98] which is displayed in the title. The iterative process of ER is shown in Fig. 4.4 (b). The disadvantage of the approach is that this algorithm could stuck in the local minima and could not get out of it because it is aimed to keep the error small (stagnation problem). The possibility to escape from the local minima requires a rise in the error value. These issues were overcome in other phase retrieval algorithms.

#### 4.3.3 Hybrid-Input-Output

Hybrid-Input-Output (HIO) algorithm is one which helps to avoid the stagnation problem of ER algorithms. As it follows from the title, HIO is based on the combination of the input and output results of the iteration. The output of the previous iteration is modified according to the constraints and then fed as the input to the next iteration. HIO uses modified



Eq. (4.25) according to the rule

$$A'_1(\mathbf{r}) = \begin{cases} A_1(\mathbf{r}) & \text{if } \mathbf{r} \in S \\ A(\mathbf{r}) - \beta\psi_1(\mathbf{r}) & \text{if } \mathbf{r} \notin S \end{cases}, \quad (4.30)$$

where  $A(\mathbf{r})$  and  $A'_1(\mathbf{r})$  are the input and the output of the iteration respectively,  $\beta$  is so-called feedback parameter and  $\beta \in (0, 1)$ . Typical values of the feedback parameter  $\beta = 0.8$ . The idea is not to put strict zeros outside of the support  $S(\mathbf{r})$  but to go there smoothly by using the input and output. It brings non-linearity to the HIO algorithm and allows to avoid local minima. Of course, the error metric from Eq. (4.22) does not display real error since the  $A'_1(\mathbf{r})$  is not the direct estimate of the object. Error, in this case, in comparison with the error in ER algorithm is not necessarily declining.

The combination of HIO and ER algorithms were proved to give good results with the phase problem in CXDI. HIO iterations are helpful in searching for the global solution and ER iterations are added as a refinement of HIO result.

#### 4.3.4 Continuous Hybrid-Input-Output

HIO was further developed into the Continuous Hybrid-Input-Output (CHIO) algorithm. It was described in Ref. [99] and was designed to solve the disadvantage of the HIO which is a big difference between the input and output results. This discontinuity between the next input and current output was addressed in CHIO modification where Eq. (4.30) was updated to the following form

$$A'_1(\mathbf{r}) = \begin{cases} A_1(\mathbf{r}) & \text{if } \mathbf{r} \in S, \alpha A(\mathbf{r}) \leq A_1(\mathbf{r}) \\ A(\mathbf{r}) - \frac{1-\alpha}{\alpha} A_1(\mathbf{r}) & \text{if } 0 \leq A_1(\mathbf{r}) \leq \alpha A(\mathbf{r}) \\ A(\mathbf{r}) - \beta A_1(\mathbf{r}) & \text{otherwise} \end{cases}, \quad (4.31)$$

where another parameter  $\alpha$  appears and typically  $\alpha = 0.4$ . Experiments show the efficiency of CHIO and in the framework of this Thesis, it will be used in combination with other phase retrieval algorithms.

#### 4.3.5 Shrink-Wrap

Shrink-Wrap (SW) [100] algorithm is an additional tool in phase retrieval which allows updating the support region  $S(\mathbf{r})$  used in Eq. (4.25). As it is clear from the title, SW is responsible for shrinking  $S(\mathbf{r})$ . Practically, good results are obtained when the support  $S(\mathbf{r})$  is known or as accurate as possible. That is why tightening the support can lead to the global solution during iterative phase retrieval.

### 4.3. Iterative phase retrieval techniques

---

As it was mentioned before initial guess of the support  $S(\mathbf{r})$  is estimated from the auto-correlation function (Eq. (4.28)). Additionally, it can be smeared by the convolution with the Gaussian function. Then the updated support  $S(\mathbf{r})$  is thresholded by the normalized amplitude of the object which gives the general way of support – the binary mask

$$S_1(\mathbf{r}) = \begin{cases} 1 & \text{if } \frac{|A_1(\mathbf{r})|}{|\max\{A_1(\mathbf{r})\}|} \geq t \\ 0 & \text{if } \frac{|A_1(\mathbf{r})|}{|\max\{A_1(\mathbf{r})\}|} < t \end{cases} , \quad (4.32)$$

where  $t$  is the threshold value. Typically, it is 10 – 20% of the maximum value of the image amplitude. Usually, the SW algorithm is applied every 10/20/30 iterations during phase retrieval.

The correct support is an important part of the reconstruction process. Of course, exact support can be measured before X-ray experiments with the sample but usually, it is not possible or the object is too small. Iterative updating of the support (as in the SW algorithm) allows the support to shrink eventually to the real boundaries of the object.

#### 4.3.6 Solvent Flipping

Solvent Flipping (SF) is an improved modification of ER phase retrieval algorithm and allows to achieve faster convergence. In contrast to the ER method which computes the wave field (Eq. (4.25) by multiplication by the support mask (support projection  $\pi_S\{A(\mathbf{r})\}$ ), in SF the support reflection is used

$$A'_1(\mathbf{r}) = 2S(\mathbf{r}) \cdot A_1(\mathbf{r}) - A(\mathbf{r}) , \quad (4.33)$$

where  $A(\mathbf{r})$  is the initial complex-valued field of the iteration. Such an approach helps to converge faster than ER but it still has the same disadvantage of falling into local minima.

There are other phase retrieval algorithms, such as difference map (DM) or relaxed averaged alternating reflections (RAAR) that are also actively used.



## Chapter 5

# Single Particle Imaging with XFELs

CXDI provides the possibility of imaging cells, cell organelles, viruses, and biological materials. CXDI in forward direction is aimed at studying non-crystalline objects and BCDI is analyzing the crystalline objects. However, these methods require a series of exposures and certain classes of samples could not survive such X-ray radiation. The focus of this Chapter will be on studying of radiation-sensitive objects – such as single biological particles, viruses, and macromolecules.

To avoid sample destruction and still perform multiple illuminations, cooling to cryogenic (cryo) temperature has been introduced in, for example, Cryo-EM technique [64–66]. It permits the reduction of radiation damage to a specimen. In most cases, this method works well and the structure of the capsid of viruses can be determined, however, it is difficult to get the internal structure. In addition, when using Cryo-EM, the samples must be cooled to cryo-temperatures which makes it difficult to understand the functioning of biological samples in the natural environment.

A different approach should be used in this case. Significant success was reached with the use of intense femtosecond pulses from XFELs. They are able to outrun the structural radiation damage (breaking of the chemical bonds), found at the intensities of XFELs [101, 102], and therefore allow, in principle, atomic-resolution structure determination of isolated macromolecules in their native environment [33]. It resulted in the development of the Serial Femtosecond Crystallography (SFX) technique used with biological objects.

SFX [34, 103, 104] is based on combining proteins and macromolecules into large crystals. This approach allows to resist X-ray damaging radiation and to enhance the scattering signal at high resolution. To do so, intense X-ray pulses from an XFEL are used as they can provide several exposures on the sample and, thus, the high-resolution structure of the crystal can be obtained.

The experimental set-up used in SFX is shown in Fig. 5.1. Intense ultra-short X-ray pulses (with femtosecond duration) are hitting the crystals coming from the liquid microjet where they are randomly oriented. Individual diffraction patterns are collected on the detector. Af-

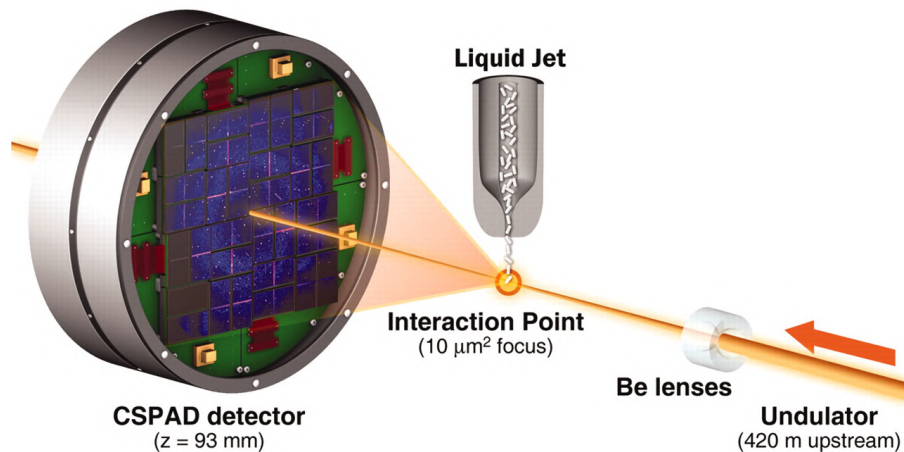


Figure 5.1: Experimental set-up for SFX experiment. Crystals of macromolecules or proteins are injected and are hit by the XFEL pulse. The diffraction pattern is recorded on the detector. Adapted from [104].

ter they are properly oriented and merged into one 3D intensity volume, the list of intensities depending on the  $hkl$ -indices (see Sec. 3.4) is retrieved. Different software was developed and is available in order to deal with the big amount of data, the most commonly used is CrystFEL [105].

Crystallization of biological particles is not always possible so SFX could not always be used. A method of reconstruction of non-crystalline single biological particles based on diffraction images obtained by the CXDI method is called Single Particle Imaging (SPI) [106, 107]. The key ingredient to studying small biological samples, viruses, and proteins in SPI is modern XFEL sources with their spatially coherent properties. They can produce  $\sim 10^{13} - 10^{15}$  photons/ $\mu\text{m}^2$  per pulse [108]. This amount is more than enough to damage the biological sample and record the diffraction signal beforehand. Contrary to the SFX method, SPI uses reproducible copies of the studied particle without crystal formation or freezing to cryo-temperatures. It allows studying single particles in a more natural environment and opens new possibilities. The recent worldwide outbreak of the COVID-19 pandemic [109] has indicated an urgency for the development of complementary imaging techniques for the study of virus structures at high resolution, and SPI provides such an opportunity.

Working with the data, its analysis from the SPI experiments in practice, and performing SPI simulations will be the main focus of this Thesis and will be described in detail in further sections.

## 5.1 SPI experiment

The typical experimental set-up in SPI experiments is shown in Fig. 5.2. The idea is rather simple – many specimens of the investigated particle are injected into the x-ray beam in ran-

## 5.1. SPI experiment

---

dom orientations. Particles are destroyed during the scattering process, however, diffraction patterns are collected before the atoms have time to noticeably change their position and correspond to the structure before radiation damage takes place – the so-called “diffraction-before-destruction” approach [33, 110].

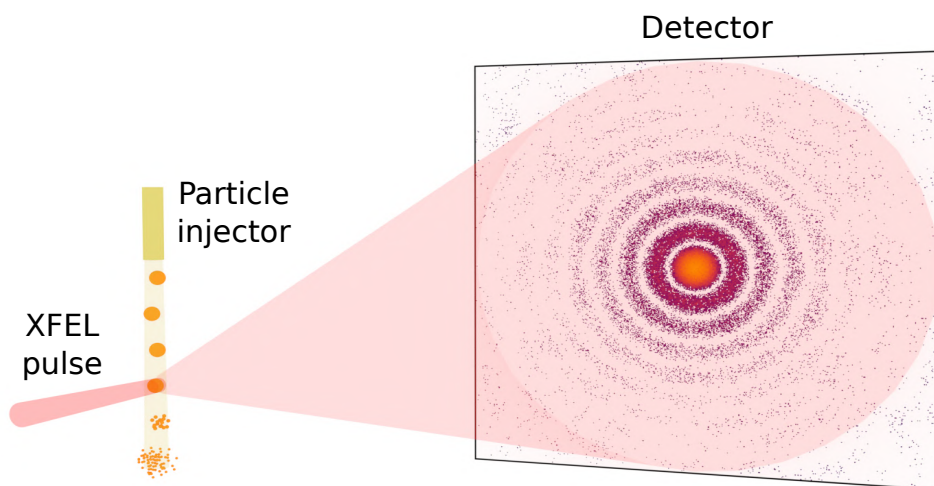


Figure 5.2: Experimental set-up for SPI experiment. An XFEL pulse hit a single particle coming out of the particle injector. After the interaction, the particle is destroyed but the diffraction pattern is recorded by the detector.

The SPI experiment requires a source of coherent radiation – XFEL, an injector that injects samples, and a detector that measures the intensity of the diffraction pattern. The intensity of the incident radiation is much higher than the intensity of the scattered radiation, so the detector has a hole in the center for the direct beam. In experiments on XFELs, it is impossible to use a direct beam stub in front of the detector as it will produce parasitic scattering illumination and noise on the detector. So the direct beam passes through the hole in the center of the detector. The radiation is focused as much as possible to increase the scattered intensity from the sample on the detector. Samples are injected into the laser beam in random orientations.

The SPI experiment is performed on XFEL, and its repetition rate plays an important role, as the detector used in the SPI experiment should be consistent with the number of pulses produced by the source in order to use the whole capabilities of the facility. Various detectors are used: for example, CSPAD [111], pnCCD [112], AGIPD [113]. Detectors must satisfy certain conditions in order to be used in X-ray experiments. They have to have high sensitivity and low noise level for individual photons detection. As it was said before in Sec. 4.2, obtained resolution highly depends on the recorded scattering angles. That is why detectors must have physical dimensions sufficient to detect a signal at large angles. In addition, the

intensity of the signal near the center of the detector is many orders of magnitude higher than the intensity at the boundaries. So the detectors must have a high dynamic range.

The detectors consist of smaller parts – panels. The panels can be configured according to the needs of the experiments. The presence of gaps between detector panels should be also considered in data analysis. Another thing that should be taken into account is that the XFEL pulse in the focusing plane has long tails with low intensity. The most likely process is that the sample will be hit by the tail of the beam with lower intensity than hitting the central part of the pulse with high intensity. This introduces a big divergence of diffraction patterns in terms of the measured intensity. It is a big task at the data analysis stage – which diffraction patterns can be used to reconstruct the electron density of the sample.

To join forces in handling big SPI experiments and to push the methodology further, a dedicated SPI consortium was formed at the Linac Coherent Light Source (LCLS) [114]. Several results were reported from the SPI consortium, covering both hard and soft X-ray experiments and focused primarily on well-characterized viruses with sizes of a few tens of nanometers. Different analysis methods have been applied to virus structure determination [35, 36, 38, 115–119].

## 5.2 Radiation damage

One of the key questions of the SPI experiments – which radiation dose the biological particle can survive before it is destroyed. Computer simulations [108] were made to study the radiation dose which allows performing SPI experiments with samples without crystallization.

Besides the ability to penetrate the studied samples, X-rays are also able to ionize elements of the sample and such effects as producing photoelectrons and Auger electrons can result in the deformation of the protein structure. Plus, the Coulomb explosion can break the chemical bonds and destroy the object. The timescale of such processes is femtoseconds. Fig. 5.3 is showing T4 lysozyme explosion [33]. The calculations indicated that after the sample was radiated by the X-ray pulse (12 keV photon energy) with the full width half maximum (FWHM) of 2 fs, it disintegrated later in time but the diffraction pattern corresponded to the undamaged structure was already recorded.

The correlation between the possible resolution and radiation dose was shown in Ref. [120]. The two main aspects were considered: the required dose for imaging and the maximum tolerable dose that the sample can survive. The experiment can be successful if the final dose is between the required dose for imaging and the maximum tolerable dose.

The required dose for imaging was estimated under the consideration of the one-voxel experiment. The X-ray scattering cross-section of this voxel with the size  $d$  into a detector is

## 5.2. Radiation damage

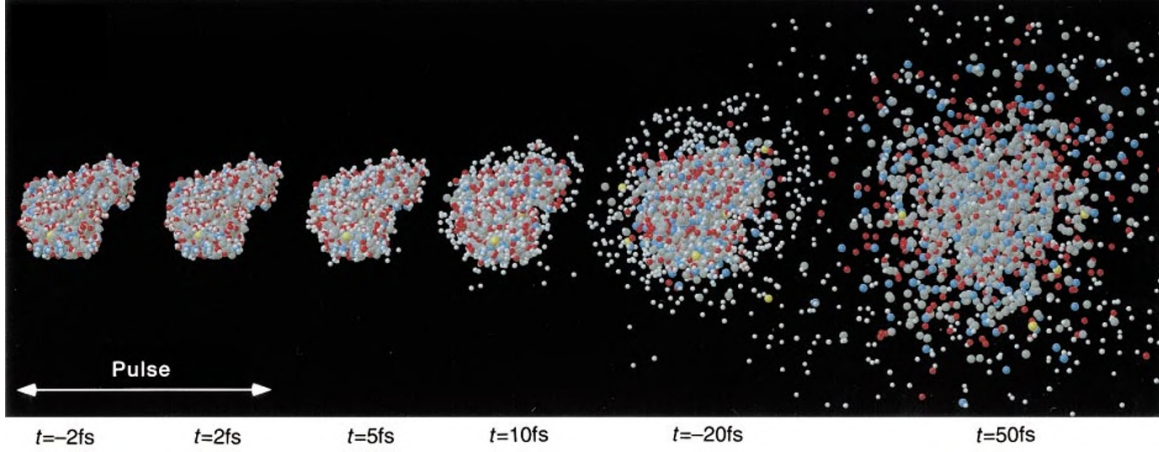


Figure 5.3: Simulations of radiation damage of T4 lysozyme. X-ray radiation was set to  $3 \times 10^{12}$  photons per 100 nm diameter spot. Adapted from [33].

then

$$\sigma_s = r_e^2 \lambda^2 |\rho|^2 d^4, \quad (5.1)$$

where  $r_e$  is the classical electron radius,  $\lambda$  is the X-ray wavelength,  $\rho$  is the electron density.

The dose absorbed by the object can be expressed as the energy deposited per unit mass of the surface

$$D = \frac{\mu \hbar \omega}{\varepsilon} N, \quad (5.2)$$

where  $\mu$  is the linear absorption coefficient from Eq. (3.47),  $\hbar \omega$  is the X-ray energy,  $\varepsilon$  is the object mass density and  $N$  is the incident number of photons per unit volume at the surface.

To calculate the required dose for imaging from the voxel element, the incident photon flux  $N$  from Eq. (5.2) can be written as the ratio of the scattered X-rays into the detector  $P$  and the cross-section  $\sigma_s$

$$N = \frac{P}{\sigma_s}, \quad (5.3)$$

then the required dose is finally

$$D_{req} = \frac{\mu \hbar \omega}{\varepsilon} \frac{P}{\sigma_s} = \frac{\mu \hbar \omega}{\varepsilon} \frac{P}{r_e^2 \lambda^2 |\rho|^2 d^4}, \quad (5.4)$$

so the required dose  $D_{req} \propto d^{-4}$  can be considered as resolution. Another question is how many detected photons  $P$  are needed to overcome the noise. The Rose criterion [121] states that the signal should be five times stronger than the rms noise of the background, the so-called signal-to-noise ratio (SNR). It is set that  $P = SNR^2 = 25$  photons is necessary to satisfy Rose criterion.

Another dependence that can be observed is  $N_{req} \propto \lambda^{-2}$ . The longer the wavelength of X-ray, the less the required fluence is needed.



The tolerable dose cannot be simply estimated and is highly dependent on the sample. It can be calculated experimentally and the following relationship was observed [120]

$$D_{tol} = 10^{17} \left[ \frac{J}{kg \cdot m} \right] \cdot r, \quad (5.5)$$

where  $r$  is the obtained resolution. The calculated examples led to the conclusion that for the frozen-hydrated biological sample the resolution cannot be better than  $r = 10$  nm with the required dose  $D_{req} = 10^8$  Gy per nm. However, a 10 nm limit can be overcome, by using, for example, many copies of the same biological particles as it is realized in SPI.

### 5.3 Challenges and limitations

Nevertheless SPI is considered to be one of the flagship experiments using XFEL pulses to image single particles, progress toward high-resolution 3D electron density images of non-crystalline biological samples has been slow compared to SFX [37, 122–125].

There are several reasons for the lower resolution achieved in SPI experiments in comparison to the SFX technique. The most important include: a lack of crystalline periodicity to amplify the signal, weak single particle signal from non-periodic nanoscale objects compared to the instrumental background, a limited number of usable data frames collected, and heterogeneity of the samples. Radiation damage processes initiated by X-ray photoionization also play an important role at high resolution. One may expect that in order to further enhance resolution, one could increase the power of XFEL pulses to boost the SNR for bio-particles such as single proteins or virus particles that scatter very weakly. Unfortunately, increased XFEL fluence strips electrons from atoms more efficiently and the scattered power from the bound electrons (that contain structural information) does not increase in proportion to the X-ray fluence [108, 126]. Decreasing the pulse length below one femtosecond may help to outrun Auger decay, but a further decrease of the pulse duration may not lead to the desired suppression of all ionization channels [108].

Besides these fundamental limitations, there are other limiting factors. The background scattering originating from the beamline obscures the weak scattering signals from biological samples, and fluctuations in the beam position and intensity can cause additional challenges that need to be accounted for in reconstruction algorithms. Moreover, single particle sample delivery remains a challenging topic. In addition, despite the strong development of detector technology [127], the dynamic range of the present detectors is still not sufficient to capture the full diffraction pattern in a strong single shot.

**Hit rate and sample delivery** One of the most difficult sides of the SPI experiment is sample delivery and interaction with the X-ray beam. Only diffraction patterns that correspond

### 5.3. Challenges and limitations

to single particle – X-ray pulse interaction, can be used in the SPI data analysis. These diffraction patterns are called single hits and the percentage of single hits over the total number of diffraction patterns taken during the experiment is called the hit rate.

The hit rate in a typical SPI experiment is usually no more than 1% as the sample delivery and proper interaction with an X-ray beam is a very sensitive process and has to be thoughtfully controlled. Most of the diffraction patterns can correspond to different scenarios: particle flow and X-ray beam are not aligned – empty diffraction patterns, the cluster of single particles hit the X-ray beam – multiple hit, other particles (from the solvent, for example) but not the particle of interest hit the X-ray beam (Fig. 5.4).

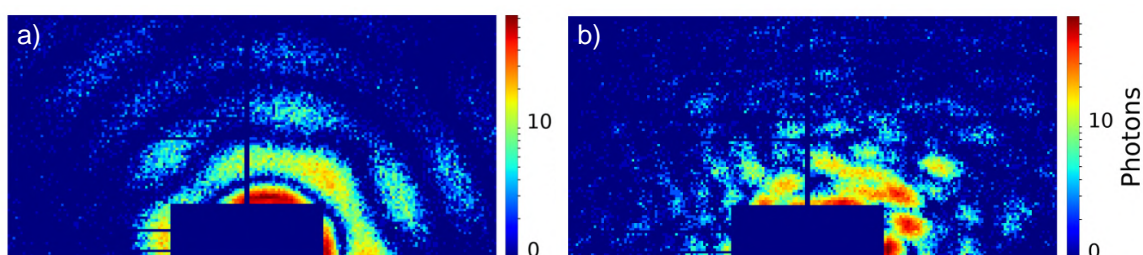


Figure 5.4: Examples of diffraction patterns recorded in SPI experiments by half of the detector. (a) Diffraction patterns from icosahedral particle – single hit. (b) Diffraction pattern from the agglomeration of the particles – non-single hit.

The first technique used for sample delivery in SPI was nano-electrosprays [122] (Fig. 5.5 (a)). The idea behind it is to drive samples through the pumping devices of the aerodynamic lens stack into the solution-free droplets containing a single particle. The method is called electrospray ionization [128] and implies solution pumped through the capillary and high voltage applied so the Taylor cone of charged liquid appears at the end of the capillary. Then the droplets are emitted. Typical sizes of the droplets are hundreds of nm from nano-electrospray nozzles of 1 – 10  $\mu\text{m}$  in diameter [129]. The important advantage of the aerodynamic lens interface is the absence of clogs which is crucial for the effective utilization of an expensive XFEL beamtime.

Another device that can be used in sample delivery is a Gas dynamic virtual nozzle (GDVN) [130]. A sheath gas (see Fig. 5.5 (b)) is focusing the liquid into the microjet. Such GDVN allows using channels of 50  $\mu\text{m}$  diameter preventing clogging for several hours and has proven to be a stable method of sample delivery.

**Detector gaps and background** Detectors used in SPI experiments usually consist of panels. In Fig. 5.4, the empty areas with no signal are seen, as the result, the final diffraction pattern does not contain photon signal everywhere and that is a complication for further data analysis. The consequence of the detector consisting of panels is that the position of the panels should be precisely calculated. The properly aligned panels give the correct diffraction

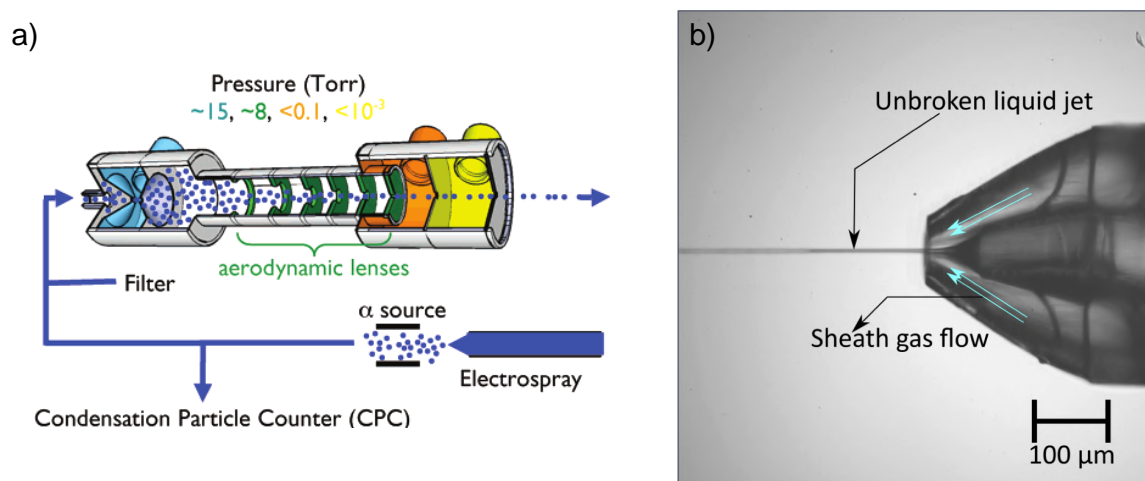


Figure 5.5: Examples of sample delivery systems in SPI experiments. (a) Electro-spray-generated aerosol. Adapted from [122]. (b) Gas dynamic virtual nozzle in operation. Adapted from [130].

pattern. This could be done at the beginning of the SPI experiment using specific samples with the known features present on diffraction patterns.

Another important feature of the detector is its dynamic range. It is seen from Fig. 5.4 that the difference in very bright and very weak signals is huge. The detector must be able to detect both thousands of photons and one photon. Each pixel of the detector accumulates an electronic charge by the capacitor. The full well capacity of the detector is primarily defined by the size of the pixel and electronic elements connected. Another challenge is single photon detection. It is shown that even a weak scattered signal recorded by the detector in SPI experiments can be used in data analysis [131], that is why detectors should be able to distinguish one-photon signal over the noise. Ability to switch between "one photon" detection mode and "one thousand photons" detection mode is realized with the adaptive gain of the readout amplifier in AGIPD [113]. Detection of thousands of photons requires the gain of the readout amplifier to be small. On the contrary, detection of one photon requires a high readout amplifier. The switching between gains is implemented in the AGIPD.

Detectors are recording not only signals from the sample but also from the set-up, solutions, and other experimental elements. Several ways to detect background exist and it can be subtracted at different stages of SPI data analysis. Primarily, it is defined after 2D diffraction patterns are assembled into one 3D volume. It is possible to compare obtained volume with the diffraction from the known particle (for example, spherical one [35]). Another way is to estimate the background from the angular average diffraction profile [132]. It is also possible to use background-aware algorithms at the stage of phase retrieval, such as Difference Map [133] or Richardson-Lucy deconvolution algorithm [134]. The latter is used in the framework of this Thesis.

The data analysis pipeline in SPI experiments will be discussed in detail in the following Section.

## 5.4 Data analysis pipeline

The data analysis pipeline of SPI experiments is aimed at obtaining the 3D structure of the object on the basis of diffraction patterns collected in the experiment. The first pipeline was proposed in Ref. [107] and is demonstrated in Fig. 5.6. The structure of the object is determined by the electron density distribution and measured 2D diffraction patterns contain information about the reciprocal space which is a 3D Fourier transform image of the electron density. By determining the orientation of the series of 2D diffraction patterns, they can be formed into one 3D volume in reciprocal space. Then the scattering phase values need to be recovered which allows obtaining the electron density distribution through the Inverse Fourier transform.

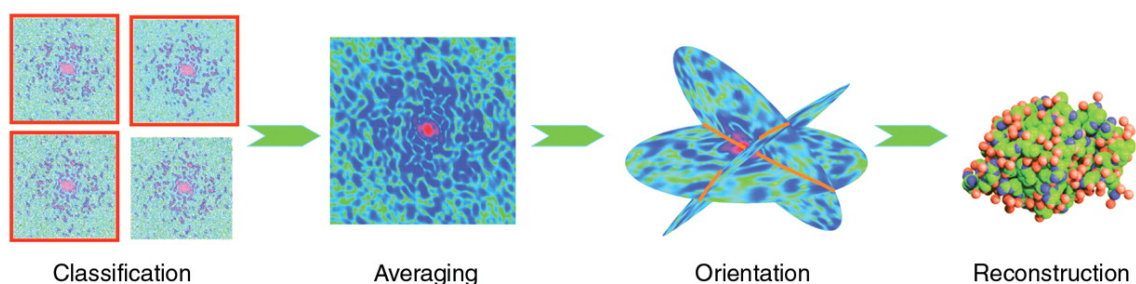


Figure 5.6: First proposed SPI data analysis pipeline demonstrated in Ref. [107].

To account for the specifics of the experiment, additional steps were included in the data analysis procedure [35, 38, 135]. First, due to the fact that only a small fraction of the images contains diffraction patterns, empty patterns are filtered out from the analysis. Second, before starting the analysis of the diffraction patterns, the position of the center of the diffraction pattern relative to the detector position should be precisely determined. Then, among all diffraction patterns, the ones which contain scattering signal from a single virus are selected. For this purpose, the size of the samples is estimated from the diffraction patterns and those with a reasonable size distribution are selected. Then, the selected patterns are classified according to the features of the diffraction patterns which are related to the features of the sample structure in real space. Thus, the single hits are distinguished. From the experience, a significant part of the diffraction patterns refers to impurities or water droplets, and a part of patterns contains diffraction from several objects combined. The quality of diffraction patterns classification directly influences the obtained structure resolution – if the quality of single hits classification is insufficient, structure reconstruction becomes impossible.

Selected diffraction patterns of the studied object are combined in reciprocal 3D space. For this purpose, the orientations of the diffraction patterns relative to each other are determined. They are respectively related to the orientations of the object inserted into the X-ray pulse during the experiment. The orientation determination and the reconstruction

of the reciprocal space volume are based on the maximum-likelihood method [136] implemented into Expand-Maximize-Compress (EMC) algorithm [137]. After the reconstruction of the reciprocal space volume, the background signal is corrected. It is usually caused by parasitic scattering on the elements of the experimental setup. The background signal does not depend on the orientation of the sample, so the background correction is performed at 3D intensity volume in reciprocal space. At this stage, the influence of the background is averaged, and it is easier to correct it.

Next, the phases of the diffraction patterns are reconstructed in reciprocal space and the structure is reconstructed in real space. For this purpose, iterative phase retrieval algorithms are used [93, 100]. Finally, the resolution of the resulting electron density distribution is evaluated.

Thus, the structure reconstruction pipeline includes the following steps (Fig. 5.7):

- filtering of empty diffraction patterns;
- position of the center of the diffraction patterns determination;
- evaluating the size of the object according to the diffraction patterns;
- filtering of diffraction patterns by size;
- clustering and classification of diffraction patterns by object type;
- merging diffraction patterns into a diffraction 3D volume in reciprocal space;
- correction of the background signal of the diffraction 3D volume;
- reconstruction of the scattering phases and reconstruction of the object structure;
- resolution estimation.

The mentioned steps can be realized with different methods and approaches. The general pipeline should consist of the most important steps: single hit classification, orientation determination and phase retrieval. Within other steps of the pipeline, a certain degree of freedom is possible. In further Chapters, we will focus on different ways of the realization of the data analysis procedure.

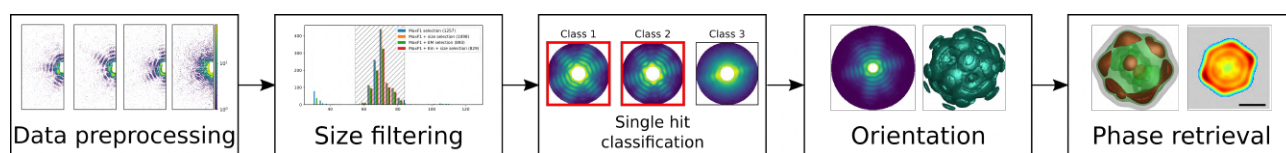


Figure 5.7: Schematic interpretation of key steps in SPI data analysis pipeline proposed in Ref. [38].

The presented in Fig. 5.7 pipeline was tested with the data of several SPI experiments on XFELs: LCLS (Stanford, USA) and European XFEL [138]. Processing the experimental

SPI data according to the presented scenario allowed obtaining a significant improvement in the structure recovery in SPI [35, 38]. The software was developed to implement all the stages of the pipeline, and a platform for automated data processing of SPI XFEL experiments was created [135]. The platform includes containerized software that is integrated into a software pipeline which provides automatic processing from the experimental data collection to structure reconstruction. The platform can be quickly run on any computing architecture with container support (e.g., Docker [139]), as well as in computing clusters running by Kubernetes. Information about the developed software [135] and the automatic data processing platform is publicly available [140].

## 5.5 Data classification in SPI

The task of single hit diffraction patterns classification is one of the most important parts of the SPI data analysis pipeline. We will focus on this task in the following Section. It will be based mainly on the following Ref. [141–143]. Nowadays, several methods for classifying in SPI experiments are known and widely used. The basic principle of most of them is classifying according to the type of structure observed on diffraction patterns.

### Existing methods in SPI data classification

Proteins in different states were studied in Ref. [144]. The approach to classify diffraction patterns according to two different conformations and also transition states was the Diffusion Map (DM) method which was based on the manifold concept. As the result, proteins in two different states can be distinguished with high accuracy. However, the patterns of the protein in the transition state were often mistaken for the patterns of the other states. According to the classification results, only about 20% of images in the transition state do not intersect with groups of patterns in one of the basic states. The idea is to project the object as a point in the N-dimensional space where relativity is described with a distance. DM method implies that data points are mapped by the eigenvalues and eigenvectors of a normalized transition matrix which is connected to the matrix describing connectivity between the points on the manifold. Authors of Ref. [144] used simulated data for their studies.

The article [145] presents classification based on the method of spectral clustering. The authors performed their testing on the experimental XFEL data from LCLS. Spectral clustering is an unsupervised method, as a Principle Component Analysis (PCA), which catches the non-linear correlations in the data set in contrast to standard PCA. The authors compared spectral clustering, standard PCA and manual classification results, the main result shows around 90% agreement between the classification based on the spectral clustering

method and the result of the manual classification. Comparison to the PCA shows better results, it is more suitable for the nonlinear structure of XFEL data sets.

The step of single hit classification may be significantly improved by the application of machine learning approaches. In recent work [146], supervised machine learning was used to map patterns into a low dimensional manifold representation in which the authors were able to separate single from non-single hits through transformation into a bimodal distribution. In general, a variety of methods to classify data in SPI exists. The next Sections will be focusing on the machine learning methods in diffraction patterns classification.

## Machine learning methods in SPI data classification

Machine learning methods are often effective in data analysis in SPI. The task can be formulated as follows:

1. Cluster analysis or the determining of groups within data where elements are more similar to each other than to elements of the other groups, without using additional information about the data elements;
2. Supervised classification algorithms or classification of a complete set based on a training data set that contains manual assignments of the data type.

### 5.5.1 Principal Component Analysis

As the diffraction patterns contain thousand of pixels, there is hardly a simple way to describe one pattern with a value or a function. If one can reduce the dimensionality of the data set without losing important feature information, it is possible to perform analysis and even visualize the data.

One of the dimensionality reduction methods is Principal Component Analysis (PCA) which is based on the feature extraction from the data. In contrast, feature selection methods are based on determining the components that carry the basic information about the structure of the sample and the rest components are not used. When features are extracted, a new basis is determined on the full space of features. Features in the new basis are a combination of features in the old basis.

In general, principal components are the set of unit vectors which form the orthonormal basis where the data dimensions are not linearly correlated. Each unit vector is a direction which fits the data the best. The best fit is defined by the minimization of the average squared distance from the points to the line of the direction.

PCA is based on the idea to project a multidimensional feature space into a new basis which is why it is also called a projection method. At the same time, principle components

can be shown as eigenvectors of the data covariance matrix. So the task of PCA can be formulated as a calculation of the eigenvectors of the data covariance matrix.

For example, we have the original space  $\mathbf{X}$  of  $n \times p$  dimensions, where  $p$  is the number of features,  $n$  is the number of experiments. We want to get the projection of each row vector  $\mathbf{x}_{(i)}$  on the space with lower dimensionality  $l$  to a new vector  $\mathbf{t}_{(i)}$  through a coefficients or weights of transformation  $\mathbf{w}_{(k)}$

$$t_{k(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)}, \quad (5.6)$$

where  $i = 1, \dots, n$ , and  $k = 1, \dots, l$ , and  $l < p$ .

New features of lower dimensions can be extracted directly. Then the algorithm starts with an empty set and features are added. In the next step, the features are determined, the addition of which minimizes the error value as much as possible. The algorithm stops when adding the remaining features reduces the error to less than the threshold value. It can be done the other way around. In the inverse extraction, the algorithm starts with a complete set of features, at each step the feature are excluded. They are the features that reduce the error as much as possible or increase it weaker than the threshold value. The algorithm stops when feature exclusion increases the error significantly.

PCA automatically determines the basis for feature extraction without any additional information. The components of the basis (principal components) are selected according to the maximum variance criterion. The main or first component is defined in such a way that the projection of the full set onto it has the maximum variance. To obtain a unique solution, the condition for the weight vector  $\|\mathbf{w}\| = 1$  is imposed. Then the main weight vector  $\mathbf{w}_{(1)}$  is determined by the formula

$$\mathbf{w}_{(1)} = \arg \max \left\{ \sum_i (t_{1(i)})^2 \right\} = \arg \max \left\{ \sum_i (\mathbf{x}_{(i)} \cdot \mathbf{w})^2 \right\}, \quad (5.7)$$

which can be rewritten in the matrix form

$$\mathbf{w}_{(1)} = \arg \max \left\{ \|\mathbf{X}\mathbf{w}\|^2 \right\} = \arg \max \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right\}, \quad (5.8)$$

where T is the transpose operation. When the main principal component could be found from Eq. (5.6):

$$t_{1(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}. \quad (5.9)$$

Principal components are the eigenvectors of the input covariance matrix  $\mathbf{X}^T \mathbf{X}$ , arranged in descending order of eigenvalues. The first principal component corresponds to the direction with the maximum variance, the second corresponds to the maximum variance of the directions orthogonal to the first principal component, and so on. If a significant part of the data variance belongs to the first two principal components, the projection of the data onto



the plane of the first two principal components allows one to visualize the data, analyze it, divide the data into groups, find outliers, and so on. The example of PCA application to the real diffraction data was published in the papers [35, 147] and illustrated in Fig. 5.8.

The set of feature vectors was projected onto the plane of the first two principal components, where each vector corresponds to a point. According to the distribution of points on the plane, the possibility of dividing the feature vectors into groups was visually evaluated. The set that was defined as a single hit diffraction pattern  $Set_{14k}$  appeared as a packed region and was used further in the data analysis pipeline. The final reconstruction of the virus studied in Ref. [35] was in a good agreement with the expectation and showed resolution below 10 nm.

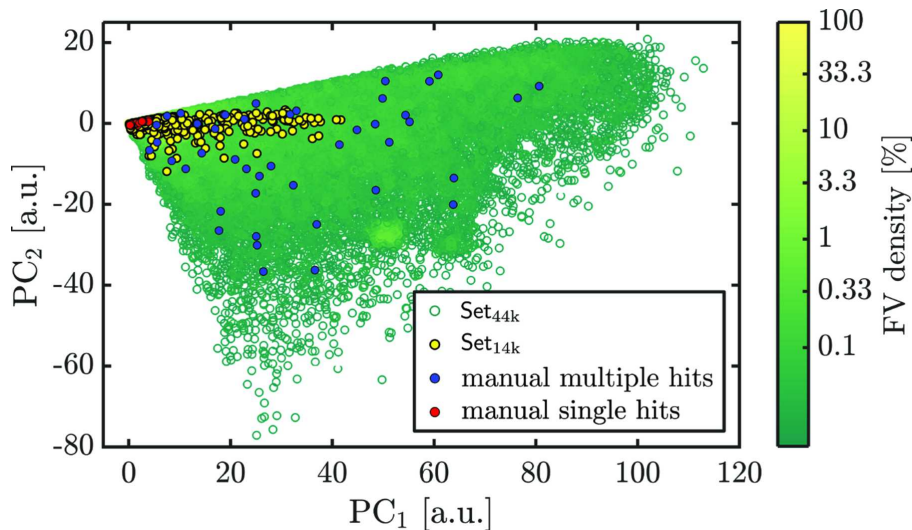


Figure 5.8: Diffraction patterns representation on the plane of principle components. Each diffraction pattern is shown by the feature vector, different sets are shown with different colors. Adapted from [35].

## Data clustering

Cluster analysis methods allow distribution of the input data into a certain number of groups, the cluster parameters are determined automatically from the input without any supervision. In the task of SPI classification, we almost always need the binary classification: single and non-single hit diffraction patterns. So we are going to describe a data set by determining the characteristics of each element of the data set.

The ultimate goal of clustering diffraction images is to divide the complete set into groups according to the type of objects under study. In addition, cluster analysis can be used to detect outliers or to compress data sets by excluding similar items.

### 5.5.2 K-means clustering

The most commonly used clustering method designed to automatically divide elements into a given number of groups is the method of k-means. In this approach, the total square deviation of the elements of each group from the center of this group is minimized. The original idea was proposed by H. Steinhaus in 1957 [148] and was formalized by J. MacQueen in 1967 [143].

Let's consider the set of vectors  $X = x_N^{t=1}$  and the set of basis vectors (or the set of clusters we want to distribute the data in)  $m_j$ , where  $j = 1, \dots, k$ . From the definition of the k-means, for each vector  $x^t$  there is a basis vector  $m_i$ , so that

$$\|x^t - m_i\| = \min_j \|x^t - m_j\|. \quad (5.10)$$

Thus, each basis vector  $m_i$  can describe a group of respective  $x$ . The goal is to find values of  $m_i$  from the condition of minimizing variance

$$E(\{m^i\}_{i=1}^k | X) = \sum_t \sum_i b_i^t \|x^t - m_i\|^2, \quad (5.11)$$

where

$$b_i^t = \begin{cases} 1 & \text{if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\ 0 & \text{else} \end{cases}. \quad (5.12)$$

So the minimization of variance corresponds to the optimal set of basis vectors  $m_i$ . To obtain this set, iterative refinement techniques are used. The iteration starts with a random assignment of  $m_i$ . Then coefficients  $b_i^t$  are calculated according to Eq. (5.12). It is seen, that if  $b_i^t = 1$ , then vector  $x^t$  contains in the cluster  $m_i$ . Then the minimization of variance is applied (Eq. (5.11)), taking the derivative over  $m_i$  we will the condition for the minimum of the variance

$$m_i = \frac{\sum_t b_i^t x^t}{\sum_t b_i^t}. \quad (5.13)$$

The basis vector is then updated as the mean of the vectors  $x^t$  assigned to the respective cluster. The iteration continues till the values of  $m_i$  converge.

The initialization process plays a role in k-means clustering. There are two major ways to do it [149]: the Forgy partition implies choosing of  $k$  random elements of data as initial  $m_i$ ; the Random partition implies random distribution of elements in  $X$  to the clusters and then mean values inside each cluster are assigned to  $m_i$ . Applying k-means to the diffraction patterns (without any dimensionality reduction) does not show a satisfying result. Classification methods for structure types based on the method of k-means noticeably lose to other methods used in SPI.

### 5.5.3 Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm is based on the finding maximum likelihood of the parameters of the data set [136, 150]. This algorithm allows for unsupervised clustering of data when neither initial data assignments to clusters nor cluster parameters are known. The data set is distributed into a pre-defined number of clusters and cluster parameters are retrieved automatically at the same time. Later, manual input is used to classify each of the cluster models. In SPI it can be based on symmetry considerations or expected fringe diffraction patterns, in order to perform data selection. This algorithm is commonly used in Cryo-EM for unsupervised single-particle clustering (see, for example, software RELION [151]).

Unsupervised EM-clustering is an iterative process; the algorithm starts from a random model for each class. At each iteration, the probability for each diffraction pattern to belong to a certain class is calculated. To accommodate for random particle orientations in the SPI experiment, the cluster model is compared with the 2D diffraction pattern rotated in-plane. After probabilities are evaluated, a new cluster model is calculated by weighted averaging of patterns that belong to each cluster in each orientation. The weights are defined by computed probabilities. Poisson noise model is used to form clusters. In fact, the algorithm imitates the Expand–Maximize–Compress (EMC) algorithm [136, 137] which will be discussed in the next Section, but, instead of rotation in 3D space, an in-plane rotation and cluster distribution are used [152]. When the EM-algorithm converges, the supervised class selection takes place. The cluster models that correspond to single hits of an investigated particle are selected manually by an expert.

This particular method is used in the framework of this Thesis, and the result of its application will be shown in Chapter 6.

## Neural Networks

The task of single hit diffraction patterns classification can be formulated as an image classification problem and solved by using Neural Networks (NN).

### Artificial Neural Networks

One of the areas of machine learning is Artificial Neural Networks (ANNs). ANN was motivated by the biological networks of neurons in the brain. Such networks consist of many nodes – neurons which are connected, and each connection transmits a signal to other neurons. Such a connection is called edge. This neuron receives a signal, processes it and transmits it further, to the neurons connected to it. ANNs are used to solve problems of pattern recognition, clustering and classification methods, and others.

The signal received by the neuron is a real number, and the output of the neuron is calculated by some non-linear function of the sum of received signals. The edges and neurons of ANN have weights, and during the learning process, these weights are adjusted. The weight can increase or decrease the influence of the signal at the edge. The sum of signals received by the neuron can be denoted as

$$S \equiv \sum_i w_i x_i + b, \tag{5.14}$$

where  $w_i$  – weight coefficient of  $i$ -neuron,  $x_i$  – its signal,  $b$  – some constant.

Neurons are usually connected into layers: input, hidden, and output (see Fig. 5.9). The dimension of the input layer is the same as the dimension of the input data. If the task for ANN is classifying the data, the number of neurons in the output layer is equal to the number of the desired classes. The classification result is determined by the number of the output neuron on which the largest signal was received.

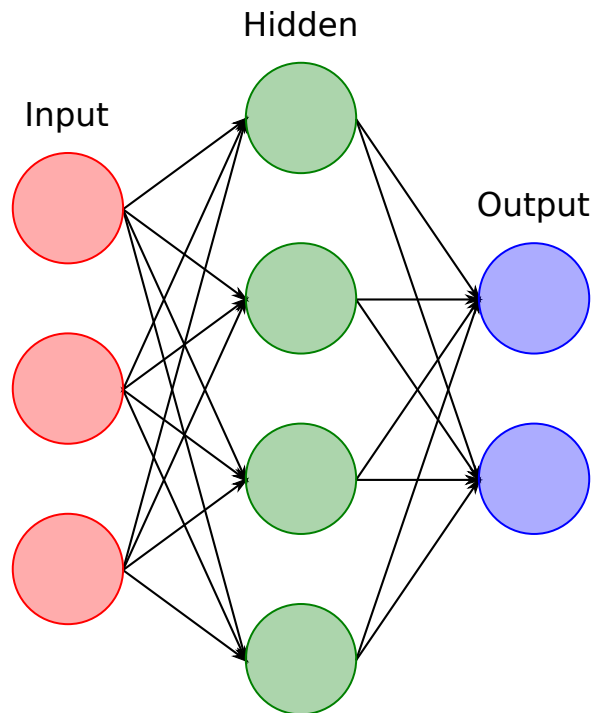


Figure 5.9: A simple example of ANN architecture. It consists of neurons and connections – edges. Neurons are connected into layers: input, hidden, and output.

### 5.5.4 Convolutional Neural Networks

In the computer vision domain, Convolutional Neural Networks (CNNs) have become the de-facto state-of-the-art in image classification [153], object detection [154], and image segmentation [155]. Thus, it is unsurprising that CNN-based solutions have been recently

successfully applied in our domain, specifically in the classification of diffraction patterns in tomography experiments at synchrotron sources [156], coherent diffraction imaging experiments at synchrotron facilities [157, 158], and at XFELs [117, 159].

CNNs consist of layers, the core building blocks being convolutional layers and with other layers, such as pooling layers, they form a network [160]. The layers are characterized by a set of filters or kernels, containing weights. In a convolutional layer, the convolutions of the input with different filters are computed.

The idea is to extract different features of the input image in order to classify it. These invariant features are then passed to the next layer. The features in the next layer are convoluted with different filters to extract more abstract features (see YOLO architecture in Fig. 5.10 – example of CNN). In this way, the convolutional layers are sensitive to features without position reference. Important parameters of CNN are weights. They are adjusted during the training of the CNN. Choosing the specific set of weights for CNN at a certain training stage creates a model which is used for testing CNN. There are also so-called hyper-parameters. They are part of CNN and are adjusted beyond the training process. For example, the number and size of the filters in the layers and the learning rate are hyper-parameters.

The CNN algorithm is a supervised deep learning algorithm. As with any neural network, it has to learn first. Training data usually represents the set of training examples with annotations. They represent the desired output which is referred to as a ground truth. These annotations are class labels in the case of classification. For object detection, they are supplemented by the position of the object [161, 162], typically represented by a bounding box surrounding the object.

The NN is trained using a gradient descent optimization algorithm [163]. The optimization algorithm minimizes the so-called loss function by updating the parameters of the model. The loss function is a specially designed function that serves as a metric of an error between predictions of the model and ground truth, it goes to a minimum when the NN correctly classifies the elements of the training set. The gradient descent [163] is based on the idea that training starts with random values of the coefficients of the loss function, at each step, the gradient of the loss function is determined, and the coefficients are adjusted so the loss function decreases. The process continues until the loss function reaches its minimum. The learning rate is the hyper-parameter that controls how much the parameters of the model are changed in response to the model error. In the framework of this Thesis, for optimizing the loss function during training, stochastic gradient descent (SGD) algorithm [164] will be used. The weights are updated based on random subsets (batches) of the training data rather than the complete training set. The period during training when the network has seen one batch is called iteration. As a result of working, the CNN algorithm produces a function that maps input and output.

## 5.5. Data classification in SPI

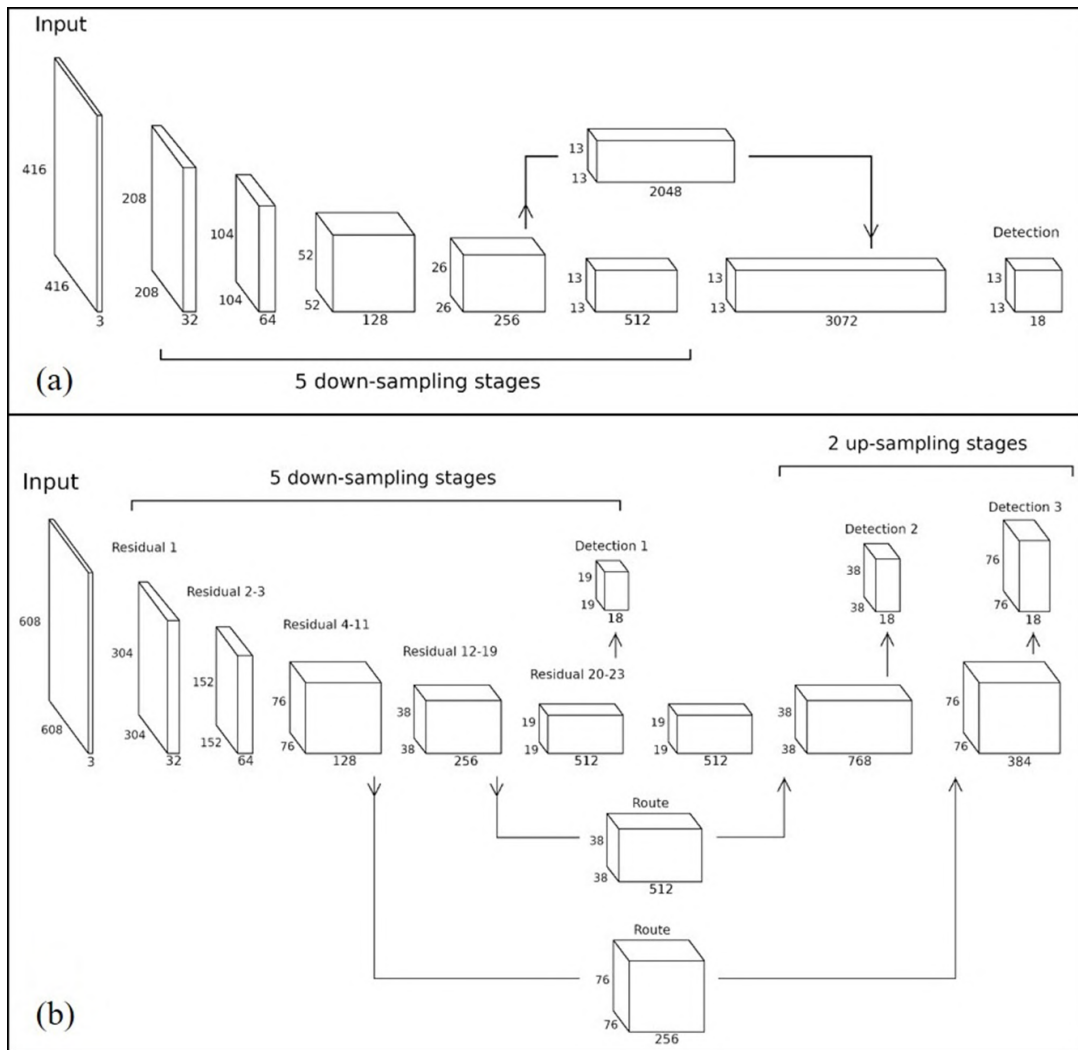


Figure 5.10: Architecture of CNN realized in fast object detector YOLOv2 (a) and YOLOv3 (b). The data undergo five down-sampling stages to limit the number of parameters. The dimensions of the input for every stage are indicated. In the case of YOLOv2 the data flow consequently from one layer to another. In the case of YOLOv3 five down-sampling stages are followed by two up-sampling stages. Adapted from [118].

Creating a representative and balanced data set to train the model is a crucial step. Many examples show that CNN output always mimics the training input, that is why it has such an important role. The training set should represent the data that the model attempts to describe as well as possible. The total number of training examples should be sufficiently large to train the CNN for the full depth.

It is possible to use so-called transfer learning with a limited amount of training data. In this case, CNN is pre-trained on some other data set which is fuller (ImageNet was used in Ref. [118]) and these pre-trained weights are used as a starting point for the specific training set with a limited amount of data. As a result of CNN training, the weights in the last layers are most affected and adjusted to the selected small training data set.

The ratio of the number of examples for each class should be close to what is expected in the experimental data set. However, in the case of SPI classification task, it could be problematic. The ratio of single hit diffraction patterns in a typical SPI experiment is less than 1% of the whole number of patterns. It is one of the bottlenecks of using supervised methods in SPI. A validation set is used to evaluate a given model, and the same requirements as for the training set apply to it as well.

## 5.6 Orientation determination of diffraction patterns

The result of the SPI data classification is the set of diffraction patterns corresponding to single hits – when the single specimen of the investigated particle hit the X-ray beam. The next step in the data analysis pipeline (see Fig. 5.7) is the orientation of 2D diffraction patterns into one 3D volume in reciprocal space.

The experimental set-up in SPI (Fig. 5.2) does not imply control over the particles after they leave the particle injector into the X-ray pulse. They hit the beam in random orientation and in order to reach the goal of SPI – to recover the electron density of the single particle – the orientation of each 2D diffraction pattern should be reconstructed. In order to solve this orientation problem, the Expand-Maximize-Compress (EMC) algorithm is used. It is based on the maximization of the likelihood [136] as it proves to work with incomplete, sparse, and noisy data. The concept of orientation determination is visualized in Fig 5.11.

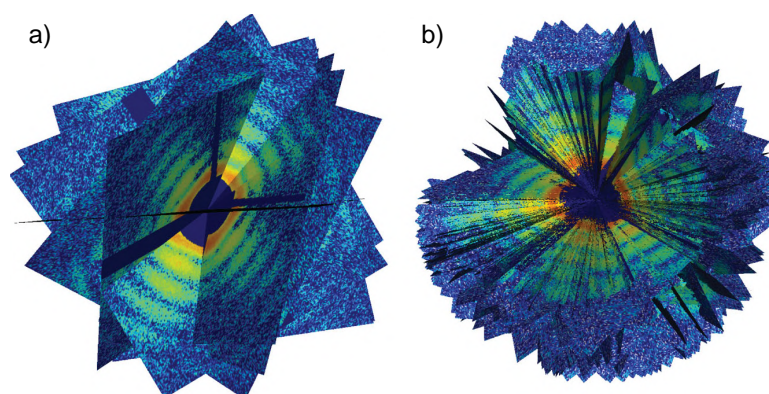


Figure 5.11: The EMC orientation determination example. (a) The first 10 diffraction patterns assembled into a 3D intensity volume. (b) 198 patterns in one 3D intensity volume. Adapted from [37].

The EMC algorithm is based on the same idea as EM classification (see Section 5.5.3). It starts with the random orientations for each 2D diffraction pattern. These angles are used to build 3D intensity volume, meanwhile, the probability for the diffraction pattern to be in this orientation is calculated. Poisson noise model is used to build the 3D intensity model.

The process starts with the Expand-step. The grid intensities are expanded into the tomographic representation. The next Maximize-step consists of the probabilistic classification

of the data and then building of a tomographic model. It is the most time-consuming step as the algorithm has to go through the whole number of detector pixels  $t$ , rotation group samples  $r$ , and diffraction patterns  $d$ . Poisson model is used to obtain the probability that the pattern is generated by a given tomogram. The maximization is done by calculating the total log-likelihood of the data that was generated by a new model  $W'_{rt}$ :

$$W_{rt} \rightarrow W'_{rt} = \frac{\sum_d P_{dr} K_{dt}}{\sum_d P_{dr}}, \quad (5.15)$$

where  $W_{rt}$  is the predicted intensity in the pixel  $t$  in orientation  $r$ ,  $W'_{rt}$  is the updated model,  $P_{dr}$  is the probability of the frame  $d$  to have an orientation  $r$ ,  $K_{dt}$  is the number of photons at the same pixel  $t$  in the same diffraction pattern  $d$ . As a result of the M-step, a new model of orientations is calculated. Then the 3D model is compressed back to the 2D with the new orientation angle probabilities. This is the Compress-step. The process is iterative and at the end, for each diffraction pattern, there is an orientation distribution.

The EMC algorithm is implemented in the Dragonfly software [137]. It can not only orient SPI data sets but also simulate diffraction patterns with different parameters for different particles. In the framework of this Thesis, we used Dragonfly only to orient 2D diffraction data from SPI experiments.

## 5.7 Resolution estimation in CXDI and SPI

The estimation of the resolution is the last step of the data analysis pipeline in SPI. It is done after phase retrieval when the electron density of the object is obtained. There are standard ways to estimate the resolution of the real space reconstructed object. The need for resolution calculations lies in the fact that the diffraction patterns obtained from the experiments are not perfect. There are missing areas, the presence of the experimental background, artifacts from the orientation determination – all of these reasons lead to the set of different reconstruction results satisfying modulus and support constraints of the phase retrieval. Thus, the general approach requires calculating several electron density reconstructions starting with random phases and then their averaging.

### 5.7.1 Phase Retrieval Transfer Function

The way to measure the resolution of the obtained averaged electron density of the sample  $\langle \rho_i(\mathbf{r}) \rangle$  is to compare the final result with the measured diffraction intensities  $I(\mathbf{q})$ . It was proposed in Ref. [165] and is called Phase Retrieval Transfer Function (PRTF). The set of independent reconstructions of electron density is made and then averaged. The modulus of the Fourier amplitudes of the result is calculated and compared with the measured in



the experiment intensities. PRTF lies in the range of  $(0, 1)$ . If all reconstructions in the set converge to similar phases, the modulus of the average will be the same as the measured data, and PRTF will be equal to unity. If all reconstructions in the set converge to different solutions and there are no correlations, PRTF will reach zero value. So the PRTF ratio

$$\text{PRTF}(\mathbf{q}) = \frac{|\mathcal{F}\{\langle \rho_i(\mathbf{r}) \rangle\}|}{\sqrt{I(\mathbf{q})}} \quad (5.16)$$

shows transfer function for the phase retrieval. The general tendency of the PRTF is to decrease with the increase of momentum transfer vector  $\mathbf{q}$ .

In practice, PRTF is calculated from the 3D volumes of the result reconstruction and measured data. It is represented in the angular averaged way depending on the momentum transfer vector  $\mathbf{q}$  expressed by the spatial frequency (Eq. (4.17)) or by the real space sampling (Eq. (4.18)). The angular averaged intensity is processed by the two angles  $\phi$  and  $\theta$  and PRTF can be written as

$$\text{PRTF}(q) = \frac{\langle \langle |A_{av}(\mathbf{q})| \rangle_{\phi} \rangle_{\theta}}{\sqrt{\langle \langle I(\mathbf{q}) \rangle_{\phi} \rangle_{\theta}}}, \quad (5.17)$$

where  $A_{av}(\mathbf{q})$  is the Fourier transform of the averaged diffraction amplitude.

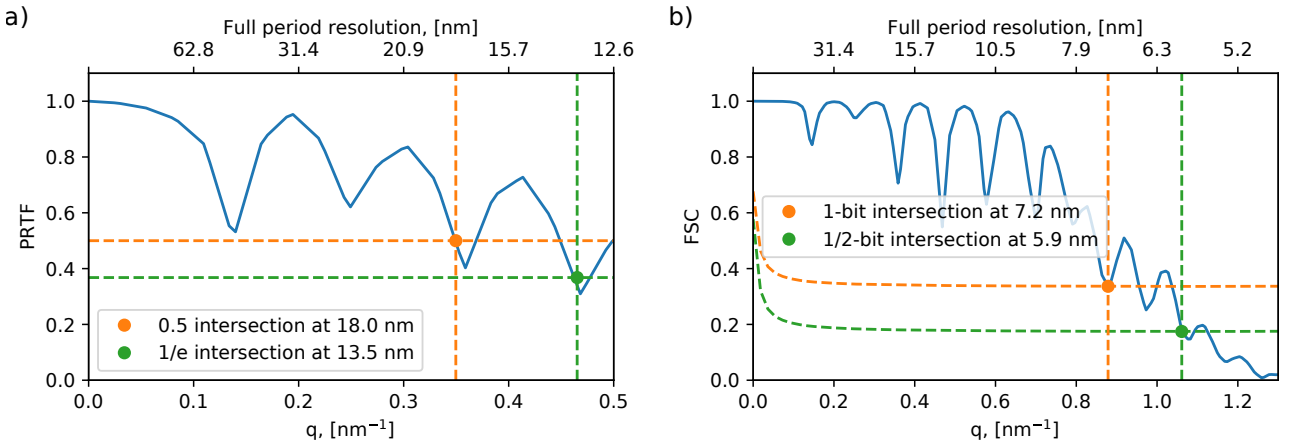


Figure 5.12: (a) An example of PRTF (blue line). Thresholds 0.5 (orange line) and  $1/e$  (green line) give different resolution values. (b) An example of FSC (blue line): 1-bit threshold (orange line) and 1/2-bit threshold (green line).

Typically, PRTF intersection with the threshold of 0.5 is used as the resolution estimation in CXDI. Sometimes, threshold  $1/e$  is used which gives a higher resolution value in comparison with the 0.5 threshold which is shown in Fig. 5.12 (a).

## 5.7.2 Fourier Shell Correlation

As PRTF is comparing the Fourier transform of the averaged electron density of the object with the measured intensities, another approach was suggested. This metric used in the

resolution estimation task is called Fourier Shell Correlation (FSC) [166]. It is based on the comparison of two reconstruction results from two halves of the data set and is representing the consistency of the obtained reconstructions. FSC measures the normalized cross-correlation coefficient between both reconstructions over corresponding shells in Fourier space. Considering we have two reconstruction results from two halves of data,  $\rho_1(\mathbf{r})$  and  $\rho_2(\mathbf{r})$ . Their representation in reciprocal space is the Fourier transform  $F_1(\mathbf{q}) = \mathcal{F}\{\rho_1(\mathbf{r})\}$  and  $F_2(\mathbf{q}) = \mathcal{F}\{\rho_2(\mathbf{r})\}$ . FSC, as PRTE, is the function of momentum transfer and is calculated in each  $i$ -shell of 3D reciprocal space of the angular averages

$$\text{FSC}(q) = \frac{\sum_{q_i} F_1(q_i) \cdot F_2^*(q_i)}{\sqrt{|\sum_{q_i} F_1(q_i)|^2 \cdot |\sum_{q_i} F_2(q_i)|^2}}. \quad (5.18)$$

The threshold criteria was proposed in Ref. [167]. The signal is represented with the "real" signal from the sample which is ideally the same for both reconstructions and random noise signal components which can be described with the signal-to-noise ratio. The resolution is estimated where FSC curve is intersecting with the so-called "1/2-bit" threshold curve. The 1/2-bit threshold corresponds to the voxel containing 1/2-bit information and SNR of each of the two reconstructions reaches a value of 0.2071. Thus, it can be calculated as

$$T_{1/2\text{-bit}}(q) = \frac{0.2071 + 1.9102 \cdot 1/\sqrt{n(q_i)}}{1.2071 + 0.9102 \cdot 1/\sqrt{n(q_i)}}, \quad (5.19)$$

where  $n(q_i)$  is the number of voxels in the shell  $q_i$ .

Besides the 1/2-bit threshold curve, it is also possible to use the 1-bit threshold curve which is shown in Fig. 5.12 (b). When it intersects the FSC curve, the SNR in the each of two final reconstructions at the intersection point will be 0.5. The 1-bit threshold is considered to be more stringent and in this case, the voxel of reciprocal space will contain 1 bit of information. Generally, the 1/2-bit threshold is used as a more standard one and in the framework of this Thesis, we will be using FSC resolution estimation approach with the 1/2-bit threshold intersection.



## Chapter 6

# An advanced workflow for SPI experiment with the limited data at an X-ray free-electron laser

Several results in SPI were published in the framework of the SPI consortium [35, 36, 115–117]. The following Chapter is based on the results presented in Ref. [38]. The SPI experiment was performed with the bacteriophage PR772 at the Atomic, Molecular and Optical Science (AMO) instrument at the Linac Coherent Light Source (LCLS) as part of the SPI initiative. We advanced the previous studies performed on PR772 [35] by extending the general analysis pipeline of the SPI data (Fig. 5.7) as first proposed in [107]. Main upgraded features include: a classification method based on the Expectation-Maximization (EM) algorithm which was described in Section 5.5.3 and mode decomposition for the final virus structure determination [15, 72, 168].

Modifications of the existing methods and application of the new ones allowed to cope with the shortcomings of the experimental data: inaccessibility of information from the half of the detector and small fraction of single hits. Below we present a detailed description of all steps which allowed obtaining the electron density of PR772 virus with an improved resolution of 6.9 nm as compared to the previous SPI studies on the same virus [35]. The obtained resolution was limited by the scattering intensity during the experiment and the relatively small number of single hits.

### 6.1 Experiment

The experiment was performed at the AMO instrument [169–171] at the LCLS at SLAC National Accelerator Laboratory using the LAMP end station (experimental details are available in Ref. [172]). PR772 bacteriophage growth and purification is described in Ref. [116].

Virus has icosahedral shape and size distribution according to the cryo-EM studies of the PR772 virus [173] used in the experiment is shown in Fig. 6.1.

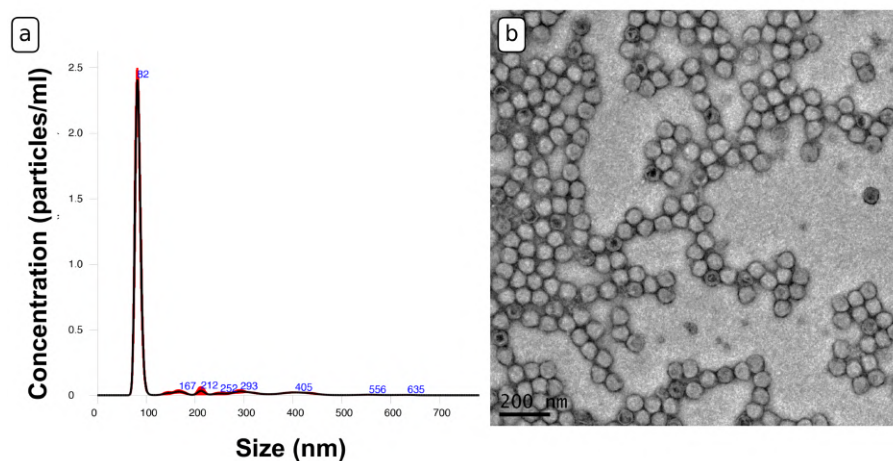


Figure 6.1: Cryo-EM structural studies of PR772 used in the experiment. (a) Virus size distribution profile. (b) PR772 visualization with screening transmission electron microscopy (TEM).

Viruses in ammonium acetate volatile buffer were aerosolized in a helium environment using a gas dynamic virtual nozzles (GDVN) that were 3D printed via two-photon polymerization photo-lithography with a Nanoscribe Professional GT printer [130]. Laboratory measurements showed reproducible jet diameters in the range of 0.5 – 2.0  $\mu\text{m}$  [174, 175]. The particles then passed through a differentially pumped skimmer for pressure reduction and were then injected/focused into the sample chamber using an aerodynamic lens injector [124, 176]. The focused particle stream intersected the focused and pulsed XFEL beam.

The experimental set-up is typical for SPI (see Fig. 5.2). The LCLS had a repetition rate of 120 Hz, for this experiment the average pulse energy was 4 mJ, with a focal diameter of 1.5  $\mu\text{m}$ , and a photon energy of 1.7 keV (wavelength 0.729 nm). Diffraction patterns were recorded by a pnCCD detector [177] mounted at a 0.130 m distance from the interaction region. In the experiment, a silver behenate salt was used as a calibration agent to determine sample detector distance and panel position.

The scattering signal was recorded only by one of the two detector panels (one panel was not operational during the experiment due to an electronic fault). The size of the working panel was  $512 \times 1024$  pixels with a pixel size of  $75 \times 75 \mu\text{m}^2$ , with the long edge closest to the interaction point. In the middle of the experiment (at run 205) the detector panel was moved one millimeter up vertically relative to the incoming X-ray beam to reduce background scattering. This research is based on the data obtained from this panel covering part of reciprocal space as shown in Fig. 6.2. The experimental data sets used here are publicly available at [178].

## 6.2. Initial classification steps

---

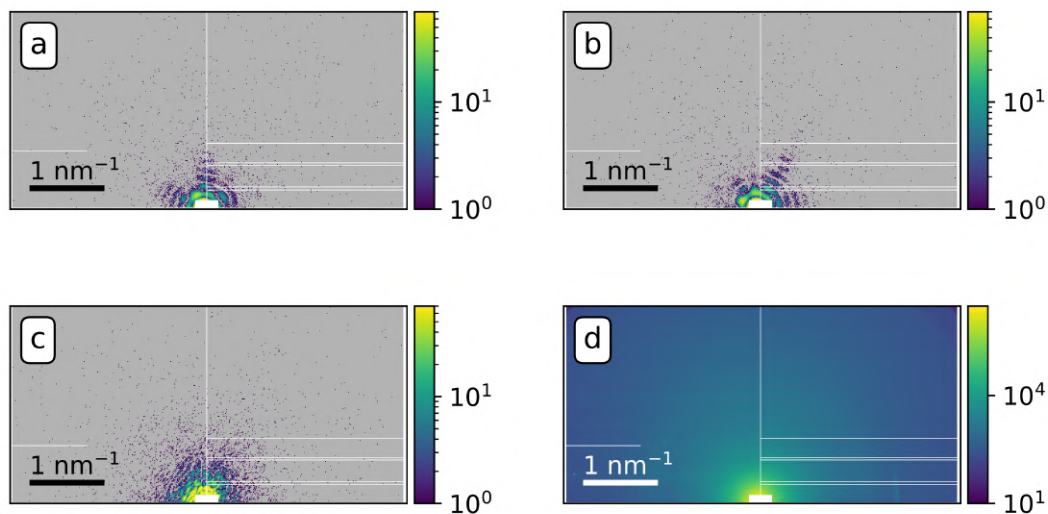


Figure 6.2: Examples of diffraction patterns from the SPI experiment. (a), (b) Diffraction patterns corresponding to a single PR772 virus hit by an XFEL beam. (c) Diffraction pattern corresponding to a non-single hit which was sorted out from the analysis at the classification step. (d) Sum of  $1.9 \times 10^5$  diffraction patterns identified as hits. White regions in the center of diffraction patterns as well as white stripes correspond to a mask introduced to reduce artifacts due to the signal exceeding the detector capabilities. The mask was the same before and after the move of the detector panel.

## 6.2 Initial classification steps

SPI data analysis involves many subsequent steps visualized in Fig. 5.7, leading to the 3D reconstructed particle structure from a large set of 2D diffraction patterns [107]. Improvements in the data analysis pipeline at early stages can result in significant enhancement in reconstruction quality. Therefore, several pre-processing methods were developed to avoid experimental artifacts on the collected diffraction patterns. Pre-processing stages include: hit finding, background correction, beam position refinement, and particle size filtering.

### 6.2.1 Hit finding

The initial experimental dataset, as collected at the AMO instrument, consists of about  $1.2 \times 10^7$  diffraction patterns ( $9 \times 10^6$  patterns before and  $3 \times 10^6$  patterns after moving the detector panel) (see Ref. [172]). The hit finding was performed using the software “psocake” in the “psana” framework [179]. As a result,  $1.9 \times 10^5$  diffraction patterns were identified as hits from the initial set of diffraction patterns, and the signal from these hits was converted to photon counts (see Table 6.1). Examples of diffraction patterns are shown in Fig. 6.2. The Power Spectral Density (PSD) function, i.e. angular averaged intensity, for the diffraction patterns identified as hits at this analysis step is shown in Fig. 6.3.

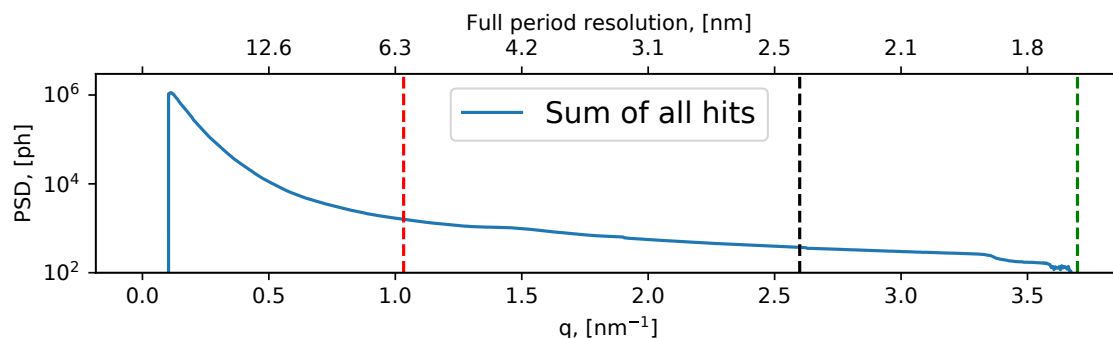


Figure 6.3: The PSD function of the scattered intensity for the sum of all diffraction patterns identified as hits collected in the experiment. The signal until the corner and edge of the detector is indicated by the green and black vertical dashed lines, respectively. For the orientation determination the data until momentum-transfer values  $1 \text{ nm}^{-1}$  (shown in red dashed line) were used.

## 6.2.2 Analysis of additional instrumental scattering

Visual inspection of the selected patterns revealed the presence of additional instrumental scattering near the center of the diffraction patterns (see Fig. 6.4). This signal remains stable from pulse to pulse which indicates that it originates from beamline scattering. Most probably, it was caused by the interaction of the tails of the XFEL beam with the upstream apertures or from the sample injector.

This additional instrumental scattering can be well seen on the averaged diffraction pattern in one of the experimental runs displayed in Fig. 6.4 (a). We analyzed histograms of intensity for individual pixels and noticed that pixels with additional instrumental scattering most often recorded a signal of several photons. Contrary to that, pixels without this additional scattering most frequently recorded a signal of zero photons. We assumed that beamline scattering follows a Gaussian distribution and it was incoherently added to particle scattering.

To correct this additional signal, we fit the first peak on histogram of intensity for each pixel by a Gaussian function (see Fig. 6.5). Then we subtract the value of the Gaussian center from the total signal of this pixel for all diffraction patterns. This instrumental scattering subtraction was crucial for further beam center position finding and particle size filtering. We did not mask this region because we would lose important information about the first diffraction minimum.

## 6.2.3 Beam center position finding

The beam center position was retrieved from the diffraction patterns. Detector consists of two identical panels with the gap between them for direct beam propagation. Due to the fact that the signal from only one panel was available, the beam center position could not be determined by centrosymmetric property of diffraction patterns. Furthermore, the detector

## 6.2. Initial classification steps

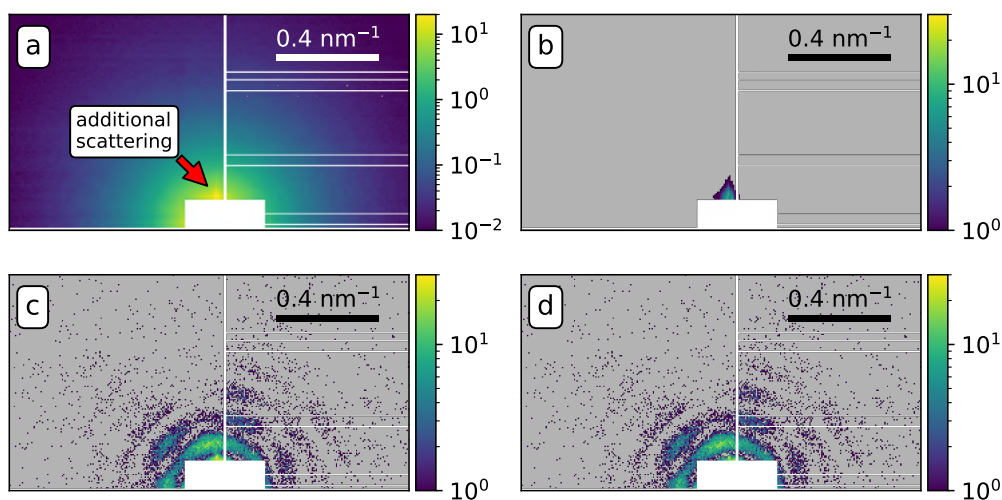


Figure 6.4: (a) An averaged diffraction pattern of one of the runs. White regions in the diffraction patterns correspond to a mask introduced to hide misbehaving pixels. Additional instrumental scattering originating from the beamline is well visible in the central part of the averaged diffraction pattern. (b) Identified additional scattering for this run. Diffraction pattern before (c) and after (d) subtraction of additional scattering.

panel was moved during the experiment, and we estimated the beam position twice – before and after the detector panel was moved.

The beam center position was determined in the following way. First, the sum of all diffraction patterns was calculated. The resulting average diffraction pattern was rotationally symmetric and allowed a rough estimate of the beam position center. For the success of this step, it was crucial to subtract parasitic scattering from the beamline as described in the previous section.

To define the beam position center more carefully on the next step, diffraction patterns with a narrow distribution of particle sizes were selected and the averaged diffraction pattern was obtained. This diffraction pattern has pronounced diffraction fringes and it was correlated with the 2D form factor of a spherical particle (see Fig. 6.6). Inspection of this method on simulated data with similar parameters showed that mean deviation of the refined center from the true center of diffraction patterns is less than half of a pixel.

### 6.2.4 Particle size filtering

Next, the particle size filtering was performed in two steps. It was based on the fitting of the PSD function of each diffraction pattern with the form factor of a sphere. A set of form factors corresponding to the spheres with the diameter in the range from 30 to 300 nm was generated first. On the next step the PSD function of each diffraction pattern was fitted with a spherical form factor function from the generated set (see Fig. 6.7 (a)). As the fit quality measure for the certain size (diameter) of the spherical particle, the mean difference was



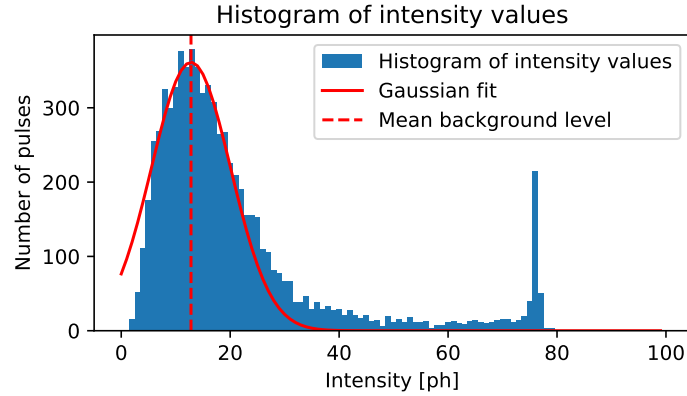


Figure 6.5: Histogram of intensity values for a selected pixel from one of the runs with strong additional instrumental scattering. The pixel shows the most frequently recorded value is 13 photons. This histogram was fitted with a Gaussian function and the mean value of this Gaussian function was subtracted from all intensity values for different pulses corresponding to that pixel. The peak on the right side of the histogram is due to limitations of the detector: if the detector pixel collects more than 75 photons, its response is always in the range of 76 – 79 photons.

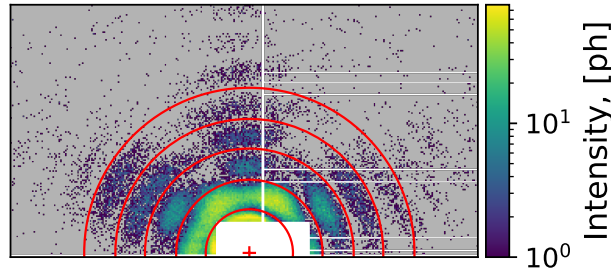


Figure 6.6: Center position on a selected diffraction pattern. Minima of the optimal spherical form factor are shown by red circles.

used

$$D_S = \frac{1}{q_{max} - q_{min}} \sum_{q_{min}}^{q_{max}} |I_{exp}(q) - I_S(q)|, \quad (6.1)$$

where  $I_{exp}(q)$  is the PSD value of the experimental intensity for selected  $q$ ,  $I_S(q)$  is the form factor of a sphere with the size (diameter)  $S$ .

In Eq. (6.1) the  $q$ -values were ranging from  $q_{min} = 0.12/0.15 \text{ nm}^{-1}$  before and after the detector panel was moved, up to  $q_{max} = 0.66 \text{ nm}^{-1}$ . An example of the mean difference function of Eq. (6.1) obtained for one of the diffraction patterns is shown in the Fig. 6.7 (b). This function has several minima, where the first minimum corresponds to a sphere with the best size. The second minimum corresponds to a sphere with the second-best size, etc. To measure fidelity of the particle size estimation we used fidelity score (FS) defined as

$$FS = \frac{D_{S_2}}{D_{S_1}}, \quad (6.2)$$

## 6.2. Initial classification steps

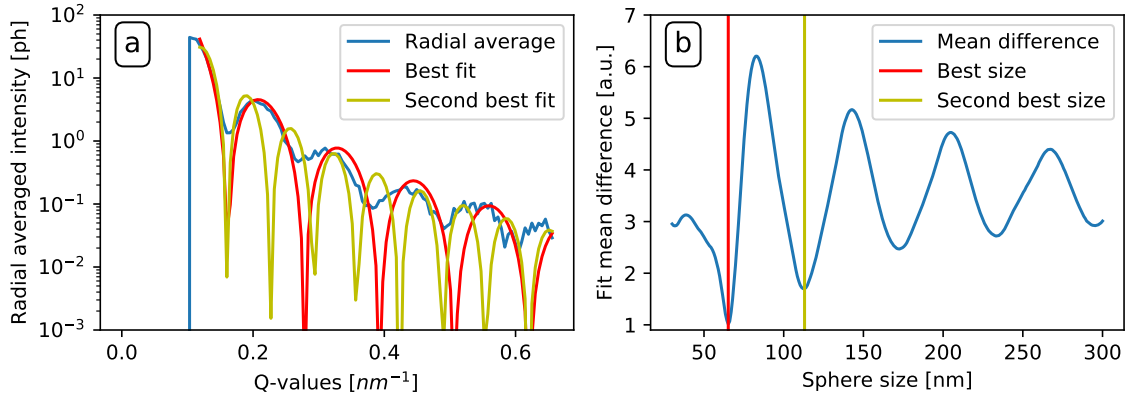


Figure 6.7: PSD fitting analysis of the diffraction pattern. (a) PSD function (blue line) was fitted with the form factors of spherical particles of different size. Red and yellow lines correspond to the form factors of the spherical particles with the best and second best size, that were used for calculation of fidelity score. (b) Mean difference function as defined in Eq. (6.1). Fidelity score value is 1.6 for this diffraction pattern.

where  $D_{S_1}$  and  $D_{S_2}$  are the values of the mean difference function  $D_S$  corresponding to the first and second minima in Fig. 6.7 (b).

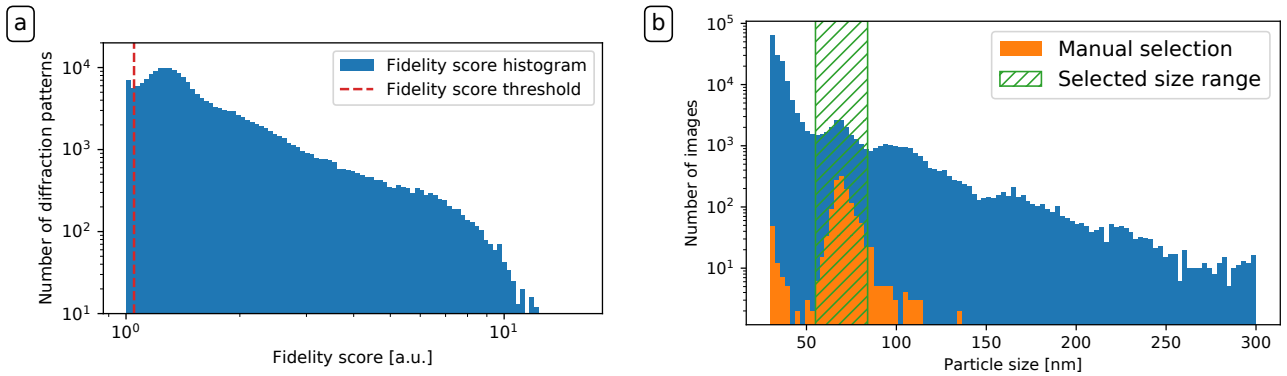


Figure 6.8: (a) Fidelity score histogram for all diffraction patterns identified as hits in the experiment. Fidelity score threshold of the value 1.05 is shown as the vertical dashed red line.  $1.8 \times 10^5$  selected diffraction patterns with fidelity score above threshold were used. (b) A particle-size histogram after the PSD-function filtering. The blue area corresponds to  $1.8 \times 10^5$  diffraction patterns fidelity score filtering. The orange area corresponds to manually selected single-hit diffraction patterns. A range of particle sizes from 55 to 84 nm ( $1.8 \times 10^4$  diffraction patterns) were selected further (green dashed area).

The fidelity score histogram for all diffraction patterns identified as hits ( $1.9 \times 10^5$  diffraction patterns) is shown in Fig. 6.8 (a). According to its definition in Eq. (6.2), if the fidelity score is equal to unity  $FS = 1$ , it means that  $D_S$  values equal for two different minima, therefore fitting cannot find an appropriate size for a particle that will correspond to a given diffraction pattern. We introduced a threshold value of  $FS = 1.05$  and considered all diffraction patterns with the fidelity score higher than this value (see Fig. 6.8 (a)). By that we

determined  $1.8 \times 10^5$  diffraction patterns that were selected for the further particle size filtering.

A histogram for different particle sizes in the selected size range is shown in Fig. 6.8 (b). An extended range of sizes observed in this figure corresponds to clusters of particles stuck together and a varying thickness of a hydration layer. One can also identify in this histogram a peak in the range from 55 nm to 84 nm. This size range agrees well with the expected virus particle size of about 70 nm [35, 173]. The considered range contains  $1.8 \times 10^4$  diffraction patterns which were selected for further single hit classification (see Table 6.1).

Table 6.1: Datasets selected at different stages of the analysis: hit finding selection, PSD-fitting score filtering, particle-size filtering and single-hit diffraction pattern selection. The percentage of the chosen dataset to the initial one  $S_0$  is given in parentheses.

Analysis step	Dataset name	Number of diffraction patterns
Initial dataset	$S_0$	$1.2 \times 10^7$
Hit-finding classification	$S_{hit}$	191,183 (1.6%)
PSD-fitting score filtering	$S_{fit}$	179,886 (1.5%)
Particle-size filtering	$S_D$	18,213 (0.1%)
First EM-based classification	$S_1^{EM}$	1,609
Second EM-based classification	$S_2^{EM}$	1,402
Third EM-based classification	$S_3^{EM}$	1,366
Fourth EM-based classification	$S_4^{EM}$	1,401
Fifth EM-based classification	$S_5^{EM}$	2,119
Final EM-based classification	$S_{EM}$	1,085 (0.009%)
Manual selection	$S_{man}$	1,393 (0.01%)

### 6.3 Single hit diffraction patterns classification

The key step of data selection in this work was single-hit classification. The angular X-ray cross-correlation analysis (AXCCA) classification that was used in the previous work [35, 147] was not as effective with the present experimental data due to the absence of the scattering signal on one half of the detector and the low single-hit rate.

To overcome this challenges of experimental data, we used Expectation-Maximization (EM) algorithm to solve single hit classification task. It was described in details in Section 5.5.3. As soon as, low contrast and low signal to noise level are common problems for Cryo-EM and SPI, we implemented in this work the original EM algorithm developed in Cryo-EM for clustering of SPI data.

Diffraction patterns remaining from the previous step of particle size filtering ( $1.8 \times 10^4$ ) were distributed into 20 classes. As an example, in one of the EM-clustering we selected

### 6.3. Single hit diffraction patterns classification

classes 1 and 2 as classes containing high contrast six fringes that we attribute to scattering from a PR772 virus particle (see Fig. 6.9).

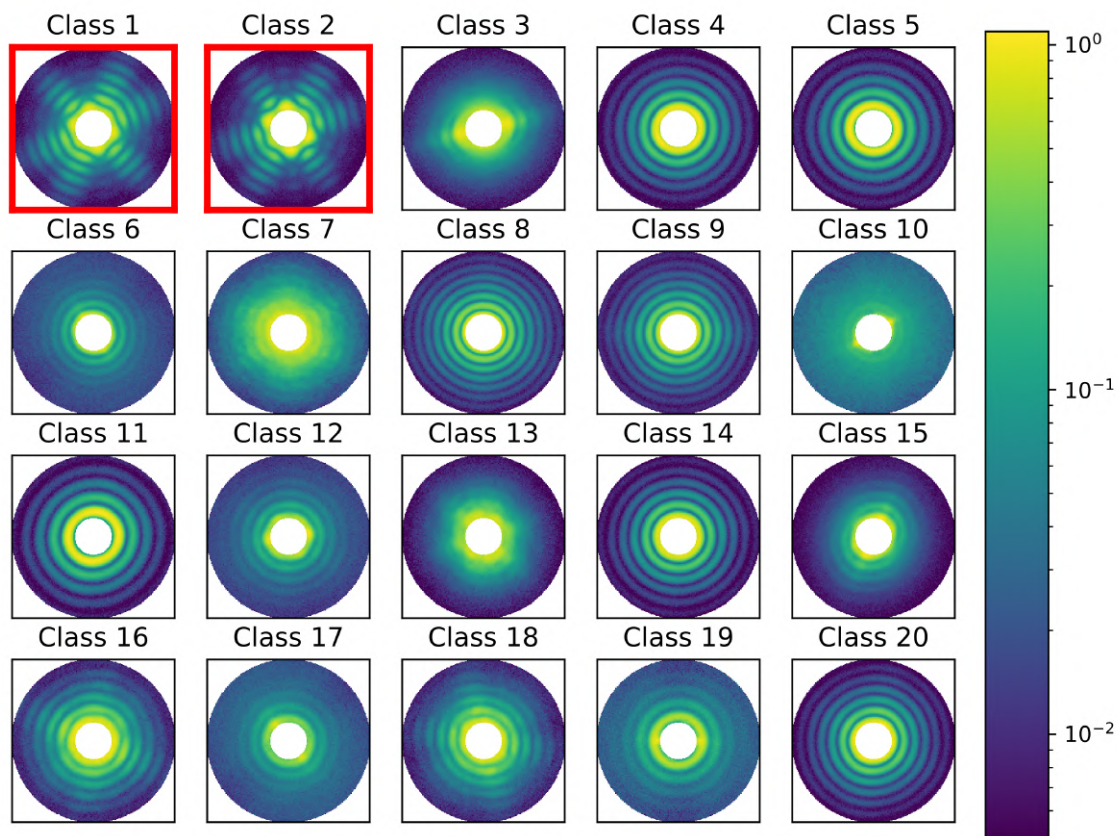


Figure 6.9: Classification of diffraction patterns by EM clustering. Diffraction patterns are distributed into 20 classes according to their features. Classes 1 and 2 were selected as they clearly contain structural features of the investigated virus and its icosahedral shape. These two classes contain 1,609 diffraction patterns.

To make classification more accurate we performed five independent EM-clusterings starting from random cluster models. Each EM-clustering produced slightly different results which are summarized in Table 6.1. The intersection of all results was considered as a stable single hit selection (see Fig. 6.10 (a)). Finally, we ended with the data set containing 1,085 single particle patterns (see Table 6.1). The PSD function of an average of all selected diffraction patterns is shown in Fig. 6.10 (b). The scattering signal from the virus particle extends up to the momentum transfer value  $q_{max} = 1 \text{ nm}^{-1}$  in reciprocal space which corresponds to a resolution ( $2\pi/q_{max}$ ) of 6.3 nm in real space.

A manual search and selection of single hits from the data set of  $1.9 \times 10^5$  diffraction patterns produced a new data set containing 1,393 diffraction patterns [172]. From this selection 574 diffraction patterns are also present in our EM-based selection. The PSD function for the manual selection shown in Fig. 6.10 (b) has lower contrast with less visible fringes. We attribute this mostly to the absence of the size filtering step in manual selection. The virus size

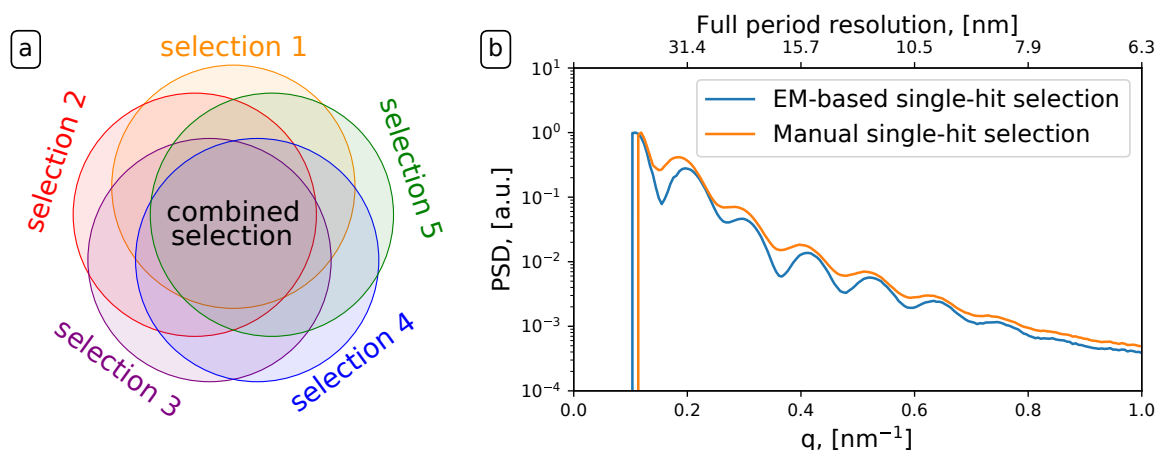


Figure 6.10: Classification of diffraction patterns by EM clustering. (a) EM-clustering was repeated five times, and intersecting selection with 1085 patterns was considered for further analysis. (c) Averaged PSD functions for EM-based single-hit selection containing 1,085 patterns (blue line) and for manual selection containing 1,393 patterns (orange line).

fluctuations could be caused by a slight change in hydration layer thickness, variation of the position where the particles interact with X-ray beam, slightly upstream or downstream of the focus or a real sample size distribution of the PR772.

We note that in previous study of PR772 [35] the number of single hit diffraction patterns was about  $7.3 \times 10^3$  which is about one order of magnitude higher than in this SPI experiment. The reason for the smaller number of patterns was a combination of downtime due to detector troubleshooting and time needed for sample-injection optimizations. The GDVN flow and pressures were needed to be optimized because of the relatively large initial droplet diameters of approximately one to two micrometers.

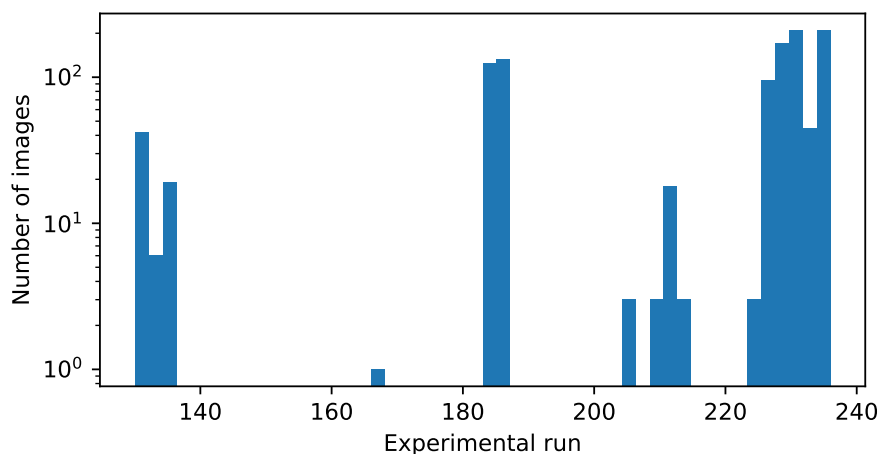


Figure 6.11: Histogram of the number of images per experimental run from the data set containing 1,085 patterns.

After the size filtering and running EM algorithm, we ended with the data set containing 1,085 patterns (see Table 6.1). In order to identify performance of single particle collection

as well as efficiency of the 3D printed nozzles we plot a histogram of selected patterns as a function of experimental run (see Fig. 6.11). This histogram shows that collection was significantly improved towards the end of the experiment.

## 6.4 Orientation determination and background subtraction

For orientation determination we used the EMC algorithm implemented in the Dragonfly software package described in [137] and mentioned in Section 5.6 of this Thesis. This iterative algorithm successfully combines 2D diffraction patterns into 3D intensity distribution of the PR772 virus.

Results presented in Fig. 6.12 (a)-(c) clearly show that the recovered 3D diffraction intensity contains a substantial high momentum background that may be caused by scattering on helium from the carrier gas. To reduce the influence of this background contribution, a background subtraction was applied. Several approaches for the background correction in SPI experiment analysis were developed earlier and were already mentioned in Section 5.3 and it was understood that the background correction method may affect the reconstructed structure.

Here we used a combined approach for the background determination, first a constant background was subtracted and then, on the reconstruction step, the contrast of diffraction patterns was additionally enhanced by applying deconvolution algorithms which will be described further. In the first step, the background level was defined as the mean signal value in the high momentum region of the 3D intensity distribution free of particle scattering contribution (see red boxes in Fig. 6.12 (a)-(c)). The histogram of intensity in this area is shown in Fig. 6.12 (d). The background level was defined as the mean signal value and was subtracted from the 3D intensity map in reciprocal space. Negative values of intensity in the final representation were set to zero.

The PSD function after background subtraction (red line in Fig. 6.12 (e)) reveals artifacts in the regions of low ( $q < 0.12 \text{ nm}^{-1}$ ) and high ( $q > 0.93 \text{ nm}^{-1}$ ) momentum transfer values. Since the data in these regions did not follow the expected spherical form factor behavior, we did not consider this part of the data. Data used for further analysis are shown with red dots in Fig. 6.12 (e).

Reciprocal space data with and without background subtraction are shown in Fig. 6.13. As it is seen in Fig. 6.12 (e), the power law dependence of data after the background subtraction is the same up to high  $q$ -range of about one inverse nm. That is in contrast to the data after EMC orientation determination (blue curve) which is clearly saturated at high  $q$ -range.

Due to the background subtraction, more features and higher contrast were revealed in the high momentum transfer region. The fringe visibility or averaged contrast  $\langle \gamma \rangle$  was

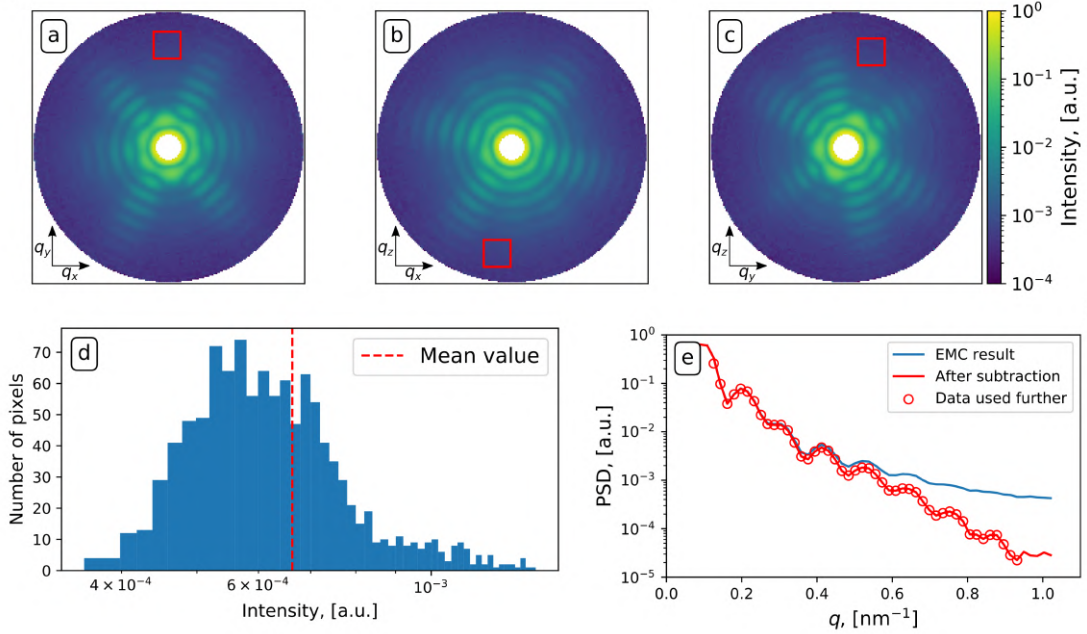


Figure 6.12: Results of the EMC orientation determination algorithm. (a)-(c) Orthogonal 2D cuts through the center of the 3D volume of reciprocal space after application of the EMC algorithm. For the background estimate the intensity values in the region of high  $q$ -values shown with red squares were analyzed. (d) Histogram of the signal from the area shown in (a)-(c). The mean value of the signal is shown with the vertical dashed red line. (e) PSD functions before (blue line) and after (red line) background subtraction. To avoid artifacts at low and high  $q$ -values a part of the curve indicated with red dots was considered for further analysis.

defined as

$$\langle \gamma \rangle = \frac{1}{N} \sum_{i=1}^N \gamma_i, \quad (6.3)$$

where

$$\gamma_i = \frac{I_{max}^i - I_{min}^i}{I_{max}^i + I_{min}^i}. \quad (6.4)$$

Here  $\gamma_i$  is the local contrast and  $I_{max}^i$  and  $I_{min}^i$  are the PSD-function values in the local maxima and following minima, respectively, and  $N$  is a number of pairs of maxima and minima considered in this analysis.

In our case the average contrast values  $\gamma_i$  were calculated for the first  $N = 6$  pairs. For the experimental data before the background subtraction we obtained the value  $\langle \gamma \rangle = 0.41$ . The fringe contrast after the background subtraction showed significant improvement with  $\langle \gamma \rangle = 0.58$ .

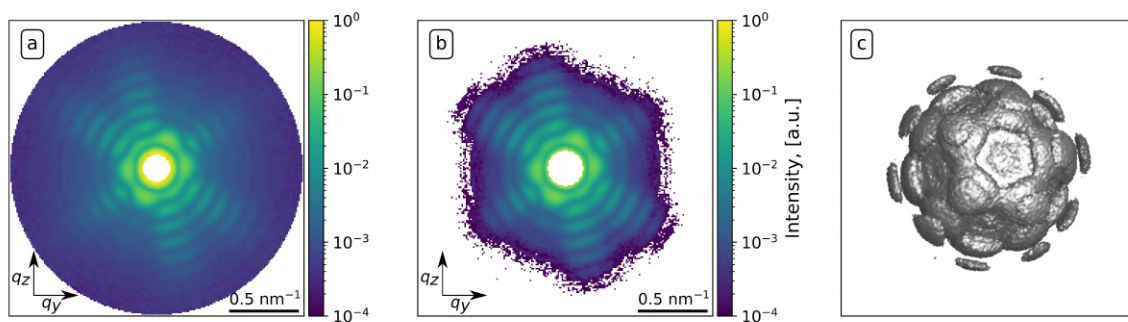


Figure 6.13: Background subtraction results. (a) A 2D  $q_y$ - $q_z$  cut of the 3D intensity distribution in reciprocal space and (b) the same intensity distribution after the background subtraction. (c) The 3D intensity distribution in reciprocal space after the background subtraction, shown at 0.5% level of the maximum value.

## 6.5 PR772 virus structure

### 6.5.1 Phase retrieval

To determine the electron density distribution of the PR772 virus the phase retrieval was performed using the 3D intensity distribution of the virus in reciprocal space described above (Fig. 6.13 (c)). Iterative phase retrieval algorithms are based on the Fourier transform between real and reciprocal space using two constraints: in reciprocal space the amplitude of the signal is equal to the experimentally measured values and in real space a finite support of the particle is used [87, 93]. They described in details in Section 4.3.

The electron density distribution of the PR772 virus was obtained with the following steps. First, the central gap, originating from the initial data masking of the diffraction patterns, was filled (Fig. 6.13). This was accomplished by running multiple 3D reconstructions of the virus with an assumption of free-evolving intensity in this part of reciprocal space. The following algorithms were considered at this stage: 90 iterations of Continuous Hybrid Input-Output (CHIO) with the feedback value 0.8 [180], 200 iterations of the Error Reduction (ER) algorithm [93] with alternation of the Shrink-Wrap (SW) algorithm each 10 iterations with the threshold value of 0.2 and Gaussian filtering with 3 to 2 sigma [100]. This combination of algorithms was repeated three times for one reconstruction with the total number of 870 iterations.

All obtained reconstructions showed identical central part and we used one of them in further analysis. In Fig. 6.14 (a) the PSD functions of the initial and one of the reconstructed data are shown. One can see from that figure that the reconstructed curve follows very well the experimental data points. For the low  $q$ -values below  $0.14 \text{ nm}^{-1}$  the experimental data were substituted with the data obtained in phase retrieval. Difference between experimental data and reconstruction in the central fringe is contributed to incorrectly reduced detector



signal for intensity above 75 photons (Fig. 6.5). This modified 3D intensity map was used for the final phase retrieval and virus structure determination (Fig. 6.14 (b)).

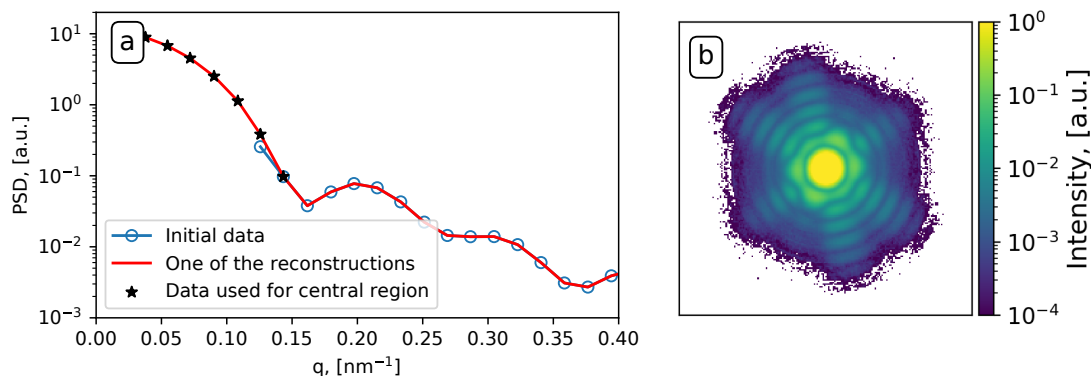


Figure 6.14: PSD functions for experimental data (blue empty dots) and one of the reconstructions (red line). The central part below  $q = 0.14 \text{ nm}^{-1}$  (black stars) was taken from this reconstruction for further analysis (b) Modified data with the filled central part. White area around the diffraction pattern is the part of reciprocal space where the data were set to zero initially but were allowed to freely evolve during the iterative phase retrieval.

On the next step 50 individual reconstructions were performed. In these reconstructions intensities at high  $q$ -values (in the regions where they were initially set to zero – white area in Fig. 6.14 (b)) were allowed to freely evolve. The initial support was taken as a Fourier transform of the 3D data used for reconstructions and had spherical shape with diameter about 90 nm. The same sequence of algorithms was used for these reconstructions as mentioned above plus it was performed 100 iterations of the Richardson-Lucy (RL) algorithm [134] with the total number of 970 iterations.

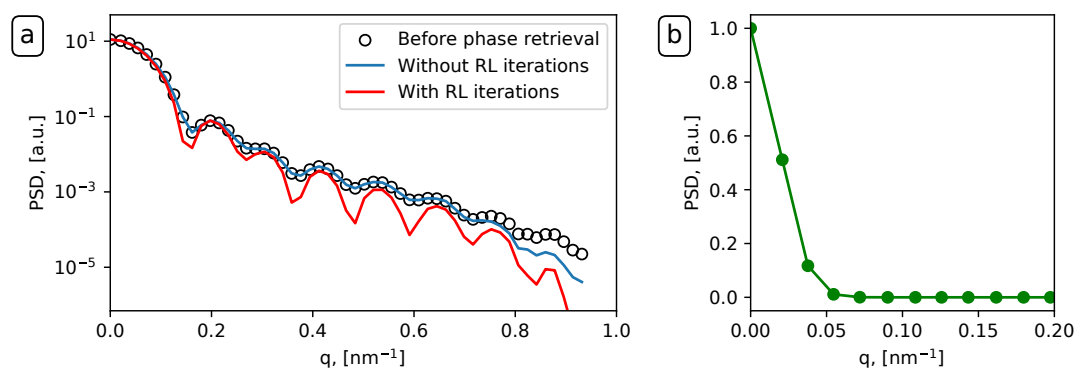


Figure 6.15: (a) PSD functions for data before phase retrieval (black empty dots) and one of the reconstructions without RL iterations (blue line) and with them (red line). This additional deconvolution allowed us to improve contrast of diffraction patterns. (b) PSF for individual reconstruction as a result of RL deconvolution algorithm. Intensity is normalized to the maximum value.

This algorithm based on deconvolution technique allowed to additionally enhance the contrast of the reconstructed diffraction patterns to the value of  $\langle \gamma \rangle = 0.87$  (see Fig. 6.15 (a)),

and by that remove the remaining background from the 3D diffraction patterns which is defined as Point Spread Function (PSF) shown in Fig. 6.15 (b). As a result, we obtained complex valued real space images for each 50 reconstructions from which the absolute value was considered as an electron densities of the virus.

## 6.5.2 Mode decomposition

On the final step, to identify electron density of the virus we used mode decomposition. As an outcome of this procedure an orthogonal set of modes was found. The whole procedure consists of the following steps (Fig. 6.16 and [15]):

- Initial 4D matrix consists of 3D amplitudes of the reconstructions ( $203 \times 203 \times 203$  pixels), where the fourth dimension is given by the number of reconstructions (50 in the present case, see Fig. 6.16 (a)).
- This 4D matrix of amplitudes is rearranged into a 2D matrix (Fig. 6.16 (b)) with 50 columns, where each 3D amplitude matrix was rearranged to a 1D column.
- Next, the mode decomposition is performed for the density matrix that is obtained by multiplication of the previously defined 2D matrix transposed complex conjugated and 2D matrix itself (Fig. 6.16 (c))

$$\rho(\mathbf{r}_1, \mathbf{r}_2) = \langle \rho^*(\mathbf{r}_1) \rho(\mathbf{r}_2) \rangle, \quad (6.5)$$

where  $\rho(\mathbf{r})$  relates to complex-valued real-space images and the brackets  $\langle \dots \rangle$  indicate ensemble averaging over different reconstructions.

By diagonalization of this matrix using Principal Component Analysis (PCA), eigenfunctions and eigenvalues of the reconstructed object are obtained

$$\rho(\mathbf{r}_1, \mathbf{r}_2) = \sum_{n=0}^N \beta_n \rho_n^*(\mathbf{r}_1) \rho_n(\mathbf{r}_2), \quad (6.6)$$

where  $\beta_n$  are eigenvalues of this decomposition.

This approach is advantageous in comparison to averaging that would make important object features blurry and, finally, may affect the final resolution. By considering the zero mode only, unique features present in all other reconstructions will be represented. In practice, we used all 50 reconstruction results and performed mode decomposition.

The fundamental mode (with the weight factor  $\beta_0$  of 99%) was considered as the final result of reconstruction. Such a high weight factor value indicates that all reconstructions converged essentially to the same result with the uncertainty level of only 1%. To determine

the electron density, we took an absolute value of this fundamental mode complex valued image.

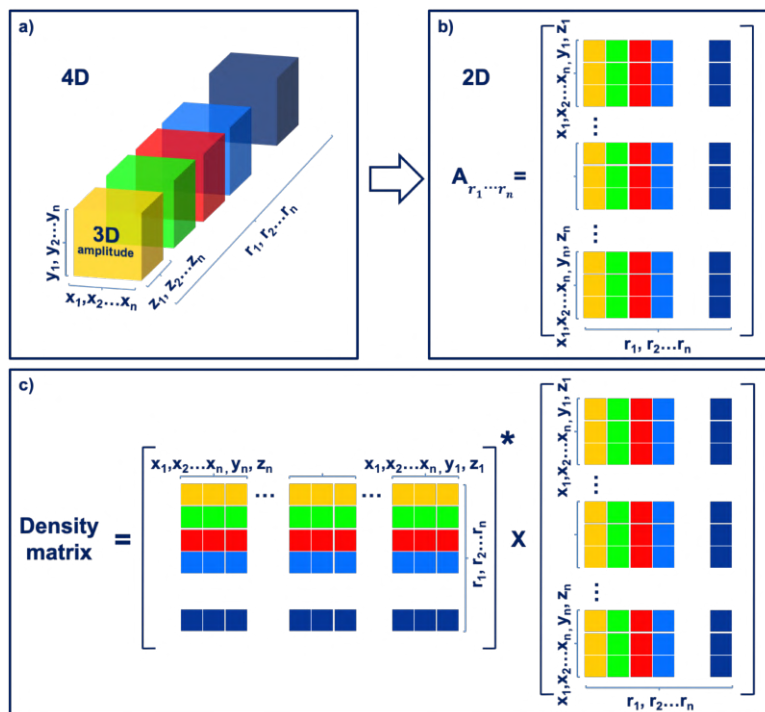


Figure 6.16: Mode decomposition procedure for the set of the reconstructions obtained by phase retrieval. (a) Initial 4D matrix consisting of 3D amplitudes of the reconstructions ( $203 \times 203 \times 203$  pixels), where the fourth dimension is the number of reconstructions. (b) 4D matrix rearranged to 2D matrix, where each 3D amplitude matrix was rearranged to 1D column, the number of columns corresponds to the number of reconstructions. (c) Density matrix obtained by the multiplication of 2D matrices (b): its transposed complex conjugated and matrix itself.

### 6.5.3 PR772 structure analysis

The real space electron density was three-times upsampled and the comparison between the initial and up-sampled structures is shown in Fig. 6.17. The electron density of the reconstructed PR772 virus was normalized to the maximum value.

Different visualizations of the retrieved virus article are shown in Fig. 6.18. As is seen, the obtained electron density of PR772 shows the expected icosahedral structure. A 2D cut of the virus structure is shown in Fig. 6.17 (d)-(f), where a higher electron density in a thin outer shell is well resolved and is attributed to the capsid proteins arrangement. As one can see, the electron density in the center of the reconstructed virus particle was reduced. The reason for this may be heterogeneity of virus particles present in solution and injected in X-ray beam.

Bacteriophage PR772, like other members of the Tectiviridae family, contains an inner proteolipid membrane that facilitates delivery of the viral genomic DNA during infection [181].

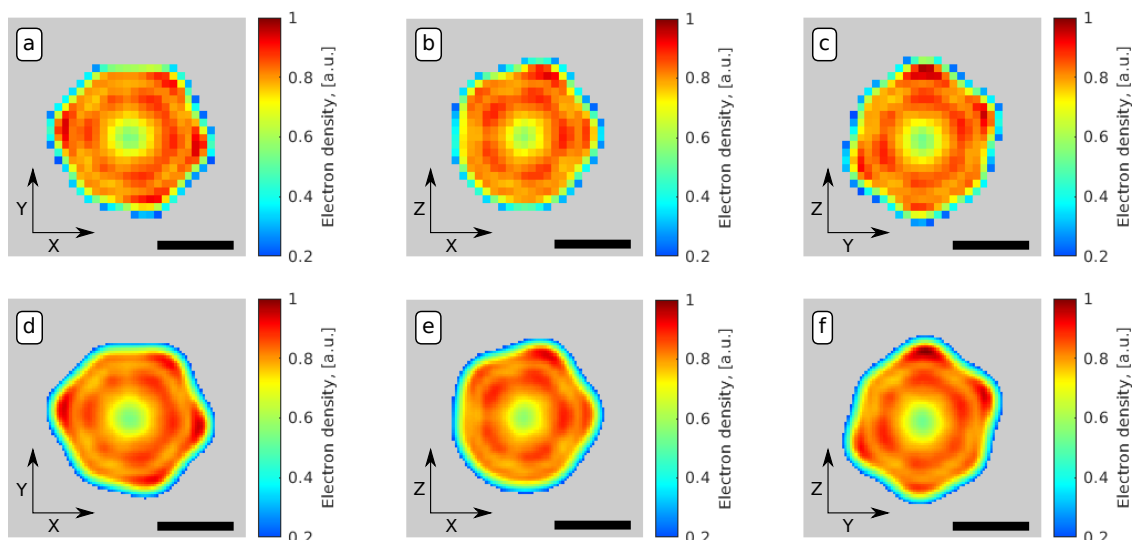


Figure 6.17: Final electron density of the virus obtained as a result of mode decomposition. Black line denotes 30 nm. (a)-(c) Results of the initial reconstruction. (d)-(f) Three times up-sampled results from (a)-(c). Electron density less than 0.2 was set to gray color scale.

When the virus binds to a bacterial host cell the inner membrane is extruded from one of the viral vertices to form a nanotube that facilitates genome delivery. Spontaneous release of the viral genome has been reported [182, 183]. Bacteriophage PR772 preparations were also analyzed by Cryo-EM (Fig. 6.1). Under the conditions used for plunge freezing, we observed particles that were more rounded than icosahedral. Preliminary volume analysis suggests that some particles had released their DNA. These particles appeared to be “triggered” during the freezing process since TEM imaging did not reveal this.

It may be possible from minor differences in virus preparation or upon XFEL sample delivery that some of the virus particles were similarly “triggered” to release their genomes during cryo-freezing, thus resulting in decrease in inner electron density. Previous analyses of XFEL snapshots on the same virus also suggested that some particles exhibit decreased inner density [117, 184].

To identify the particle size the electron density profiles in different directions were investigated. In this work, we used the same approach as developed earlier [35] and analyzed the electron density profiles in the directions from facet to facet and from vertex to vertex of the reconstructed virus particle (Fig. 6.19). For the particle size estimate we selected the electron density threshold value of 0.2 as it was considered in the SW algorithm during the phase retrieval. From this criterion we determined the particle size in different directions (see Table 6.2). Thus, the obtained mean particle size between facets was  $61 \pm 2$  nm and between vertices  $63 \pm 2$  nm, respectively. These sizes correspond well to the initial range of particle sizes (from 55 nm to 84 nm) considered at the initial classification step (see Section 6.2.4) and other XFEL data performed on the same virus [35, 36].

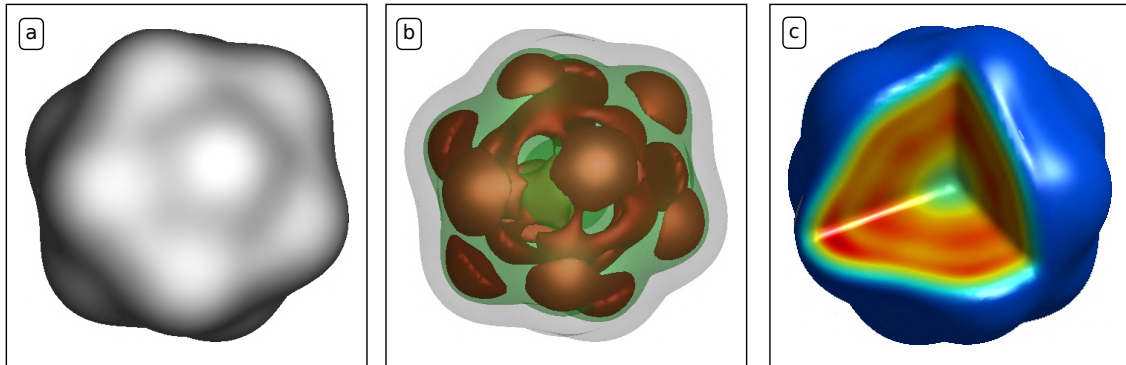


Figure 6.18: Electron density of the reconstructed PR772 virus normalized to the maximum value. (a) The outer structure of the PR772 virus at the isosurface value of 20% of the maximum electron density. (b) The inner 3D structure of the PR772 virus at the isosurface values of 85% (brown area), 75% (green area) and 20% (gray area). (c) A 3D section of the virus.

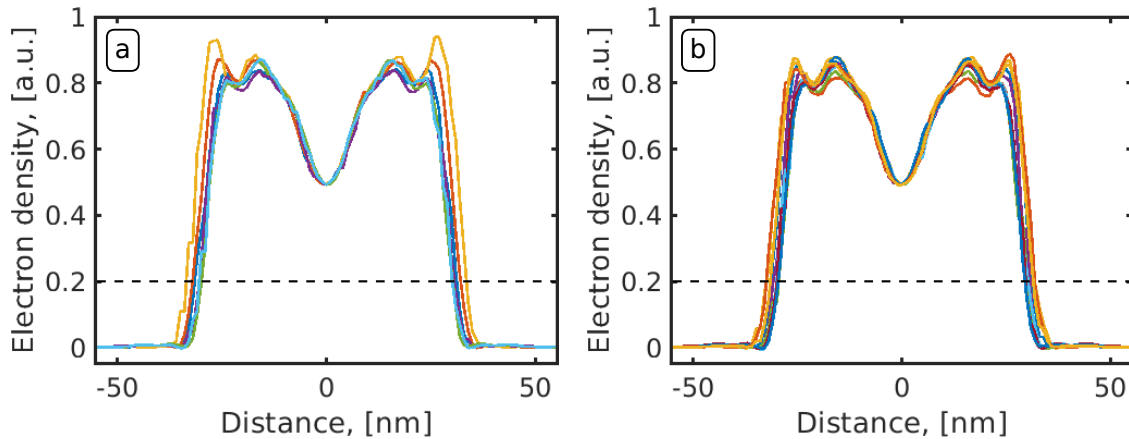


Figure 6.19: Electron-density profiles of the reconstructed virus PR772 normalized to the maximum value for the cut between vertices (a) and facets (b). The horizontal black dashed lines denote a particle-size threshold of 0.2. The mean virus size is 63 and 61 nm for the distance between vertices and between facets, respectively.

Similar to previous studies on PR772 [35, 36] we observed also a certain elongation of the particle shape which might be inherently present in the viruses in solution or appear due to the aerosol injection conditions. We defined elongation of the particle by the following measure

$$\alpha = \frac{D_{max} - D_{min}}{D_{mean}}, \quad (6.7)$$

where  $D_{max}$ ,  $D_{min}$ , and  $D_{mean}$  are maximum, minimum, and mean particle sizes, respectively. The virus structure obtained in the current work showed elongation value  $\alpha = 11\%$  for sizes taken between vertices which is similar to the result of the previous SPI experiments [35] with  $\alpha = 9\%$ .

The mean capsid thickness was obtained from the Gaussian fitting of the electron density profiles with well pronounced features. The area of fitting was considered according to

## 6.5. PR772 virus structure

Table 6.2: The virus sizes in the directions from facet to facet and from vertex to vertex shown in nm. The mean sizes in each direction are shown in the last row.

	Facet-to-Facet	Vertex-to-Vertex
Sizes	$62 \pm 2$	$63 \pm 2$
	$60 \pm 2$	$64 \pm 2$
	$64 \pm 2$	$67 \pm 2$
	$62 \pm 2$	$61 \pm 2$
	$60 \pm 2$	$60 \pm 2$
	$60 \pm 2$	$61 \pm 2$
	$60 \pm 2$	
	$59 \pm 2$	
	$64 \pm 2$	
	$63 \pm 2$	
Mean size	$61 \pm 2$	$63 \pm 2$

the electron density threshold 0.2, again similar to the SW value during the reconstruction. Fitting the result for the electron density profile is shown in Fig. 6.20. Left and right Gaussian functions correspond to the capsid part of the virus structure. Taking the full width half maximum (FWHM) values of these curves we determined the capsid size to be  $7.6 \pm 0.3$  nm. The thickness of the capsid in recent Cryo-EM studies [173] was identified to be below 10 nm which is in good agreement with the thickness determined in this experiment.

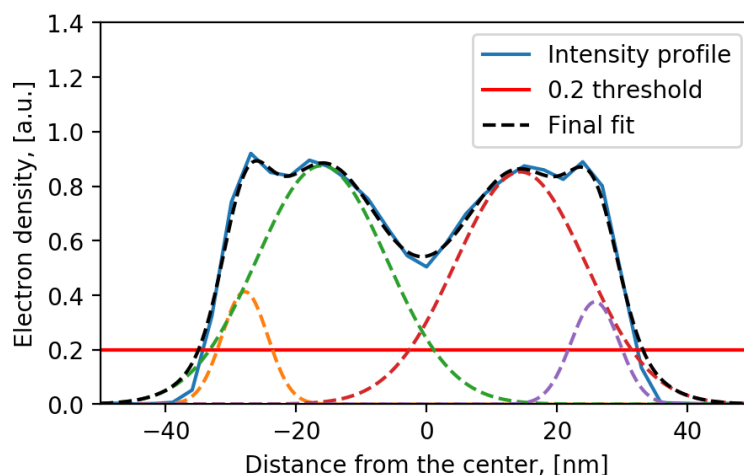


Figure 6.20: Analysis of the electron density profile. For the capsid size estimate fitting of the electron density line profile with four Gaussian functions was performed. Left and right (orange and purple) Gaussian functions correspond to the capsid. The mean size of the capsid was determined as FWHM of these Gaussian functions and is equal to  $7.6 \pm 0.3$  nm.

### 6.5.4 Resolution estimation

Finally, we determined resolution of the reconstructed electron density of the PR772 virus. We used the Fourier Shell Correlation (FSC) approach [166] to determine resolution of the reconstructed virus. For this method two independent sets of reconstruction are required, usually each is based on half of the available data set. This resolution estimation method was described in details in Section 5.7.2. Results of the FSC analysis are shown in Fig. 6.21. To estimate the achieved resolution we used  $\frac{1}{2}$ -bit threshold criteria [167]. Its intersection with the FSC-curve gave a resolution value of 6.9 nm.

The obtained result is better than previously reported value of  $\sim 8$  nm for the same virus [35], though the number of diffraction patterns used for the final analysis was much lower (15% of the previous data set). In this case the resolution is limited by the number of diffraction patterns and moderate scattered intensity where previous reconstructions were limited by the extent of the detector.

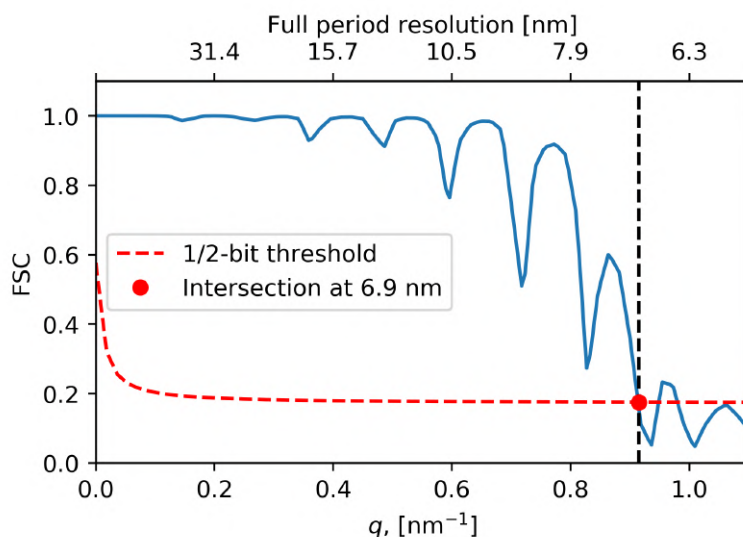


Figure 6.21: FSC of the final reconstruction (blue line) that shows a 6.9 nm resolution (red dot) with a  $\frac{1}{2}$ -bit threshold (red dashed line).

## 6.6 Summary

The implementation of the SPI data analysis workflow from diffraction patterns measured at the AMO instrument at LCLS to reconstruct the electron density of bacteriophage PR772 was presented. We implemented several methods into the workflow including EM-based classification and mode decomposition that were crucial for the high-resolution final reconstruction. Although only half of the detector was operational, implementation of all these steps allowed to determine the PR772 virus structure with a higher resolution as compared to previous SPI studies.

From the initial set of  $1.2 \times 10^7$  experimentally measured diffraction patterns, the final single hit selection set contained 1,085 diffraction patterns. About 53% of this final set was also present in the single hit selection made manually. The number of diffraction patterns classified for further analysis was substantially lower than in the previous experiment [35], improvements in sample delivery to increase the number of diffraction patterns are crucial for advancements in reconstruction resolution.

The combination of all methods implemented in the workflow allowed to obtain the 3D electron density of the virus with the resolution of 6.9 nm based on the FSC analysis, that is better than obtained in the previous studies [35]. The obtained mean PR772 virus size in this experiment was 61 nm (between facets) and 63 nm (between vertices) like in the previous SPI experiments performed on the same PR772 virus [35, 36]. We also observed a similar elongation of about 11% of the virus structure as it was determined in the previous experiments.

This research is another step forward in the SPI data analysis. Implemented methods may become especially important when SPI experiments are performed at high repetition rate XFELs such as the European XFEL [185, 186] and LCLS-II [187] facilities. The first experiments performed at the European XFEL demonstrated the possibility of collecting the SPI data at megahertz rate [131, 138] that might be crucial for the future progress in single particle imaging experiments performed at XFELs.





## Chapter 7

# Classification of diffraction patterns using a Convolutional Neural Network in SPI experiments

Artificial intelligence (AI) and machine learning methods are rapidly becoming an important tool in physics research. We have witnessed an increased interest in these approaches, especially during recent years. This is also related to the large amount of data collected nowadays in experiments not only in particle physics but also in astronomy and X-ray physics. For example, petabytes of data can easily be collected within just a few days at a single beamline of the megahertz European XFEL [185]. Machine learning approaches can help to use this enormous quantity of data effectively.

Typical SPI data analysis pipeline can highly benefit from applying machine learning techniques as it was described in Sections 5.5.1, 5.5.2, 5.5.3, 5.5.4, in particular, the step of single-hit classification. Fig. 7.1 shows the possible implementation of Convolutional Neural Networks (CNNs) in the general scheme. Single hit classification proposed in [38] was based on Expectation-Maximization (EM) method (black arrows in Fig. 7.1). CNN also successfully solve the problem of binary classification "single hit – non-single hit" as it was shown in [118] with the data from the same experiment (blue arrows in Fig. 7.1).

This Chapter is based on Ref. [119]. We develop CNN implementation in the pipeline (red arrows in Fig. 7.1). By classifying single hits first, computationally intensive steps, such as size filtering and EM-based selection, need to be performed on a fraction of the initially collected patterns, saving substantial computational resources. In addition, the proposed scheme allows the classification of newly collected patterns independently, without the need to recompute from the beginning (as would be required by pure EM-based selection). This is particularly useful as experimentalists have the possibility to plan the experiment as it goes and stop it whenever a sufficient number of single hits have been collected, thereby saving time at the XFEL facility.

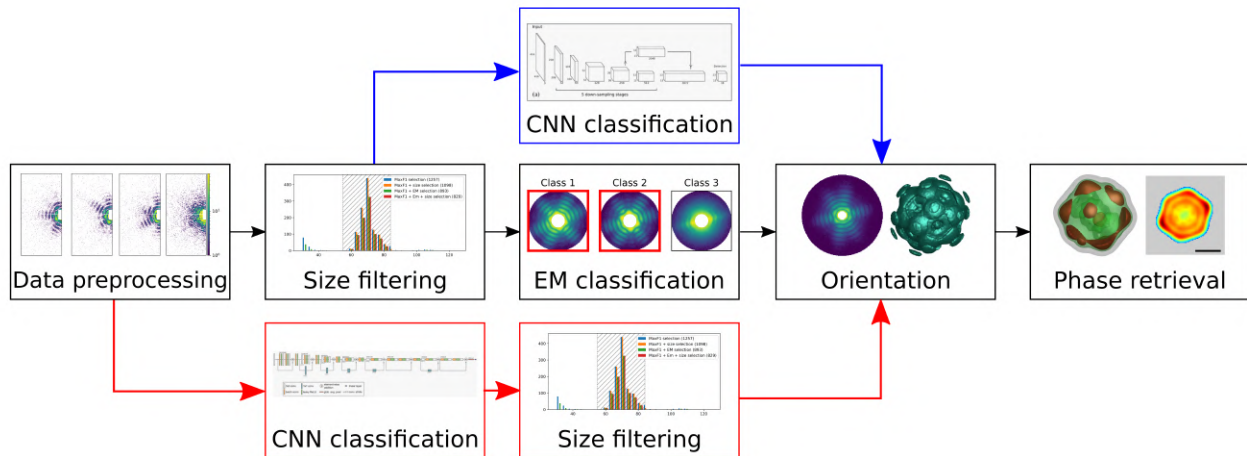


Figure 7.1: Different approaches in SPI analysis workflow. In [38], the steps in black arrows were used. In [118], the steps in blue arrows were used. In [119], the steps in red arrows were used.

We used the data from the same SPI experiment as in Chapter 6. Data description is given in Section 6.1. The total number of diffraction patterns collected during the experiment was  $1.2 \times 10^7$  (data set  $D_0$  in Table 7.1 [172]). Out of those images, only a small fraction contained any scattering patterns. To isolate such patterns, hit finding was performed using the software “psocake” in the “psana” framework [179]. (see Section 6.2 for pre-processing steps). As a result, 191,183 diffraction patterns (data set  $D$  in Table 7.1) were selected as hits from the initial set of experimental data [172]. Manual selection of single-hit diffraction patterns was performed on the data set  $D$  (data set  $D_M$  in Table 7.1) which resulted in 1,393 single-hit diffraction patterns [172]. This selection was used as a ground truth for training and evaluating the CNN in this study. In the previous research [38], we used the EM-classification step to select single-hit diffraction patterns which gave us the 1,085 diffraction patterns ( $D_{EM}$  selection in Table 7.1). We will be comparing CNN single hit selection with EM-based approach.

Table 7.1: Number of diffraction patterns obtained at different SPI analysis steps. This Table mimics the numbers of diffraction patterns in Table 6.1.

Analysis step	Data set name	Number of diffraction patterns
Initial data set	$D_0$	$1.2 \times 10^7$
Hit-finding classification	$D$	191,183
Manual selection	$D_M$	1,393
EM-based classification	$D_{EM}$	1,085

## 7.1 CNN description and architecture

A CNN consists of a succession of convolutional layers, interlaced with nonlinearities. Like most supervised machine learning models, CNNs need to be trained using a set of annotated data stemming from the task that they are intended to solve. As part of the training process, the parameters of the CNN will be tuned to enable it to learn the requested task. Here, the vast majority of parameters are represented by the weights of the convolutional kernels. Training takes place via stochastic gradient descent (SGD) described in Section 5.5.4, where images from the training set are given to the network (forward pass) and the output of the network is compared with the reference annotation through a loss function. Then, the gradients of that loss function with respect to each of the model's parameters are computed (backwards pass) and used to update the weights. This process is repeated many times until the model converges, i.e. the training loss no longer decreases.

The advantage of CNNs over traditional image analysis methods is that the experimenter no longer needs to manually define and compute informative feature representations of the input. This is handled intrinsically by the convolutional layers and learned automatically as part of the training process. As a consequence, CNNs have far greater capabilities in terms of the complexity of tasks they can solve but often require a larger number of annotated example images.

The network architecture used in this work is shown in Fig. 7.2. It is inspired by the pre-activation ResNet-18 [188] and was selected on the basis of initial experiments on the training data set. The network processes patches of size  $192 \times 96$  and is initialized with 16 convolutional filters. The number of filters is doubled with each downsampling up to a maximum of 256. Downsampling is implemented as strided convolution. We use leaky ReLU activation functions [189] and standard batch normalization [190]. The final feature map has a size of  $6 \times 6$  which is aggregated through global average pooling into a vector that is then processed by a linear layer to distinguish single and nonsingle hits.

As evaluation metrics we used precision, recall and the F1 score. These values are defined through true positive (TP), false positive (FP) and false negative (FN) predictions. The definition of the evaluation metrics is as follows

$$P = \frac{TP}{TP + FP}, \quad (7.1)$$

$$R = \frac{TP}{TP + FN}, \quad (7.2)$$

where  $P$  is the precision and  $R$  is the recall metrics. The F1 score is the harmonic mean of the precision and recall:

$$F1 = 2 \frac{PR}{P + R} = \frac{2TP}{2TP + FP + FN}, \quad (7.3)$$

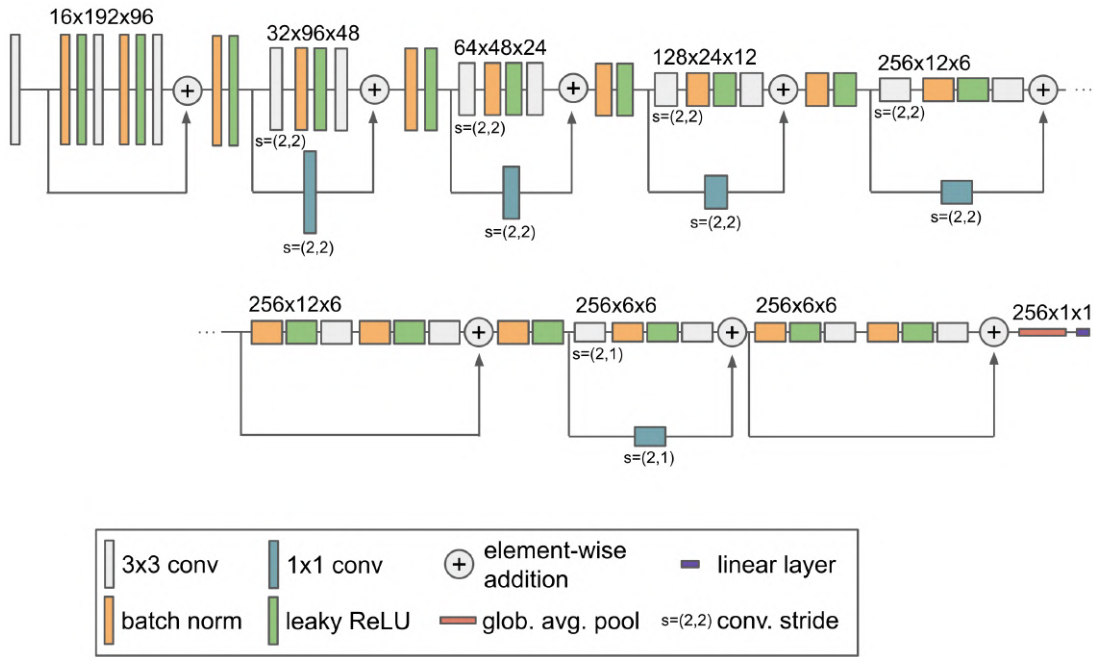


Figure 7.2: Network architecture. We use a pre-activation ResNet-inspired architecture. It takes patches of size  $192 \times 96$  as input and processes them in a sequence of eight pre-activation residual blocks. Downsampling is implemented via strided convolution. The architecture is initialized with 16 filters and doubles the number of filters with each downsampling operation up to a maximum of 256. Global average pooling reduces the final feature representation (shape  $6 \times 6$ ) to a vector that is then used by the classification layer to distinguish single from non-single hits. The size of the feature representations is indicated above each residual block.  $16 \times 192 \times 96$  here denotes 16 convolutional filters with a feature representation of size  $192 \times 96$ .

Owing to the pronounced class imbalance in our data set (a small number of single hits in comparison with a large number of non-single hits), we mainly use the F1 score for evaluating our models. In addition, we report the number of single hits.

## 7.2 Training, validation and test procedure in CNN classification

We use a training data set that is representative of the modified workflow introduced in previous Section, where the experimentalist identifies a limited number of single hits at the beginning of the experiment. Taking into account the annotation effort that would be required, we chose to use 100 single hits and a number of non-single hits that corresponds to the number of images the experimentalist would have seen until the required number of single hits was collected (see Table 7.2). In accordance with the class ratio of the data set used here (approximately 1:200), our training set  $D_{tr}$  consists of 100 single and 19,900 non-single hits. All hits were sampled randomly without replacement. We used the manual selection  $D_M$  as a ground truth.

## 7.2. Training, validation and test procedure in CNN classification

Table 7.2: Number of diffraction patterns for training and test of CNN.

	Data set name	Single hits	Non-single hits
Training and validation data set	$D_{tr}$	100	19,900
Test data set	$D_{test}$	1293	169,890

To prepare our data for the CNN, all diffraction patterns were cropped to the region of interest of size  $192 \times 96$  pixels (Fig. 7.3). All images were normalized by subtraction of the training-data-set (20,000 data) mean value  $\mu = 0.342$  and divided by the standard deviation of the same data set  $\sigma = 2.336$ .

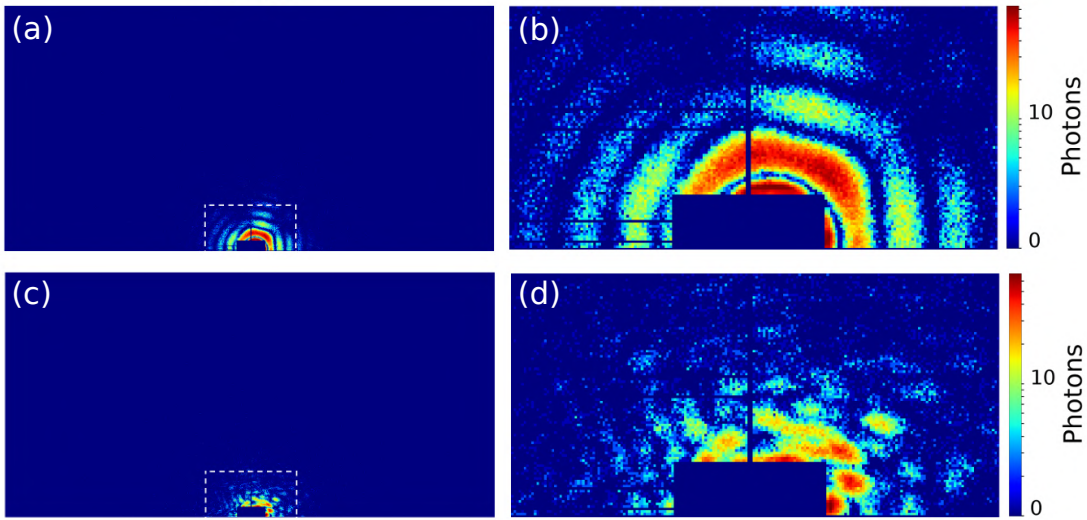


Figure 7.3: (a), (c) Illustration of data cropping before sending to the input of a CNN: single hit example (a), non-single hit example (c). The center of diffraction pattern is located around the center of the bottom part of the only operational detector plane with dimensions in pixels  $1024 \times 512$ . The area surrounding the center of each diffraction pattern with dimensions in pixels  $192 \times 96$  is cropped (white dotted rectangle). (b), (d) The cropped part is used as an input of a CNN: single hit example (b), non-single hit example (d). The diffraction patterns are shown in logarithmic scale.

We trained the network with stochastic gradient descent using the Adam optimizer [191], a minibatch size of 64 and an initial learning rate of  $10^{-4}$ . The standard cross-entropy loss function was used. Samples within minibatches were sampled randomly with replacement. We modified the sampling probabilities such that on average 2% of the presented samples are single hits. We defined an epoch as 50 training iterations and trained for a total of 1,000 epochs (50,000 iterations). The learning rate was reduced each epoch according to the polynomial-learning rate schedule presented in [192].

### 7.2.1 Polynomial learning rate (polyLR) policy

Learning rate is one of the most important hyper-parameters in any neural network optimization process. It controls the speed of network convergence in the training process. Min-

imization of loss function was done with SGD which first computes the gradients of the loss function with respect to all model parameters using an algorithm called back-propagation and then updates the model weights  $w$  as follows

$$w^{i+1} = w^i - \eta \cdot \frac{\partial L}{\partial w}, \quad (7.4)$$

where  $L$  is the loss function,  $i$  is iteration number,  $\eta$  is learning rate. A conventional approach to control convergence of a model is to set an initial value of learning rate and let it decrease over time. Here we use a learning rate scheduler called polynomial learning rate policy (polyLR) [192]. The learning rate is changed during training according to the equation

$$\eta = \eta_0 \cdot \left(1 - \frac{i}{t_i}\right)^{power}, \quad (7.5)$$

where  $t_i$  is the total number of iterations during training.

## 7.2.2 Data augmentation

Owing to the limited number of training cases, extensive data augmentation is performed on the fly during training using the "batchgenerators" framework [193]. Data augmentation is a powerful tool to improve the robustness of models trained on a limited number of training cases. By running transformations on the training cases, new images are generated that direct the models to learn better generalizing features and thus ultimately improve their generalization capabilities on the test set. Specifically, we used random rotations, scaling, elastic deformation, gamma augmentation, Gaussian noise, Gaussian blur, mirroring, random shift, and cutout [194].

Random rotation is a common augmentation technique when a source image is rotated clockwise or counterclockwise by some number of degrees. This changes the position of the object in the image. In random rotation of the image its corners are cut off, after rotation the new corners are filled with padding.

Scaling can be done outward or inward. When scaling inward, the resultant image size is larger than the original image size. A section is cut out from the resultant image to make the size equal to the original image. When scaling outward, the size of the image is reduced, the missing part is filled with padding.

Obtaining an augmented image using elastic deformations is done in two parts. First, a random stress field is generated for horizontal and vertical directions with randomly sampled values

$$\Delta_x = G(\sigma) \cdot (\alpha \cdot \text{Rand}(n, m)), \quad (7.6)$$

$$\Delta_y = G(\sigma) \cdot (\alpha \cdot \text{Rand}(n, m)), \quad (7.7)$$

where  $G(\sigma)$  is the strength of the smoothing operation given by the standard deviation of the Gaussian filter  $\sigma$ ,  $\sigma$  is a parameter defining the maximum value for the random initial displacement,  $n$  and  $m$  are the image dimensions. After that, the stress field is applied to the image by moving each pixel to a new position using spline interpolation of order one to obtain pixel values at integer coordinates

$$I_{deformed}(j + \Delta_x(j, k), k + \Delta_y(j, k)) = I(j, k), \quad (7.8)$$

where  $I$  and  $I_{deformed}$  are the initial and deformed images,  $j$  and  $k$  are pixel coordinates.

Gamma augmentation is a nonlinear operation used to encode and decode luminance in images, it is defined by power-law expression

$$V = AU^\gamma, \quad (7.9)$$

where  $V$  is resultant pixel value,  $U$  is initial pixel value,  $A$  is a constant,  $\gamma$  is a parameter.

Gaussian noise is an additive noise type, where the intensity value in a pixel with the coordinates  $(x, y)$  for the noisy image is given by the expression

$$N(x, y) = A(x, y) + B(x, y), \quad (7.10)$$

$$N(x, y) = A(x, y) + B(x, y), \quad (7.11)$$

where the  $A(x, y)$  in the pixel value of the original image,  $B(x, y)$  is the added noise. The added value of noise is defined by probability density function of Gaussian random value is indicated in equation

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ \frac{-(z - \mu)^2}{2\sigma^2} \right], \quad (7.12)$$

where  $\sigma$  and  $\mu$  are standard deviation and mean values,  $z$  is pixel value.

Gaussian blur is a type of image-blurring filter that uses a Gaussian function for calculating the transformation to apply to each pixel in the image. The response of the Gaussian filter in two dimensions is described by

$$g(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ \frac{x^2 + y^2}{2\sigma^2} \right], \quad (7.13)$$

where  $x$  and  $y$  are the distances from the filter origin in the horizontal and vertical directions, respectively,  $\sigma$  is the standard deviation of the Gaussian distribution.

Mirroring implies flipping images along vertical and horizontal axes.

Random shift is a transformation when the image as a whole is shifted horizontally and vertically by a random number of pixels. The missing parts at the edges appeared due to these shifts being filled with padding.



Cutout is a data augmentation technique that randomly masks out square regions in images. These regions are of random size, appear in random positions in the image and filled with padding.

### 7.2.3 *K*-fold cross-validation

During method development, our models were trained and validated through stratified fivefold cross-validation on the set of 20,000 training examples.

Cross-validation is a procedure used to evaluate machine learning models on a limited data set size, i. e. the amount of data is too small to draw robust conclusions using a conventional training and validation split. The procedure of *k*-fold cross-validation is the following. The entire data set available for training and validation is shuffled and split into *k* groups. For each unique group, the data from this group becomes the validation data set; the respective training data set consists of the other *k* - 1 groups. As a result, there are *k* individual trained models. The performance metrics are then defined by average performance of these models.

We chose *k* = 5 for developing our models. Final performance on the test set is obtained by using the resulting five models as an ensemble, as described in the following section.

### 7.2.4 Ensembling via softmax averaging

Ensembling refers to combining predictions from multiple machine learning models. It is a commonly used strategy to reduce the variance of the models and increase the overall quality of the predictions. In the case of image classification, ensembling can be implemented via softmax averaging.

Here, this is implemented in the following way: each CNN model issues a prediction for each diffraction pattern of the test data set providing single hit probability (ranging from 0 to 1). The average of five predictions, one for each model, is then the single hit probability for the ensemble. We put a threshold for the average single hit probability to obtain the final prediction. Diffraction patterns with final probability above 0.5 are classified as single hits.

### 7.2.5 Inference

For model development we used stratified fivefold cross-validation on the training set. The resulting five models are used as an ensemble for test set predictions. We report final results on the test set  $D_{test}$  consisting of the 171,183 remaining patterns (1,293 single and 169,890 non-single hits). We further use test-time data augmentation. Ensembling was implemented via softmax averaging, followed by thresholding at 0.5 to obtain the final predictions.

## 7.3 CNN variant: identifying more single hits

The CNN model described above is optimized for maximizing the F1 score on our training cross-validation. We subsequently refer to it as "MaxF1". In addition, we trained a second CNN model that predicts a larger number of single hits ("moreSH") and leans more towards higher recall values. To achieve that, we made modifications to the sampling strategy as well as the loss function. Specifically, we increased the probability of selecting single hits when constructing the minibatches from 2 to 5% and made use of a weighted cross-entropy loss which weights samples of ground-truth single hits higher during loss computation (weights 0.1 and 0.9 for non-single hits and single hits, respectively). For both models (MaxF1 and moreSH), we used the same augmentation and inference scheme.

## 7.4 Particle size determination

Particle size filtering is also an important part of the SPI data analysis workflow (see Fig. 7.1). It can help to remove unnecessary diffraction patterns corresponding to other particles apart from the viruses under investigation. In the previous approach (Fig. 7.1, black arrows), particle size determination was carried out on the entire data set  $D$  prior to applying the EM classification, and thus the single-hit classification was performed only on particle sizes between 55 and 84 nm [38].

In this work we used the CNN classification after the initial preprocessing step and particle size filtering was applied afterwards. Particle size determination was implemented by fitting the Power Spectral Density (PSD) function, i.e. the angular averaged intensity, of each diffraction pattern with the PSD of diffraction pattern from the spherical particles in a range of sizes from 30 to 300 nm and was described in [38] and in Section 6.2.4. Here we used the same results, and the same virus size range (55 – 84 nm) was considered here.

## 7.5 Results

### 7.5.1 CNN performance

Table 7.3 summarizes the performance of our CNNs on the training set cross-validation. The MaxF1 configuration obtains balanced precision and recall and an F1 score of 0.645. The number of predicted single hits (120) is close to the number of single hits (100) in this data set. The moreSH configuration, however, trades a higher recall with lower precision, resulting in an overall decreased F1 score of 0.536. As expected, the number of predicted single hits is higher, being 221 in this case.

Table 7.3: Five-fold cross-validation results ( $N = 20,000$  training samples).

	MaxF1	moreSH
F1 score	$0.645 \pm 0.074$	$0.536 \pm 0.018$
P (precision)	$0.591 \pm 0.062$	$0.391 \pm 0.023$
R (recall)	$0.710 \pm 0.096$	$0.960 \pm 0.065$
Predicted single hits	120	221

Test set predictions (see Table 7.4) were obtained by ensembling the five models obtained during cross-validation (as described in Section 7.2). On the test set (171,183 patterns), the MaxF1 configuration obtained an F1 score of 0.731 with balanced precision and recall. Interestingly, the F1 score is substantially higher than that on the training set cross-validation which we attribute to the use of ensembling. The predicted number of single hits (1,257 patterns) is close to the number of single hits (1,393 patterns) in the reference set  $D_M$ .

Table 7.4: Test set results ( $N = 171,183$  test samples).

	MaxF1	moreSH
F1 score	0.731	0.644
P (precision)	0.741	0.522
R (recall)	0.721	0.841
Predicted single hits	1,257	2,086

The moreSH configuration, as expected, again displays an imbalance between precision and recall. Overall, its recall is higher (0.841 versus 0.721), but its F1 score is lower at 0.644 (versus 0.731). Again, as expected, the number of predicted single hits is larger (2,086 patterns).

On a workstation equipped with an AMD Ryzen 5800X CPU, 32 GB of RAM and an Nvidia RTX 3090 GPU, training each individual model took less than 25 min ( $< 2.5$  h for all five models in the cross-validation). The inference speed was  $\sim 450$  diffraction patterns per second for the ensemble and with test time data augmentation (five models and mirroring along all axes for a total of 20 predictions per pattern). Predicting the 171,183 test patterns took less than 7 min. If faster inference is required, single-model prediction without test-time augmentation can be used to increase the throughput to  $\sim 8700$  patterns per second. Training required merely 3.5 GB of VRAM, and a much smaller GPU than the RTX3090 used here would have been sufficient as well. The code for training the CNN and running predictions on our test set is available in Ref. [195], data available in Ref. [196].

### 7.5.2 PSD comparison, EM and particle size filtering

As a result of single-hit classification, we obtained data selections with different numbers of diffraction patterns. In order to compare these selections, we plotted and analyzed the PSD function. To quantify the contrast values of the PSD functions for each selection, we introduced contrast metric which describes the mean difference between the local minima and maxima (similar as in Eq. (6.3) and Eq. (6.4)) over the first three pairs

$$\gamma = \frac{1}{N} \sum_{i=1}^N \frac{I_{max} - I_{min}}{I_{max} + I_{min}}, \quad (7.14)$$

where  $N = 3$  is the number of pairs, and  $I_{max}$  and  $I_{min}$  are values of the PSD function for the maxima and minima, respectively. By looking at the PSD functions and the corresponding contrast values we can compare various single-hit selections and analyze which one has more features.

As a result of CNN classification, we obtained two data sets: MaxF1 and moreSH with the number of single-hit diffraction patterns 1,257 and 2,086, respectively (see Table 7.4). Plotted PSD functions for both selections are shown in Fig. 7.4 (a), (c) (blue dashed lines). Additionally, we plotted the PSD functions for the  $D_M$  and  $D_{EM}$  selections [38], containing 1,393 and 1,085 diffraction patterns, respectively (purple and brown solid lines).

The corresponding number of diffraction patterns and PSD contrast values for all four data sets (MaxF1, moreSH,  $D_M$  and  $D_{EM}$  selection) are given in Table 7.5. From Fig. 7.4 (a), (c) we observe the same number of fringes as in our previous paper. However, the contrast values were lower in the case of CNN classification in comparison with EM classification. As expected, the PSD functions for MaxF1 and moreSH mimic the behaviour of the PSD function of the  $D_M$  selection which was used as the ground truth for CNN training.

Table 7.5: Number of diffraction patterns in different data sets of single hits and PSD contrast values for each of them.

Data set	No. of diff. pat.	PSD contrast
MaxF1	1,257	0.63
MaxF1 + EM	893	0.64
MaxF1 + size selection	1,098	0.64
MaxF1 + EM + size selection	829	0.64
moreSH	2,086	0.59
moreSH + EM	1,204	0.64
moreSH + size selection	1,617	0.62
moreSH + EM + size selection	1,090	0.65
$D_M$	1,393	0.59
$D_{EM}$	1,085	0.71

In order to increase the PSD contrast of the CNN selection, we applied EM-based selection to the MaxF1 and moreSH data sets. This method was described in details in Section 5.5.3. The EM classification algorithm is designed to distribute the whole data set into a predefined number of clusters. On each iteration, probabilities of patterns to be assigned to each cluster are calculated and cluster models are updated by weighted averaging of the associated patterns, where weights are determined by obtained probabilities. After the algorithm converges, one can manually select the required clusters which correspond to the particle under investigation.

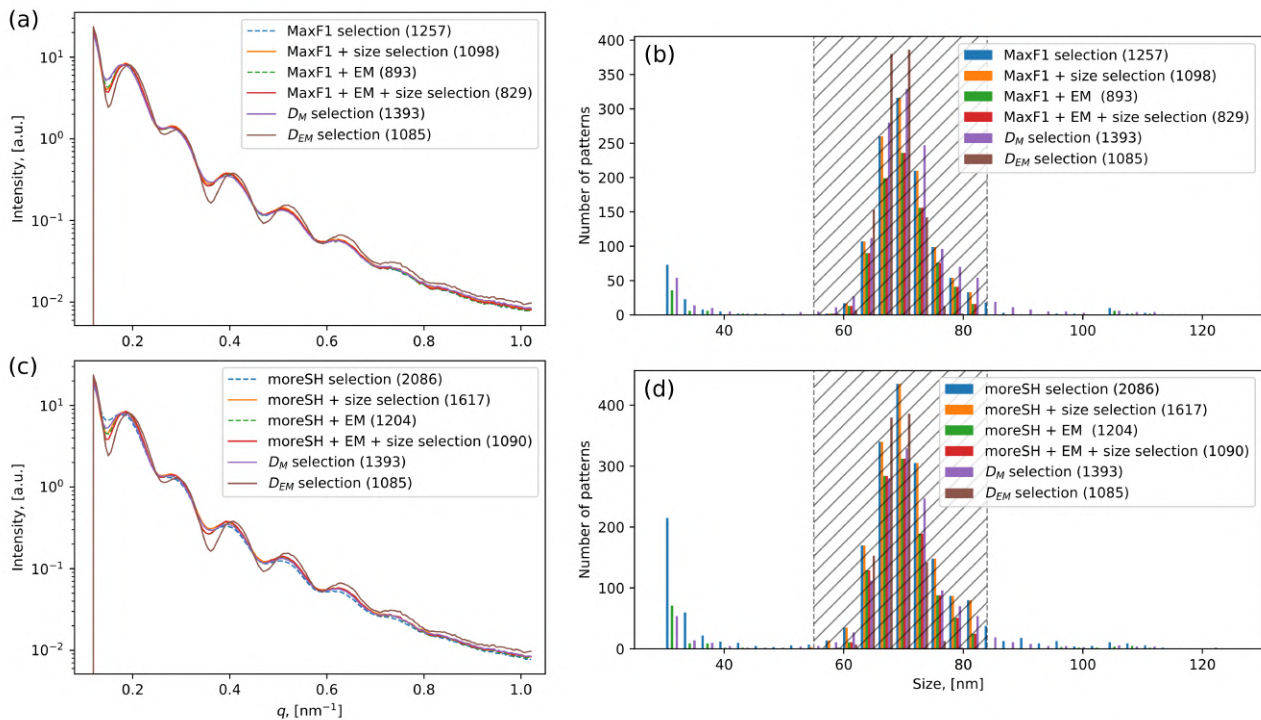


Figure 7.4: (a), (c) PSD functions for the different data sets: for the MaxF1 data selection (a) and for the moreSH data selection (c). (b), (d) Particle size histograms for different data sets: for the MaxF1 data selection (b) and for the moreSH data selection (d). Notations are the following: blue line/histogram – the whole selection, orange line/histogram – selection with size filtering applied, green line/histogram – selection with the EM algorithm applied, red line/histogram – selection with the EM algorithm and size filtering applied. All panels contain the PSD functions/histograms of the  $D_M$  (purple line/histogram) and  $D_{EM}$  (brown line/histogram) selections. In panels (b) and (d), the dashed areas indicate the particle size range from 55 to 84 nm. In the legend, the number of diffraction patterns for each selection is given in brackets.

If one considers the contrast of the PSD function as a criterion for best reconstruction, the EM algorithm outperforms CNN classification. EM-based algorithm was applied to the diffraction patterns selected by CNN: MaxF1 and moreSH data sets, containing 1,257 and 2,086 patterns respectively. Both selections were distributed into 20 classes (example of distribution for MaxF1 data set is in Fig. 7.5) and after 10 iterations of the algorithm, the obtained classes were inspected. Some of them clearly contained diffraction patterns of the

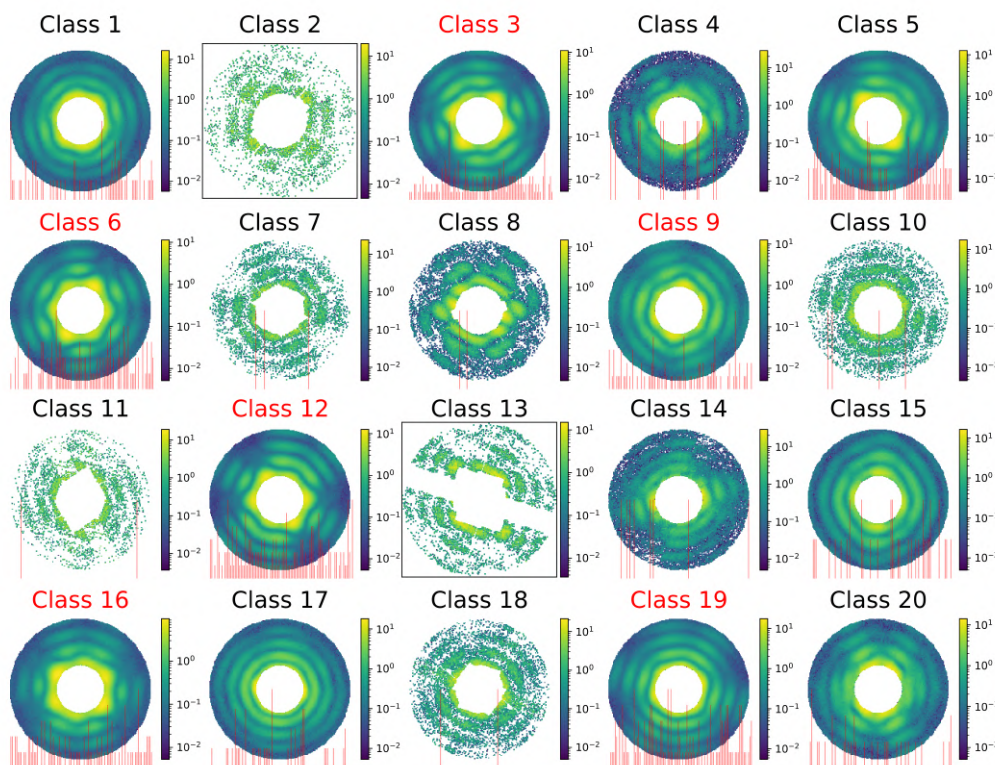


Figure 7.5: EM-based classification of single hit diffraction patterns for MaxF1 data set. Data were distributed into 20 classes, Classes 3, 6, 9, 12, 16, 19 were selected as containing diffraction patterns of PR772. These classes contain 893 patterns in total.

virus PR772 and the rest ones contained other scattering. Classes of interest were selected manually by the 6-fold symmetry expected from the virus. In the case of the MaxF1 data set, classes 3, 6, 9, 12, 16, 19 (highlighted with red title in Fig. 7.5) were considered to contain patterns of interest.

The PSD results of this additional selection are summarized in Fig. 7.4 (a), (c) (green dashed lines) and the final numbers of the diffraction patterns before and after applying EM-based algorithm are presented in Table 7.5 with notation "+ EM". The contrast for moreSH + EM selection showed a substantial improvement (0.64 versus 0.59 without EM), and we also observed a slight improvement for the MaxF1 + EM selection (0.64 versus 0.63 without EM). At the same time, the EM selection [38] still has the best result in terms of contrast.

The EM classification carried out in [38] was performed on a size range of viruses from 55 to 84 nm which was determined prior to EM classification. To perform particle size analysis in this study, we first plotted histograms of the particle size distribution for each data set (MaxF1 with/without EM algorithm applied, moreSH with/ without EM algorithm applied) in Fig. 7.4 (b), (d). Each data selection consists of diffraction patterns within a wide size range. This means that, even after single-hit classification (with/without EM algorithm), the data sets contain diffraction patterns that correspond to particles of different sizes.

To be consistent with the previous research, the size range from 55 to 84 nm was considered for further analysis and particle size selection was applied. The corresponding PSD functions are plotted in Fig. 7.4 (a), (c) (solid orange and red lines), and the resulting numbers of diffraction patterns and contrast values are summarized in Table 7.5 with notation "+ size selection".

Fig. 7.4 (a) and Table 7.5 show that for the MaxF1 data set the particle size filtering did not change the contrast values (= 0.64). However, for the selection moreSH with the EM algorithm applied the particle size filtering Fig. 7.4 (c) gave the best PSD contrast value (= 0.65).

Even though we were able to increase the PSD contrast through different classification strategies and particle size filtering, we, unfortunately, reduced the number of diffraction patterns along the way. For the MaxF1 data set we started from a data set of 1,257 patterns and finally came to 829 patterns. For the moreSH selection, we started with 2,086 patterns and finally came to 1,090 patterns. In the context of our data processing pipeline, where a large number of single hits is required to get reliable results, this can be detrimental.

In the following, we will consider four final data sets:

- MaxF1 with size filtering applied: Fig. 7.4 (a), orange solid line; Fig. 7.4 (b), orange histogram;
- MaxF1 with the EM algorithm and size filtering applied: Fig. 7.4 (a), red solid line; Fig. 7.4 (b), red histogram;
- moreSH with size filtering applied: Fig. 7.4 (c), orange solid line; Fig. 7.4 (d), orange histogram;
- moreSH with the EM algorithm and size filtering applied: Fig. 7.4 (c), red solid line; Fig. 7.4 (d), red histogram.

Below are computing times to obtain  $D_{EM}$  selection by size filtering of 191k diffraction patterns and performing the EM algorithm on 18k patterns in the size range 55 – 84 nm. Size estimation takes 16 min 26 s. It is single threaded and do not really benefit from many cores. Extraction and saving of filtered data take: 20 min 37 s. It is limited by storage read and write speed. EM classification takes 26 min 16 s for 10 iterations. For 5 classifications it is 2 h 11 min 20 s. Calculations were performed on a computer cluster node (max-exfl027) with 2 Intel E5-2698 v4 @ 2.20GHz. It is 40 cores and 80 threads total. The node also has 512GB of memory, but it is barely used by EM.

### 7.5.3 Intersection over union comparison

To compare different data selections, we also looked at the intersection over union  $\alpha$  metric which can be described as

$$\alpha = \frac{A \cap B}{A \cup B}. \quad (7.15)$$

Here  $A$  and  $B$  are two sets of data, and signs  $\cap$  and  $\cup$  mean intersection and union of these two data sets.

Table 7.6: Number of diffraction patterns in intersections of different pairs of data sets. The initial number of diffraction patterns in the sets is shown in brackets. In the second line, the intersection over union  $\alpha$  is shown.

	MaxF1 + size selection (1098)	MaxF1 + EM + size selection (829)	moreSH + size selection (1,617)	moreSH + EM + size selection (1,090)	$D_M$ (1,393)	$D_{EM}$ (1,085)
MaxF1 + size selection (1098)	1,098 – 100%	829 – 75%	1097 – 68%	878 – 67%	875 – 54%	575 – 36%
MaxF1 + EM + size selection (829)	829 – 75%	829 – 100%	829 – 51%	730 – 61%	678 – 44%	485 – 34%
moreSH + size selection (1,617)	1097 – 68%	829 – 51%	1617 – 100%	1090 – 67%	1006 – 50%	686 – 34%
moreSH + EM + size selection (1,090)	878 – 67%	730 – 61%	1090 – 67%	1090 – 100%	791 – 47%	651 – 43%
$D_M$ (1,393)	875 – 54%	678 – 44%	1006 – 50%	791 – 47%	1393 – 100%	574 – 30%
$D_{EM}$ (1,085)	575 – 36%	485 – 34%	686 – 34%	651 – 43%	575 – 30%	1085 – 100%

The values obtained for different pairs of data sets are shown in Table 7.6. We also calculated the intersection over union over three selections – MaxF1 with size filtering applied, moreSH with size filtering applied and  $D_{EM}$  selection – which gave the intersection over union  $\alpha = 29\%$  with 575 diffraction patterns in the intersection. Another three selections – MaxF1 with EM algorithm and size filtering applied, moreSH with the EM algorithm and size filtering applied, and  $D_{EM}$  selection – gave the intersection over union  $\alpha = 29\%$  with 469 diffraction patterns. We think that this choice of diffraction patterns in the intersection



of three data selections is providing us with the most important diffraction patterns that contain the features of virus structure from all data selections.

### 7.5.4 Orientation determination

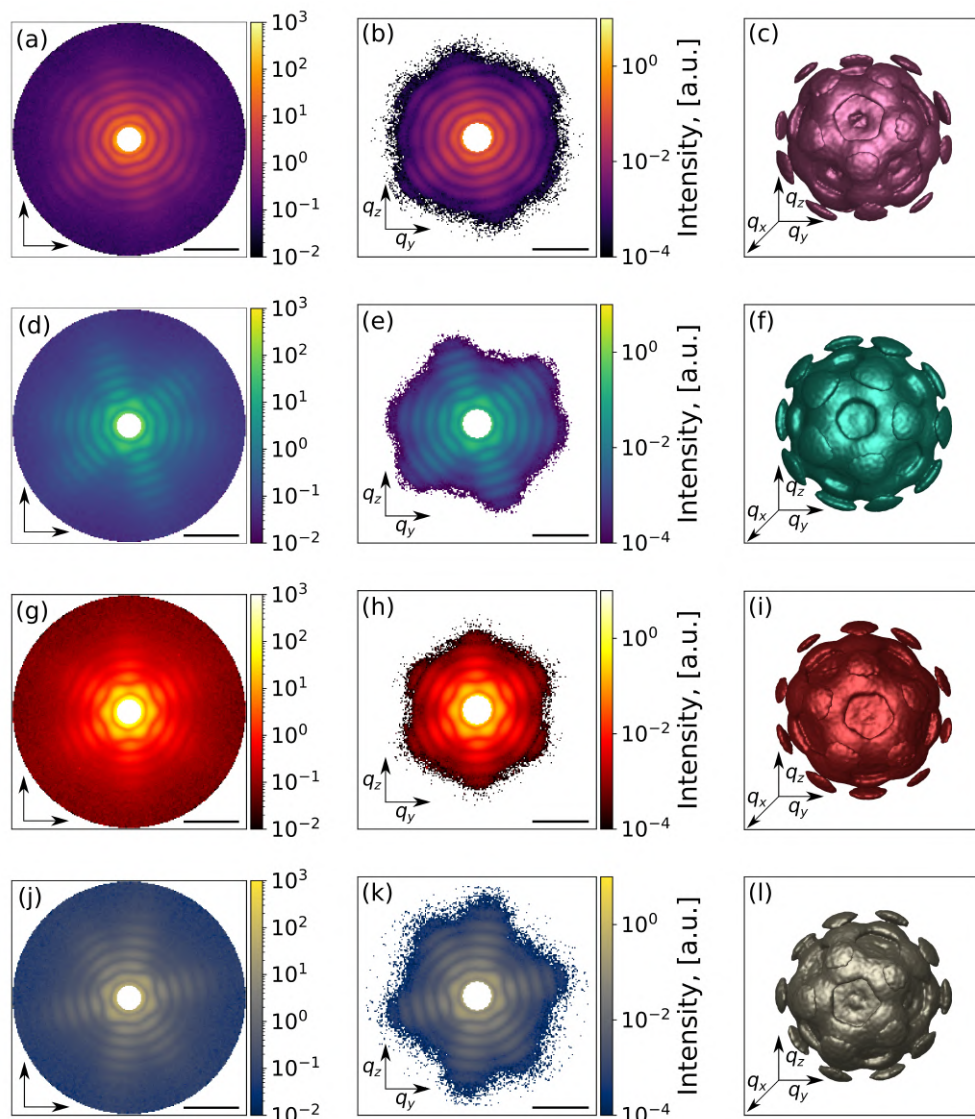


Figure 7.6: 2D central slice of 3D intensity distribution for MaxF1 with the size filtering applied (a), MaxF1 with the EM algorithm and size filtering applied (d), moreSH with the size filtering applied (g), moreSH with the EM algorithm and size filtering applied (j). Vertical and horizontal axes denotes  $q_z$  and  $q_y$  directions, respectively. 2D cuts and reciprocal volume after background subtraction for MaxF1 with the size filtering applied (b)-(c), MaxF1 with the EM algorithm and size filtering applied (e)-(f), moreSH with the size filtering applied (h)-(i), moreSH with the EM algorithm and size filtering applied (k)-(l). Black scale bar in denotes  $0.5 \text{ nm}^{-1}$ .

The next step of the workflow for SPI analysis after single-hit classification is orientation determination of the diffraction patterns (Fig. 7.1). The Expand–Maximize–Compress algorithm [136] in Dragonfly [137] was used to retrieve the orientation of each diffraction pattern

and to combine them into one 3D intensity distribution of the PR772 virus (described in Section 5.6).

We retrieved the orientation of all previously selected data sets with the size filtering applied, with and without the EM classification. Visual inspection does not allow us to see a significant difference between data sets (MaxF1 and moreSH with/without the EM algorithm applied, and with size filtering applied). However, for all four data sets the background at high  $q$  values is clearly seen as in [38]. Here we used the same approach for background subtraction – we defined the level of the background as the mean signal in the high  $q$  region, where the presence of meaningful signal from the particle is negligible. The orientation determination results before/after background subtraction on four data sets is shown in Fig. 7.6.

### 7.5.5 Phase retrieval and reconstructions

The next and the final step in our workflow is phase retrieval and reconstruction of the electron density of our virus particle from the 3D reciprocal space data (Fig. 7.1). Since the experimental measurements provide only the amplitude of the complex-valued scattered wavefield, we applied iterative phase retrieval algorithms (Section 4.3) in order to determine the 3D structure of the virus particle.

We proceeded in the same way as in Section 6.5.1 and [38]. The phase retrieval procedure consisted of two steps. In the first step, the central gap in the 3D intensity map of the virus that originated from the masking of the initial 2D diffraction patterns was filled. In the second step, the 3D intensity maps with the filled central part were used to perform phase retrieval. We first performed 50 reconstructions for each intensity map and then used mode decomposition (Section 6.5.2) to determine the final 3D electron density structure of the virus.

The final virus structure for each data selection, obtained in the described way, is shown in Fig. 7.7. All expected features are present in these reconstructions: the icosahedral structure of the virus, higher density in the capsid part of the virus and reduced density in the central part.

The resolution was evaluated by the Fourier Shell Correlation (FSC) method. Obtained FSC resolution for all four data sets (MaxF1 and moreSH with/without EM algorithm applied, with/without size filtering applied) fluctuates from 5.8 nm to 8 nm and is shown in Table 7.7. Applied EM algorithms for the CNN-based classification could improve the reconstruction result by several nanometers in terms of FSC resolution. And CNN-based single hit diffraction patterns classification by itself with size filtering applied could give quite good resolution.

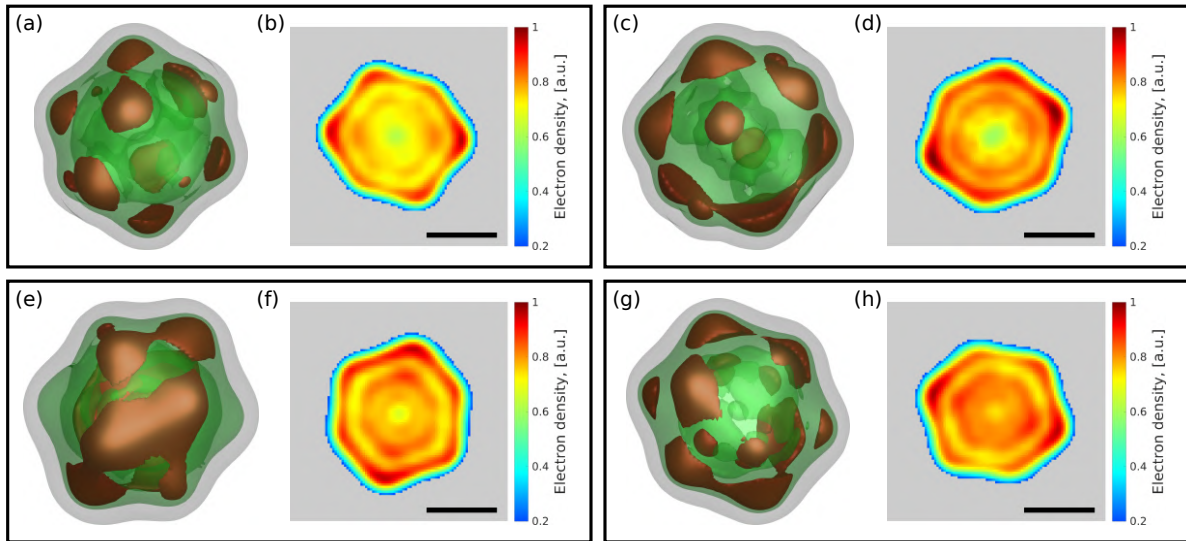


Figure 7.7: PR772 virus reconstructed from the different data sets. (a)–(d) Reconstruction of single-hit diffraction patterns selected by MaxF1 with size filtering applied (a), (b) and MaxF1 with the EM algorithm and size filtering applied (c), (d). (e)–(h) Reconstruction of the single-hit diffraction patterns selected by moreSH with size filtering applied (e), (f) and moreSH with the EM algorithm and size filtering applied (g), (h). (a), (c) The 3D inner structure of the virus with 88% (brown), 75% (green) and 20% (grey) levels of intensity for the MaxF1 selections. (e), (g) The 3D inner structure of the virus with 86% (brown), 75% (green) and 20% (grey) levels of intensity for the moreSH selections. (b), (d), (f), (h) 2D slices of the corresponding structure with the same scale bar of 30 nm. For visual representation, each virus structure was upsampled three times.

Table 7.7: FSC resolution for different data selections.

Data set	FSC resolution, nm
MaxF1 + size selection	8
MaxF1 + EM + size selection	5.8
moreSH + size selection	6.4
moreSH + EM + size selection	5.9

In comparison with the previous EM selection [38] with 6.9 nm resolution, the results obtained in this work showed overall agreement in virus structure (Fig. 7.7) and FSC resolution, the difference varies  $\pm 1$  nm. The best result appeared to be MaxF1 with the EM algorithm and the size filtering applied selection – with the FSC resolution of 5.8 nm. Corresponding inner structure (Fig. 7.7 (c)) and 2D central slice (Fig. 7.7 (d)) demonstrated only slight variance from the previous work [38] (see Fig. 6.18).

The slightly higher resolution determined in this work relative to our previous work (6.9 nm in Chapter 6) may be related to the comparatively small number of diffraction patterns used in the FSC method. As we observe in Fig. 7.7 (a)–(d), the electron densities of the virus in the CNN MaxF1 selection with size filtering and MaxF1 selection with EM selection plus size filtering are practically identical. We see small differences from the previous electron

density in the CNN moreSH selection with size filtering and moreSH with EM selection plus size filtering (Fig. 7.7 (e)–(h)). At the same time, the central slice in all four reconstructions (Fig. 7.7 (b), (d), (f) and (h)) is practically the same, the capsid layer being the same size. Since we have 400–500 diffraction patterns in common with the considered data selections and our previous work [38], we can assume that these were the ones that contributed to and shaped the final reconstructed results in such a common way for all five data selections.

## 7.6 VGG-style network

The main studies in the field of CNN classification of single hits were carried out with the network architecture pre-activated ResNet-18 described above. In order to investigate an effect of CNN depth required for the specific task of single hit classification, a VGG-style network was implemented within the same pipeline. This network is realized as a plain sequence of convolutional layers organized in four downsampling stages (Fig. 7.8). The activation function is ReLU. Batch normalization layer precedes each convolutional layer, except the first one.

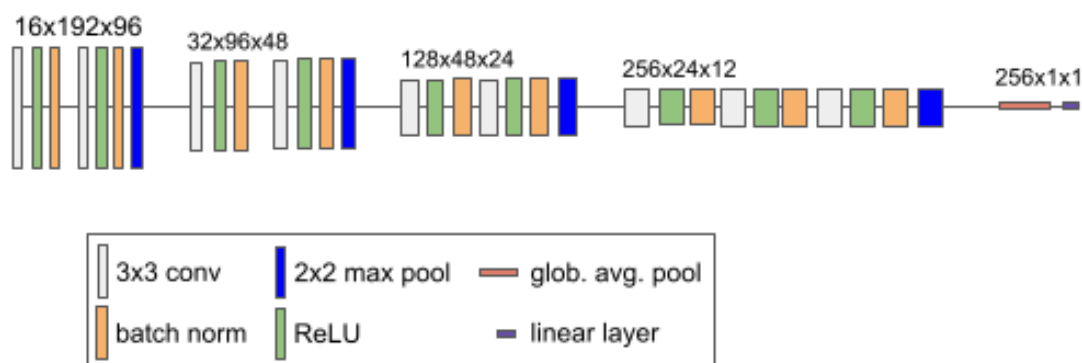


Figure 7.8: VGG-style network architecture. We use a simple VGG-style network for comparison. It has the same input size of  $192 \times 96$ . It processes the input in four downsampling stages. Downsampling is implemented via maximal pooling. The convolutional layer of the first stage has 16 filters. The number grows up to 256 filters for the fourth stage. Global average pooling is used to linearize the final feature representation of the shape  $12 \times 16$  to a feature vector used for classification.

Dimensionality reduction is realized via maximal pooling. The number of filters in the convolutional layers of the first stage is 16. It rapidly grows up to 256 at the last stage. This growth is intentionally fast. It allows to extract more higher level features while preventing the network from growing in depth. Global average pooling is used to linearize the final feature representation of the shape  $12 \times 16$  to a feature vector used for classification.

Training, validation and test follow the same procedure described for ResNet-18. The results for relevant metrics for five-fold cross-validation and test performance metrics for the VGG-style network are shown in Table 7.8. It is similar to that of ResNet-18 (Table 7.4). This

is an indication that the choice of network depth within the investigated limit has negligible effect. Thus, a simple VGG-style network can be sufficient for the task.

Table 7.8: VGG five-fold cross-validation results ( $N = 20,000$  training samples) and test set results ( $N = 171,183$  test samples).

	Cross-validation results	Test set results
F1 score	0.678	0.727
P (precision)	0.656	0.78
R (recall)	0.72	0.681
Predicted single hits	113	1,130

## 7.7 Summary

Our studies with the CNN-based single-hit classification implemented within the SPI data analysis workflow resulted in a reasonable structure reconstruction of the virus PR772 (see Fig. 7.7).

We compared two competing CNN selections, MaxF1, and moreSH. The MaxF1 selection was intended to select single hits with an optimal F1 score. The selection moreSH was optimized for finding more single-hit diffraction patterns (high recall). Both selections were refined by applying the EM algorithm and limiting the selection to particle sizes in the range 55 – 84 nm (Table 7.5). Driven by the need for many single hits in the reconstruction pipeline, the moreSH configuration was conceived with the intention of missing as few single hits as possible; the selection was cleaned up afterwards using EM selection and size filtering, in the hope of achieving a higher resolution than could be obtained with the MaxF1 counterpart.

Unfortunately, this goal was missed: MaxF1 yielded approximately the same resolution even though the moreSH approach resulted in 1,090 selected single hits instead of the 829 found by MaxF1 (with EM and size selection applied). We therefore conclude that optimizing balanced precision and recall through maximizing the F1 score is a suitable target for model development.

CNNs learn from their given training data set. The provided selection [172] which was used for this purpose here, as any other manual selection, may be subjective. In addition, the task of identifying single hits is not necessarily identical to the task of finding the ideal set of patterns needed for reconstruction. In an ideal world, the CNNs should be trained with the patterns ideally suited for reconstruction. Until we identify a way of obtaining ideal patterns from a subset of our data, subjectively selected single hits are the next-best solution.

The particle size filtering step is quite important and has to be applied throughout the SPI analysis pipeline. A real experiment might run in the following way. A trained person

will select a number of single hits and non-single hits and then will run the CNN selection on the diffraction patterns coming from the experimental stream. After size filtering, this selection will be uploaded to the SPI workflow as shown in Fig. 7.1, and the electron density of a single particle will be obtained as a result.

Reconstructing the 3D structure from a selection of single hits is expensive: both computationally and in terms of manual labour. We introduced the PSD contrast in the hope that it would constitute a good substitute measure for the quality of a selection. If successful, this would have allowed us to optimize our CNNs more directly towards identifying an optimal set of single hits for reconstruction through maximizing their PSD contrast. Comparing the PSD contrast between CNN selections,  $D_M$  and  $D_{EM}$  (Chapter 6 and [38]) revealed that the contrast in the CNN and  $D_M$  selections is always lower than that in the  $D_{EM}$  selection. We initially thought that this may be problematic for the reconstructions. However, as the results in Fig. 7.7 demonstrate, this is not the case and our CNN selection (which mimics  $D_M$ ) is working well, resulting in an electron density of the PR772 virus that is similar to that obtained in our previous work [38]. These results indicate that the PSD contrast may not be a good substitute for reconstruction fidelity. Deviations from a circular shape, as are present in PR772, might explain this observation.

We have proposed an SPI workflow that uses a CNN-based single-hit classification at an early stage of the data analysis pipeline. This approach can be beneficial not only because it can be run during SPI experiments but also because it can significantly reduce the number of diffraction patterns for further processing. That is important for data storage, as the size of collected data sets during one experiment at a megahertz XFEL facility can easily reach several petabytes. Another convenience of using CNNs for single-hit classification is that the network can be trained on a relatively small quantity of data at the beginning of the SPI experiment and can be simply applied throughout the rest of the experiment.

Introducing non-standard AI-based solutions into an established SPI analysis workflow may be beneficial for the future development of SPI experiments. Here we have demonstrated the use of CNNs at the single-hit diffraction pattern classification step which can be applied not only after the experiment but, importantly, also during the experiment and can significantly reduce the size of data storage for further analysis stages. That could be an important advantage with the development of XFELs [185] with data collection at the megahertz rate [131, 138]. Handling experimental data with CNNs also saves computational time: once the CNN is trained and new data are obtained, there is no need to retrain the CNN again as is needed with other classification approaches.



## Chapter 8

# Simulation of SPI experiment with Tick-borne encephalitis virus

Understanding the structure and functionality of viruses has become an important task. Nowadays, we can observe it in the present pandemic of COVID-19. The society has realized that without the knowledge of the structure and functionality of viruses, it is difficult or even impossible to struggle with the SARS-CoV-2 virus.

To solve this problem, different methods may be applied. X-ray crystallography is the predominant method for determining the structure of biomolecules with high resolution. But since it is necessary to crystallize a protein or virus, its application is not always possible. More often this method is used for single viral proteins or compact, homogeneous, symmetric viral particles [109, 197, 198]. Small viral proteins, especially unstructured ones, are investigated by nuclear magnetic resonance (NMR) [199]. Some membrane proteins are difficult to crystallize due to their hydrophobicity and the laborious process of manufacturing the quantities of proteins required for crystallization [64]. However, knowledge of the structure of these proteins is essential for understanding the functioning of viruses.

One way to solve these problems is the method of studying the spatial structure of biological particles – Single Particle Imaging (SPI) – using cryogenic electron microscopy (Cryo-EM) [64–66] which was briefly discussed in Section 4.1. In most cases, this method works well for obtaining the structure of a virus capsid, but the internal structure is rather difficult to determine. In addition, when using Cryo-EM, the samples must be cooled to liquid nitrogen temperature which makes it difficult to understand the functionality of viruses in their native environment. These limitations can be circumvented using SPI approach based on the use of X-ray free-electron laser (XFEL).

Currently, XFELs are the most powerful X-ray sources and can produce strong X-ray radiation with the pulse length of several tens of femtoseconds [101, 102, 185, 200] described in Section 2.2. These sources have a high degree of coherence [201–203]. It is this impor-



tant property that makes it possible to use methods of Coherent X-ray Diffractive Imaging (CXDI) (Section 4.2, [68, 204]) in order to obtain the spatial structure of biological particles.

For a successful outcome, each SPI experiment needs careful planning. This Chapter is based on Ref. [205]. We will discuss the SPI experiment planned at European XFEL with the Tick-borne encephalitis virus (TBEV). Various experimental conditions will be discussed in details: incident photon flux incoming on the sample, sample-detector distance, and others. The study also presents a general data analysis pipeline and the use of the existing structure reconstruction platform [135]. Following a well-established data processing pipeline, the structure of TBEV was obtained and the efficiency of the used methods was demonstrated.

## 8.1 Tick-borne encephalitis virus

Tick-borne encephalitis is a viral infectious disease transmitted through tick bites. The endemic area extends from west to east from the Rhine to the Urals and from north to south from Scandinavia to Italy and Greece. Tick-borne encephalitis is usually asymptomatic, but can also cause serious complications, mainly in the form of the nervous system damage. The disease can result in disability or even death. There is no cure for tick-borne encephalitis, the main preventive measure is vaccination.

The pathogen of tick-borne encephalitis is a virus belonging to the family Flaviviridae, genus *Flavivirus*. In addition to tick-borne encephalitis, flaviviruses cause a number of serious human diseases, including long-known infections – yellow fever, Dengue fever, West Nile fever, Japanese encephalitis, as well as newly discovered and capable of rapid spread to new territories, such as Zika fever [206]. Several million cases of flavivirus infections are reported worldwide each year [206, 207].

All viruses of this family are enveloped viruses with a virion diameter of  $\sim 50$  nm. The virion core consists of a single-stranded (+)RNA molecule surrounded by protein C. It is covered on top by a lipid membrane, in which two proteins are embedded: the membrane protein M and the virion surface protein E, but protein M does not form the outer surface of the virion. Glycoprotein E is mainly responsible for the first stages of viral infection and is the target of most neutralizing antibodies [207]. The structure of these viruses accounts for their natural heterogeneity: mature, immature, semi-mature and so-called "broken", i.e. deformed particles that are formed in the samples during maturation [208, 209]. This makes it difficult to obtain the structures of flavivirus virions by X-ray crystallography, since heterogeneity prevents obtaining ordered crystals. In this regard, the method for obtaining flavivirus virion structures is Cryo-EM. Currently, the Protein Data Bank (PDB) [210] contains more than 40 structures of various flaviviruses.

The Cryo-EM method requires careful sample preparation [209] and the maintenance of a fairly high concentration of homogeneous particles of the same type, usually mature

## 8.2. Data simulations for the SPI experiment

---

or immature virions which are the most symmetrical. Viral particles with antigen-binding fragments (Fab-fragments) that neutralize antibodies [211, 212] are also studied. Two TBEV structures were obtained by the Cryo-EM method with 3.9 Å resolution [208]: the structure of the mature virion complex (structure code in PDB is 5O6A [213]) shown in Fig. 8.1 (a) and the structure of the mature virion complex with the Fab-fragment of the mouse monoclonal antibody 19/1786 (structure code in PDB is 5O6V [214]) shown in Fig. 8.1 (b).

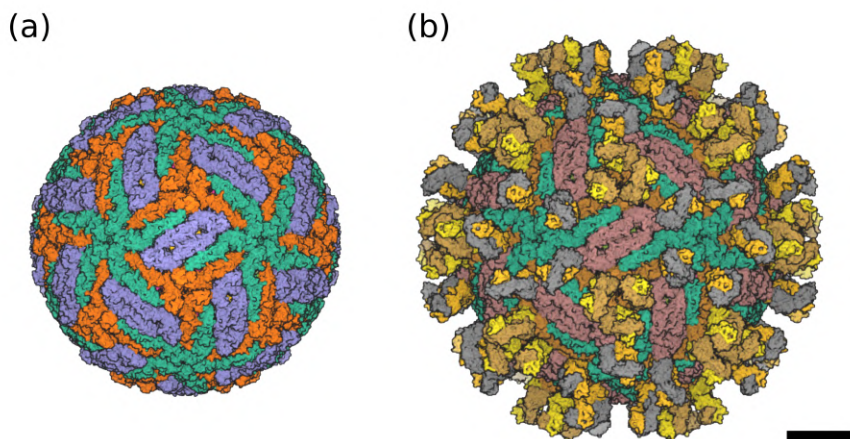


Figure 8.1: Cryo-EM structure of TBEV. (a) The structure of the mature TBEV particle 5O6A. (b) Structure of the complex with Fab-fragment of neutralizing monoclonal antibody 5O6V. The structures of TBEV were taken from the Protein Data Bank [213, 214]. The size of the scale bar is 10 nm.

To model the diffraction data, we used both structures of the virus.

## 8.2 Data simulations for the SPI experiment

One of the goals of data simulation for TVEB was to plan the SPI experiment at the European XFEL at the Single Particles, Clusters, and Biomolecules (SPB) beamline. Reconstructing the spatial structure of a TBEV on the basis of 2D diffraction patterns is the main task of SPI analysis.

Typical experimental set-up is known and shown in Fig. 5.2. The success of complex experiments such as SPI depends on many parameters. Parameters that can be evaluated in advance by simulation include the incident photon flux and the sample-detector distance. The scattered signal, clearly, depends on the intensity of the incident X-ray beam. At the European XFEL its intensity is 1 – 4 mJ per pulse which corresponds to  $10^{11}$  –  $10^{12}$  photons per pulse. Since in the planned experiment the size of the focal spot of the X-ray beam is 300 nm, it is natural to assume that there will be no more than  $10^{12}$  photons per pulse in the focal spot.

In this experiment the planned photon energy is 6 keV (wavelength 2.07 Å). On the one hand, this energy is the lowest possible energy at the SPB station. On the other hand, the

lower the energy of the incident photons, the stronger the scattered radiation. The European XFEL SPB beamline uses an AGIPD 1 Mpx detector [127] with a size of  $1024 \times 1024$  pixels (one pixel size is  $200 \times 200 \mu\text{m}^2$ ). The parameters of the SPB beamline described above were used in the simulation of diffraction patterns using the MOLTRANS program developed at DESY.

First of all, it is of interest to compare the diffraction patterns of the two available virus types: 5O6A and 5O6V. The global symmetry of both is icosahedral, but the 5O6A structure has a pronounced spherical shape, and the diffraction pattern from such an object consists of concentric rings. The structure of 5O6V looks different than that of 5O6A in reciprocal space. Due to the antigen-binding fragments, sixth-order symmetry can be seen in the diffraction patterns, indicating the appearance of characteristic features of the structure. Examples of diffraction patterns from two structures in random orientations are shown in Fig. 8.2.

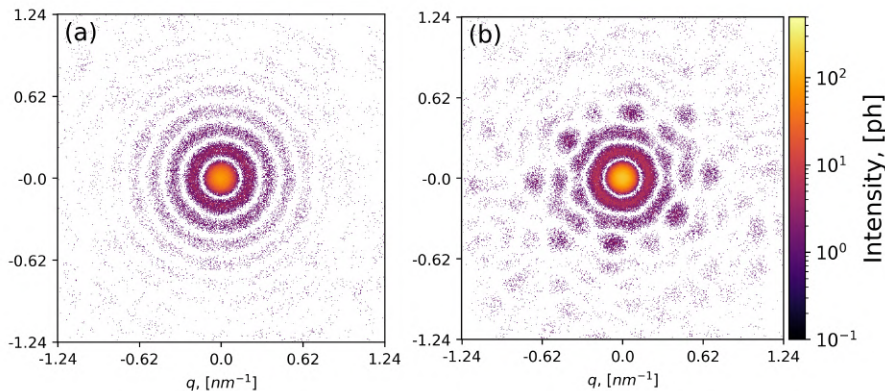


Figure 8.2: Diffraction patterns from single objects in random orientation: (a) 5O6A structure (for TBEV in Fig. 8.1 (a)); (b) 5O6V structure (for TBEV with Fab-fragments in Fig. 8.1 (b)). Simulations parameters: wavelength  $2.07 \text{ \AA}$ , X-ray beam focus  $300 \text{ nm}$ , detector  $512 \times 512$  pixels, pixel size  $400 \times 400 \mu\text{m}^2$ , distance  $2.5 \text{ m}$ , signal intensity  $10^{12}$  photons in focus.

To demonstrate other parameters, we used the 5O6V structure (Fig. 8.1 (b)). The scattered signal recorded by the detector depends on the parameters of the experimental setup as well as on the sample – the larger the object, the higher the intensity of the scattered signal. Note that the Cryo-EM structures taken from the PDB bank and used in the present work describe the surface protein E and membrane protein M and do not characterize the inner nucleocapsid formed by protein C and the RNA structure chain. Naturally, in the SPI experiment, the presence of the RNA nucleocapsid in the particles will contribute to the scattered signal on the detector.

Fig. 8.3 shows three diffraction patterns from a single virus with different photon flux in the focus of the X-ray beam. The figure shows that the photon flux from the virus must reach the necessary level for successful analysis of the SPI experiment. If the scattering signal from the virus is too weak (Fig. 8.3 (a)), the background from the experimental setup will be dominant, making further analysis difficult or impossible. If the intensity of the incident

## 8.2. Data simulations for the SPI experiment

radiation is  $10^{12}$  photons at the focus of the X-ray beam (see Fig. 8.3 (c)), the scattered signal is well distinguishable. From the simulations (see Fig. 8.3) we can conclude that the maximum signal intensity of  $10^{11} - 10^{12}$  photons at the X-ray beam focus which is achievable at the European XFEL, is necessary for the experiment to succeed.

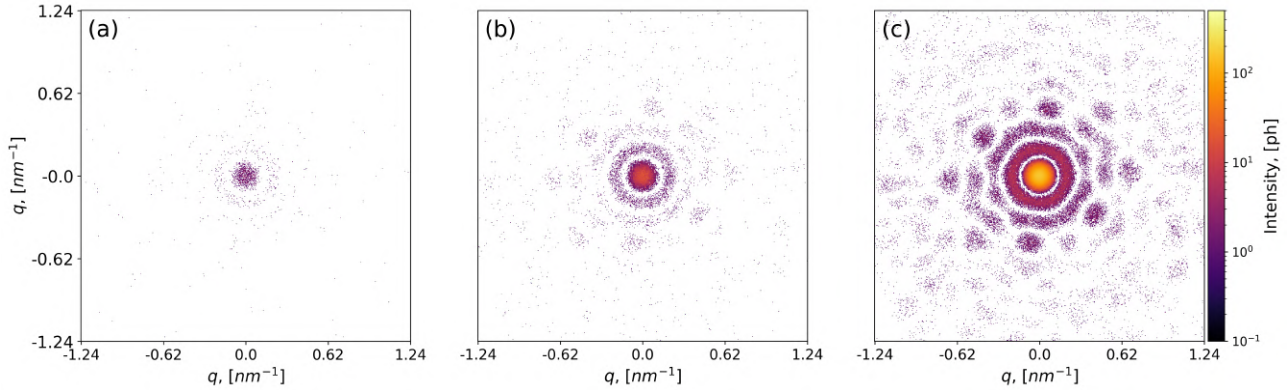


Figure 8.3: Diffraction patterns from a single TBEV in random orientation. Signal intensities: (a)  $10^{10}$ , (b)  $10^{11}$ , (c)  $10^{12}$  photons in focus. The sample-detector distance is 2.5 m.

Another important parameter of the SPI experiment that can be analyzed with the simulations is the sample-detector distance. If the distance is too small, it will not allow to obtain a pronounced diffraction from the sample, but it will allow to obtain a high resolution. If the distance is too large, it will not allow to achieve the desired resolution. Examples of diffraction patterns with different distances from 1 m to 3 m are shown in Fig. 8.4 (a)-(c). An angle-averaged intensity was plotted for each case and is shown in Fig. 8.4 (g)-(i). As it was expected, at the shortest distance of 1 m, the diffraction pattern shows the characteristic features of the virus structure, and the resolution in the real space reaches 2 nm. As the distance increases to 2 m, these features become more pronounced, but the resolution in real space drops to 4 nm. At the maximum distance of 3 m, the diffraction pattern from the virus is clearly distinguishable; in Fig. 8.4 (i) the characteristic rings are clearly visible. But the resolution in reciprocal space is limited to 6.3 nm.

Another important factor to consider when choosing the optimal sample-detector distance in an SPI experiment is the structure of the detector panels. In practice, the detector panels are not placed together; there is a distance between them. There is also a gap in the center of the detector for the direct (central) beam to pass through.

Experimental diffraction patterns with the superimposed geometry of gaps between the detector panels are shown in Fig. 8.4 (d)-(f). The detector geometry was taken from one of the SPI experiments at the SPB beamline of European XFEL. The figure shows that a large part of the central diffraction peak at a distance of 1 m is not determined because of the panel positions in the central part of the detector. Information about the size of the central peak is essential when reconstructing the object. It is also important to take this into account when planning the experiment, in particular, when choosing the optimal distance. From the

analysis of the performed simulations (see Fig. 8.4) with different sample-detector distances, we can conclude that a distance of 2 – 3 m is preferable. In this case, the detector geometry makes it possible to distinguish all structural features of the virus, and the momentum transfer vector  $q$  reaches a value of  $1.04 - 1.55 \text{ nm}^{-1}$  in reciprocal space which corresponds to a resolution of 4 – 6 nm in real space.

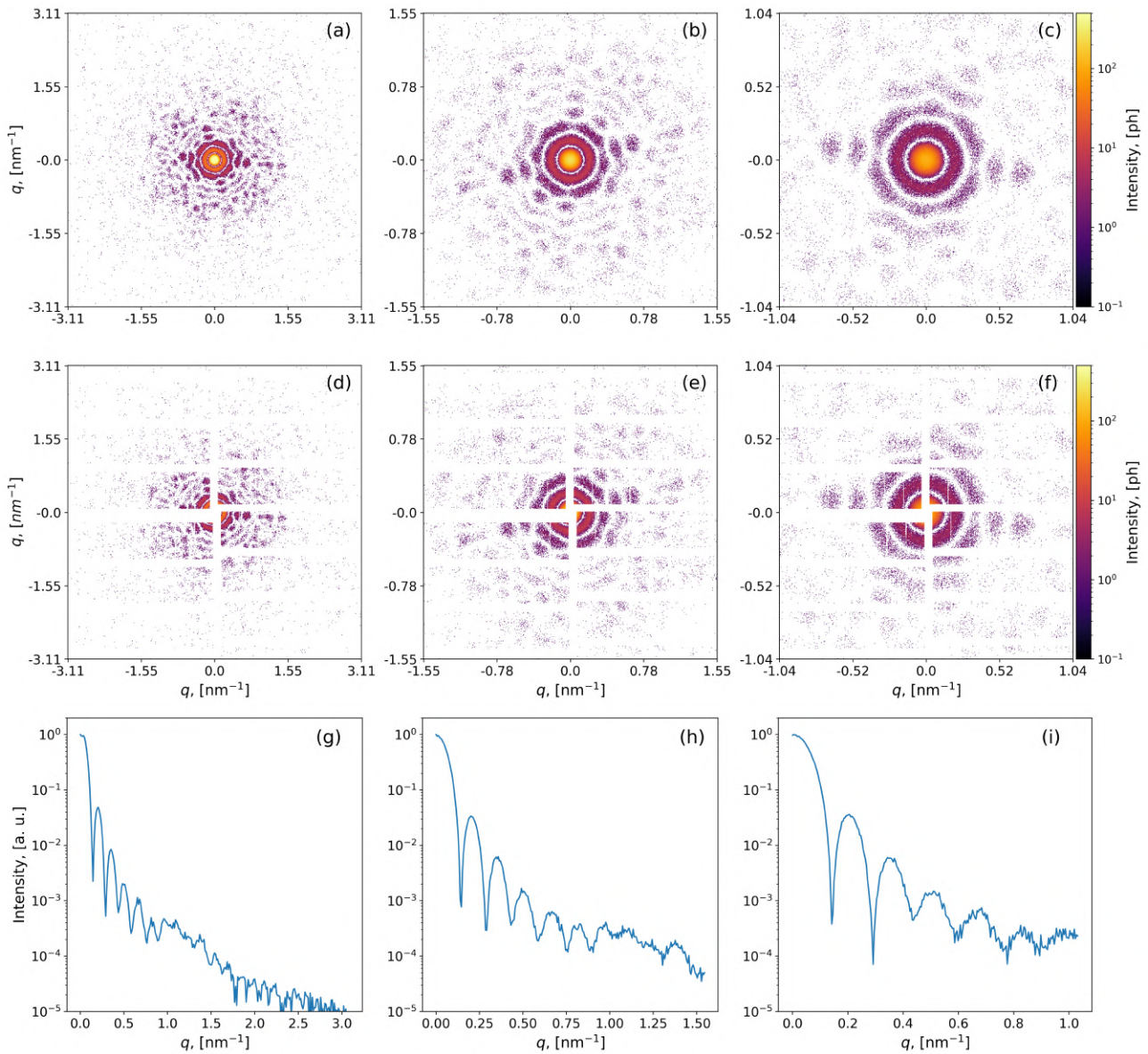


Figure 8.4: Diffraction patterns from a single TBEV in random orientation. At distances of 1 (a, d, g), 2 (b, e, h) and 3 (c, f, i) m. Examples of diffraction patterns with the detector mask superimposed on them (d)-(f). Functions (g)-(i) corresponding to the angular averaged intensity in (a)-(c).

### 8.3 Spatial structure of the TBEV from simulated data

In this work, we adapted the mentioned above data analysis pipeline (Section 5.4) for simulated diffraction patterns from TBEV with and without Fab-fragments. A 1,000 diffraction patterns in random orientations were created for each TBEV type (see Fig. 8.1) with the following parameters: wavelength 2.07 Å, focal beam size 300 nm, detector size  $128 \times 128$  pixels, pixel size  $1.6 \times 1.6 \text{ nm}^2$ . Such dimensions were set in the simulations in order to match the real size of the AGIPD 1 Mpx detector but also to save computational time. Other parameters are the sample-detector distance 2.1 m, signal intensity  $10^{11}$  photons in focus.

As only the structure of the TBEV obtained by Cryo-EM [208] was used in the simulation, clustering and classification of the diffraction patterns by object type was not required. Since the orientation of the particle in the simulations is known, combining the data into a diffraction 3D intensity volume in reciprocal space was done according to the identified orientations of the virus. The result of the simulations in reciprocal space is shown in Fig. 8.5. For virus 5O6A (see Fig. 8.5 (a)), as expected, concentric rings are observed. Due to the presence of Fab-fragments in the structure of virus 5O6V, diffraction fringes in reciprocal space are observed (see Fig. 8.5 (b)).

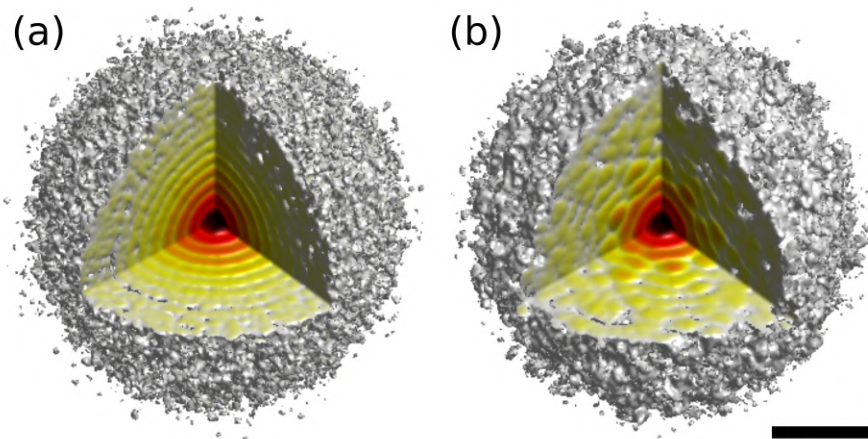


Figure 8.5: Diffraction 3D intensity volume in reciprocal space: (a) 5O6A structure (for TBEV in Fig. 8.1 (a)); (b) 5O6V structure (for TBEV with Fab-fragments, Fig. 8.1 (b)). The size of the scale bar is  $1 \text{ nm}^{-1}$ .

Since the simulations do not contain the experimental background signal, the step of its correction was not necessary. The next step is to reconstruct the scattering phases and structure of the object. As described earlier, iterative phase retrieval algorithms are used for this task (Section 4.3). These algorithms are based on the Fourier transform between real and reciprocal spaces using two constraints: in reciprocal space, the signal amplitude is set equal to the experimentally measured values, and in real space, the object occupies a limited volume whose approximate size is known in advance.

To obtain the spatial structure of the virus (for 5O6A and 5O6V) the following combination of algorithms was used: 100 iterations of Continuous Hybrid-Input-Output (CHIO), followed by 200 iterations of Error Reduction (ER) with an alternating Shrink-Wrap (SW) algorithm every 10 iterations with a threshold value of 0.2. This combination of algorithms was repeated 4 times for one reconstruction with the total number of iterations 1,200. A total of 30 reconstructions were made. Then they were averaged using the mode decomposition method described in [38] and Section 6.5.2. The main mode was further considered as the final spatial structure of the TBEV, shown in Fig. 8.6.

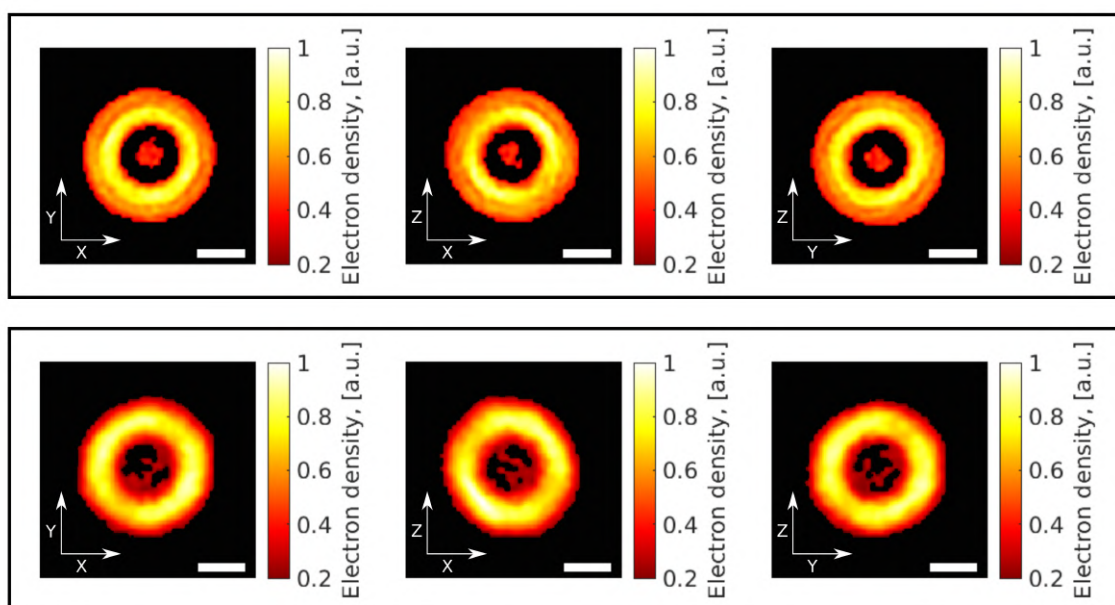


Figure 8.6: Central slices of the TBEV reconstructions. Upper row for structure 5O6A, lower row for structure 5O6V. Electron density values are normalized to the maximum, values less than 0.2 are shown in black. The size of the scale bar is 20 nm.

The spatial structure of 5O6A and 5O6V has a ring shape with reduced density inside the particle, as in the data used for simulation of diffraction patterns. From the obtained result, we can see that due to the low resolution Fab-fragments in the structure of 5O6V cannot be distinguished. Thus, the size of both structures is about 60 nm. This corresponds to the size of the 5O6V structure ( $\sim 57$  nm) obtained with Cryo-EM [208]. At the same time, the size of the virus obtained by the reconstruction is larger than the size of the 5O6A ( $\sim 47$  nm [208]). Note, that the structures 5O6A and 5O6V used for simulations did not contain electron density inside (internal RNA).

## 8.4 Summary

An analysis of diffraction patterns simulation for the SPI experiment with TBEV is presented. SPI method allows obtaining the spatial structure of biological nanoparticles using

intense XFEL femtosecond pulses. Such experiments require careful preparation and planning. To prepare the experimental set-up and efficiently use the beamtime, some parameters can be estimated in advance, for example, by means of diffraction patterns simulation.

A TBEV was chosen as a studied object, for which the structure of the outer envelope was known from Cryo-EM [208]. The size and relative homogeneity of TBEV make it a good object of study in SPI experiments on XFEL.

Two TBEV structures were used to simulate diffraction patterns: a mature virion complex (structure 5O6A in PDB) and a mature virion with a Fab-fragment (structure 5O6V in PDB). These two structures give different diffraction patterns in reciprocal space. In the case of 5O6A, only concentric rings were distinguishable in diffraction patterns. Whereas for 5O6V one can observe characteristic features of the structure associated with "spikes" of Fab-fragments on the surface of the viral particle.

In order to prepare the SPI experiment at the European XFEL, the following parameters were varied during diffraction patterns simulations: the X-ray beam intensity at the focus and the sample-detector distance. With the help of simulations, it was possible to determine the optimal parameters and use them in the preparation of the experiment. For the SPB beamline of European XFEL the following parameters were identified as optimal: signal intensity  $10^{11} - 10^{12}$  photons in the focus, sample-detector distance 2-3 m.

Only necessary steps of SPI data analysis pipeline were used for simulated data: merging the data into a diffraction 3D volume; scattering phases reconstruction, and object structure reconstruction. Using iterative phase retrieval algorithms, 30 virus reconstructions were obtained, they were averaged by mode decomposition, and the final structure for 5O6A and 5O6V was chosen as the main mode of decomposition. For the 5O6V structure it was impossible to distinguish Fab-fragments, for both structures (5O6A and 5O6V) a ring corresponding to the virus membrane was present with reduced density inside which corresponded to the original structure used for simulations.

The presented study shows the efficiency of analysis methods for SPI experiments. Megahertz facilities, such as the European XFEL [185, 186], allow to obtain the necessary amount of data in a shorter time. In recent SPI experiments at the European XFEL with gold nanoparticles, a resolution of 2 nm was achieved [215]. Thus, the resolution in future SPI experiments with viruses of similar sizes can be expected to be between 2 nm and 10 nm.





# Chapter 9

## Summary

This Thesis describes the application of coherent X-ray diffraction techniques to the structural investigation of biological particles, particularly, single particle imaging experiments at X-ray free-electron lasers. The results of three projects were presented.

The first study is based on the SPI experiment performed at the AMO instrument at the LCLS. PR772 bacteriophage was studied in order to obtain a three-dimensional structure from the two-dimensional diffraction patterns collected during the beamtime. The typical data analysis pipeline was adjusted to the challenges of the experiment, such as a not operational panel of the detector and a low percentage of usable data. Modified workflow included necessary preprocessing steps: instrumental background subtraction, beam center estimation, and advanced particle size determination. After data was prepared, the next step followed – single hit classification, performed using an expectation-maximization algorithm. It distributed the data into a predefined number of classes according to its features. This machine learning technique is widely used in data clustering with the presence of latent features. With its implementation, the number of final single hit diffraction patterns was reduced from millions to 1,085. Mode decomposition was used at the stage of combining phase retrieval reconstructions of the virus, keeping the most important features of the structure. The final three-dimensional electron density of PR772 was obtained with the 6.9 nanometer resolution which was better than previous studies reported. The outer and inner structures of the final reconstruction also corresponded well to the results of cryogenic microscopy. These results showed the potential of the SPI general workflow modification which provides a higher resolution of the final electron density of the biological particle.

The second study was formulated as a logical follow-up to the first research topic. Using the data from the same SPI experiment, we tried to broaden the utilization of modern machine learning algorithms by applying a convolutional neural network in the SPI data analysis pipeline. It was implemented to perform the binary classification of the diffraction patterns recorded at XFEL: single hit and non-single hit. Two models of CNN were developed. The first one was optimized for maximizing the F1 score which combines precision

and recall (standard metrics in performance estimation) as their harmonic mean. The second model optimized the recall value. As a training set, the manual selection of single hits was used. It was done in order to reproduce the possible CNN application in the online regime during the SPI experiment. Thus, the training set has to be representative and balanced, and forming such a training set still has room for improvement. Another benefit of CNN is the ability to reduce data size for further processing which can potentially save space, time, and money. After particle size filtering and EM clustering, the final number of diffraction patterns in two CNN models were 829 and 1,090 single hits. Both data sets led to reasonable electron density reconstructions of PR772 in agreement with the previous studies showing a similar resolution. The SPI experiments and their data analysis can highly benefit from the successful realization of CNN solutions and potentially lead to the ultimate goal of SPI – atomic resolution and observing chemical processes in biological samples.

In the third part, simulations of SPI experiments with tick-borne encephalitis virus (TBEV) at the European XFEL were demonstrated. Being a challenging experiment, SPI requires a high level of preparation. The way to conduct a successful experiment and spend expensive time of the experiment efficiently is to use simulations in advance. Two models of TBEV were considered as the base: a virion and a virion with a Fab-fragment. As a result of diffraction data simulations, the preferred experimental conditions were determined. One thousand diffraction patterns were simulated as the TBEV structures hit the X-ray beam in random orientations. They were combined into one three-dimensional reciprocal space volume, and the electron density for both virus particles was reconstructed. The conformity between obtained results and used models of TBEV, such as empty internal volume due to the RNA absence, served as a reliable indicator that experimental parameters are optimal and can be used during the real beamtime at the European XFEL.

In summary, this Thesis gives an overview of the theoretical and experimental basis of the SPI experiment and emphasizes its capabilities and future potential at different XFELs. Automation of data processing and analysis is a promising direction for further studies and can contribute to the development of X-ray technologies at XFELs. Furthermore, the current epidemiological situation in the world sets new challenges for scientific society in the field of structural studies of biological particles. In the face of these new demands, the megahertz repetition rate XFELs and their ability to record millions of diffraction patterns by high-dynamic-range novel detectors can bring new insights into understanding the morphology and functions of biological particles, particularly viruses. This can be done in conjunction with methodological development of the data analysis algorithms, including machine learning techniques which have proved to be state-of-art approaches in everyday life.

# Publications

## Scientific publications related to this thesis

1. **Assalauova, D.**, Kim, Y. Y., Bobkov, S., Khubbutdinov, R., Rose, M., Alvarez, R., Andreasson, J., Balaur, E., Contreras, A., DeMirci, H., Gelisio, L., Hajdu, J., Hunter, M. S., Kurta, R. P., Li, H., McFadden, M., Nazari, R., Schwander, P., Teslyuk, A., Walter, P., Xavier, P. L., Yoon, C. H., Zaare, S., Ilyin, V. A., Kirian, R. A., Hogue, B. G., Aquila, A. and Vartanyants, I. A. An advanced workflow for single-particle imaging with the limited data at an X-ray free-electron laser. *IUCrJ* 7, 1102-1113 (2020).
2. **Assalauova, D.**, Ignatenko, A., Isensee, F., Trofimova, D. and Vartanyants I. A. Classification of diffraction patterns using a convolutional neural network in single-particle-imaging experiments performed at X-ray free-electron lasers. *J. Appl. Crystallogr.* 55, 444-454 (2022).
3. **Assalauova, D.** and Vartanyants, I. A. The structure of Tick-borne Encephalitis virus determined at X-Ray Free-Electron Lasers. Simulations. *J. Synchrotron Radiat.* (accepted) (2022).
4. Ignatenko, A., **Assalauova, D.**, Bobkov, S. A., Gelisio, L., Teslyuk, A. B., Ilyin, V. A., and Vartanyants, I. A. Classification of diffraction patterns in single particle imaging experiments performed at x-ray free-electron lasers using a convolutional neural network. *Mach. Learn.: Sci. Technol.* 2(2), 025014 (2021).
5. Bobkov, S. A., Teslyuk, A. B., Baymukhametov, T. N., Pichkur, E. B., Chesnokov, Y. M., **Assalauova, D.**, Poyda, A. A., Novikov, A. M., Zolotarev, S. I., Ikonnikova, K. A., Velikhov, V. E., Vartanyants, I. A., Vasiliev, A. L. and Ilyin, V. A. Advances in Modern Information Technologies for Data Analysis in CRYO-EM and XFEL Experiments. *Crystallogr. Rep.* 65(6), 1081-1092 (2020).

## Scientific publications not directly related to this thesis

1. Maier, A., Lapkin, D., Mukharamova, N., Frech, P., **Assalauova, D.**, Ignatenko, A., Khubbutdinov, R., Lazarev, S., Sprung, M., Laible, F., Löffler, R., Previdi, N., Bräuer, A., Güntel, T., Fleischer, M., Schreiber, F., Vartanyants, I. A. and Scheele, M. Structure–Transport Correlation Reveals Anisotropic Charge Transport in Coupled PbS Nanocrystal Superlattices. *Adv. Mater.* 32(36), 2002254 (2020).
2. Schlotheuber né Brunner, J., Maier, B., Thomä, S. L., Kirner, F., Baburin, I. A., Lapkin, D., Rosenberg, R., Sturm, S., **Assalauova, D.**, Carnis, J., Kim, Y. Y., Ren, Z., Westermeyer, F., Theiss, S., Borrmann, H., Polarz, S., Eychmüller, A., Lubk, A., Vartanyants, I. A., Cölfen, H., Zobel, M., and Sturm, E. V. Morphogenesis of Magnetite Mesocrystals: Interplay between Nanoparticle Morphology and Solvation Shell. *Chem. Mater.* 33(23), 9119-9130 (2021).
3. Khubbutdinov, R., Gerasimova, N., Mercurio, G., **Assalauova, D.**, Carnis, J., Gelisio, L., Le Guyader, L., Ignatenko, A., Kim, Y. Y., Van Kuiken, B. E., Kurta, R. P., Lapkin, D., Teichmann, M., Yaroslavtsev, A., Gorobtsov, O., Menushenkov, A. P., Scholz, M., Scherz, A. and Vartanyants, I. A. High spatial coherence and short pulse duration revealed by the Hanbury Brown and Twiss interferometry at the European XFEL. *Struct. Dyn.* 8(4), 044305 (2021).
4. Lapkin, D., Mukharamova, N., **Assalauova, D.**, Dubinina, S., Stellhorn, J., Westermeyer, F., Lazarev, S., Sprung, M., Karg, M., Vartanyants, I. A. and Meijer, J. M. In situ characterization of crystallization and melting of soft, thermoresponsive microgels by small-angle X-ray scattering. *Soft Matter* 18, 2591-1602 (2022).
5. Lapkin, D., Kirsch, C., Hiller, J., Andrienko, D., **Assalauova, D.**, Braun, K., Carnis, J., Kim, Y. Y., Mandal, M., Maier, A., Meixner, A. J., Mukharamova, N., Scheele, M., Schreiber, F., Sprung, M., Wahl, J., Westendorf, S., Zaluzhnyy, I. A. and Vartanyants, I. A. Spatially resolved fluorescence of caesium lead halide perovskite supercrystals reveals quasi-atomic behavior of nanocrystals. *Nat. Commun.* (2022).
6. Wahl, J., Haizmann, P., Kirsch, C., Frecot, R., Mukharamova, N., **Assalauova, D.**, Kim, Y. Y., Zaluzhnyy, I., Chassé, T., Vartanyants, I. A., Peisert, H. and Scheele, M. Mitigating the photodegradation of all-inorganic mixed-halide perovskite nanocrystals by ligand exchange. *Phys. Chem. Chem. Phys.* 18. (2022).
7. Ekeberg, T., **Assalauova, D.**, Bielecki, J., Boll, R., Daurer, B. J., Eichacker, L.A., Franken, L.E., Galli, D.E., Gelisio, L., Gumprecht, L., Gunn, L. H., Hajdu, J., Hartmann, R., Hasse, D., Ignatenko, A., Koliyadu, J., Kulyk, O., Kurta, R., Kuster, M., Lugmayr, W.,

## PUBLICATIONS

---

Lübke, J., Mancuso, A. P., Mazza, T., Nettelblad, C., Ovcharenko, Y., Rivas, D. E., Rose, M., Samanta, A. K., Schmidt, P., Sobolev, E., Timneanu, N., Usenko, S., Westphal, D., Wollweber, T., Worbs, L., Xavier, P. L., Yousef, H., Ayyer, K., Chapman, H. N., Sellberg, J. A., Seuring, C., Vartanyants, I. A., Küpper, J., Meyer, M., Maia, F.R.N.C. Observation of a single protein by ultrafast X-ray diffraction. *Nat. Photon.* (submitted) (2022).



# Bibliography

- <sup>1</sup>W. C. Röntgen, “On a new kind of ray, a preliminary communication”, Wurzburg Physico-Médical Society on December **28** (1895).
- <sup>2</sup>W. Friedrich, P. Knipping, and M. Laue, “Interferenzerscheinungen bei Röntgenstrahlen”, *Annalen der Physik* **346**, 971–988 (1913).
- <sup>3</sup>W. H. Bragg and W. L. Bragg, “The reflection of X-rays by crystals”, *Proc. R. Soc. Lond. A Math. Phys.* **88**, 428–438 (1913).
- <sup>4</sup>J. Als-Nielsen and D. McMorrow, *Elements of modern x-ray physics* (John Wiley & Sons, 2011).
- <sup>5</sup>J. Stöhr, “Two-photon X-ray diffraction”, *Phys. Rev. Lett.* **118**, 024801 (2017).
- <sup>6</sup>D. Attwood, *Soft x-rays and extreme ultraviolet radiation: principles and applications* (Cambridge university press, 2000).
- <sup>7</sup>*PETRA III Beamlines*, [https://photon-science.desy.de/facilities/petra\\_iii/beamlines/index\\_eng.html](https://photon-science.desy.de/facilities/petra_iii/beamlines/index_eng.html).
- <sup>8</sup>M. Eriksson, J. F. Van der Veen, and C. Quitmann, “Diffraction-limited storage rings—a window to the science of tomorrow”, *J. Synchrotron Radiat.* **21**, 837–842 (2014).
- <sup>9</sup>C. G. Schroer, I. Agapov, W. Brefeld, R. Brinkmann, Y.-C. Chae, H.-C. Chao, M. Eriksson, J. Keil, X. Nuel Gavalda, R. Röhlberger, et al., “PETRA IV: the ultralow-emittance source project at DESY”, *J. Synchrotron Radiat.* **25**, 1277–1290 (2018).
- <sup>10</sup>L. Mandel and E. Wolf, *Optical coherence and quantum optics* (Cambridge university press, 1995).
- <sup>11</sup>K.-J. Kim, “Characteristics of synchrotron radiation”, in *AIP conference proceedings*, Vol. 184, 1 (1989), pp. 565–632.
- <sup>12</sup>G. Geloni, S. Serkez, R. Khubbutdinov, V. Kocharyan, and E. Saldin, “Effects of energy spread on brightness and coherence of undulator sources”, *J. Synchrotron Radiat.* **25**, 1335–1345 (2018).
- <sup>13</sup>P. F. Tavares, S. C. Leemann, M. Sjöström, and Å. Andersson, “The MAX IV storage ring project”, *J. Synchrotron Radiat.* **21**, 862–877 (2014).



- <sup>14</sup>P. Raimondi et al., “Hybrid multi bend achromat: from SuperB to EBS”, Proc. IPAC’17, 3670–3675 (2017).
- <sup>15</sup>R. Khubbutdinov, A. Menushenkov, and I. Vartanyants, “Coherence properties of the high-energy fourth-generation x-ray synchrotron sources”, J. Synchrotron Radiat. **26**, 1851–1862 (2019).
- <sup>16</sup>T. E. Fornek, *Advanced Photon Source upgrade project final design report*, tech. rep. (Argonne National Lab.(ANL), Argonne, IL (United States), 2019).
- <sup>17</sup>E. Weckert, “The potential of future light sources to explore the structure and function of matter”, IUCrJ **2**, 230–245 (2015).
- <sup>18</sup>J. M. Madey, “Stimulated emission of bremsstrahlung in a periodic magnetic field”, J. Appl. Phys. **42**, 1906–1913 (1971).
- <sup>19</sup>D. A. Deacon, L. Elias, J. M. Madey, G. Ramian, H. Schwettman, and T. I. Smith, “First operation of a free-electron laser”, Phys. Rev. Lett. **38**, 892 (1977).
- <sup>20</sup>A. Kondratenko and E. Saldin, “Generating of coherent radiation by a relativistic electron beam in an undulator”, Part. Accel. **10**, 207–216 (1980).
- <sup>21</sup>R. Bonifacio, C. Pellegrini, and L. Narducci, “Collective instabilities and high-gain regime free electron laser”, in AIP conference proceedings, Vol. 118, 1 (1984), pp. 236–259.
- <sup>22</sup>J. Murphy and C Pellegrini, “Generation of high-intensity coherent radiation in the soft-x-ray and vacuum-ultraviolet region”, JOSA B **2**, 259–264 (1985).
- <sup>23</sup>R Abela, A Aghababayan, M Altarelli, C Altucci, G Amatuni, P Anfinrud, P Audebert, V Ayvazyan, N Baboi, J Baehr, et al., *XFEL: the European X-ray free-electron laser-technical design report*, tech. rep. (DESY, 2006).
- <sup>24</sup>*European XFEL beamlines*, [https://www.xfel.eu/facility/beamlines/index\\_eng.html](https://www.xfel.eu/facility/beamlines/index_eng.html).
- <sup>25</sup>D Garzella, T Hara, B Carre, P Salieres, T Shintake, H Kitamura, and M. Couprie, “Using VUV high-order harmonics generated in gas as a seed for single pass FEL”, Nucl. Instrum. Methods Phys. Res. A: Accel. Spectrom. Detect. Assoc. Equip. **528**, 502–505 (2004).
- <sup>26</sup>L.-H. Yu, M Babzien, I Ben-Zvi, L. DiMauro, A Doyuran, W Graves, E Johnson, S Krinsky, R Malone, I Pogorelsky, et al., “High-gain harmonic-generation free-electron laser”, Science **289**, 932–934 (2000).
- <sup>27</sup>E Allaria, D Castronovo, P Cinquegrana, P Craievich, M. Dal Forno, M. Danailov, G D’Auria, A Demidovich, G De Ninno, S Di Mitri, et al., “Two-stage seeded soft-x-ray free-electron laser”, Nat. Photonics **7**, 913–918 (2013).

- <sup>28</sup>J Feldhaus, E. Saldin, J. Schneider, E. Schneidmiller, and M. Yurkov, "Possible application of x-ray optical elements for reducing the spectral bandwidth of an x-ray sase FEL", *Opt. Commun.* **140**, 341–352 (1997).
- <sup>29</sup>G. Geloni, V. Kocharyan, and E. Saldin, "A novel self-seeding scheme for hard x-ray FELs", *J. Mod. Opt.* **58**, 1391–1403 (2011).
- <sup>30</sup>J Amann, W Berg, V Blank, F.-J. Decker, Y Ding, P Emma, Y Feng, J Frisch, D Fritz, J Hastings, et al., "Demonstration of self-seeding in a hard-x-ray free-electron laser", *Nat. Photonics* **6**, 693–698 (2012).
- <sup>31</sup>M. Yabashi and T. Tanaka, "Self-seeded FEL emits hard X-rays", *Nat. Photonics* **6**, 648–649 (2012).
- <sup>32</sup>D. Ratner, R Abela, J Amann, C Behrens, D Bohler, G Bouchard, C Bostedt, M Boyes, K Chow, D Cocco, et al., "Experimental demonstration of a soft x-ray self-seeded free-electron laser", *Phys. Rev. Lett.* **114**, 054801 (2015).
- <sup>33</sup>R. Neutze, R. Wouts, D. Van der Spoel, E. Weckert, and J. Hajdu, "Potential for biomolecular imaging with femtosecond x-ray pulses", *Nature* **406**, 752–757 (2000).
- <sup>34</sup>H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, et al., "Femtosecond x-ray protein nanocrystallography", *Nature* **470**, 73–77 (2011).
- <sup>35</sup>M. Rose, S. Bobkov, K. Ayyer, R. P. Kurta, D. Dzhigaev, Y. Y. Kim, A. J. Morgan, C. H. Yoon, D. Westphal, J. Bielecki, et al., "Single-particle imaging without symmetry constraints at an X-ray free-electron laser", *IUCrJ* **5**, 727–736 (2018).
- <sup>36</sup>R. P. Kurta, J. J. Donatelli, C. H. Yoon, P. Berntsen, J. Bielecki, B. J. Daurer, H. DeMirici, P. Fromme, M. F. Hantke, F. R. Maia, et al., "Correlations in scattered x-ray laser pulses reveal nanoscale structural features of viruses", *Phys. Rev. Lett.* **119**, 158102 (2017).
- <sup>37</sup>T. Ekeberg, M. Svenda, C. Abergel, F. R. Maia, V. Seltzer, J.-M. Claverie, M. Hantke, O. Jönsson, C. Nettelblad, G. Van Der Schot, et al., "Three-dimensional reconstruction of the giant mimivirus particle with an x-ray free-electron laser", *Phys. Rev. Lett.* **114**, 098102 (2015).
- <sup>38</sup>D. Assalauova, Y. Y. Kim, S. Bobkov, R. Khubbutdinov, M. Rose, R. Alvarez, J. Andreasson, E. Balaur, A. Contreras, H. DeMirici, et al., "An advanced workflow for single-particle imaging with the limited data at an x-ray free-electron laser", *IUCrJ* **7**, 1102–1113 (2020).
- <sup>39</sup>P. Vester, I. A. Zaluzhnyy, R. P. Kurta, K. B. Møller, E. Biasin, K. Haldrup, M. M. Nielsen, and I. A. Vartanyants, "Ultrafast structural dynamics of photo-reactions observed by time-resolved x-ray cross-correlation analysis", *Struct. Dyn.* **6**, 024301 (2019).

- <sup>40</sup>M. Minitti, J. Budarz, A. Kirrander, J. Robinson, D. Ratner, T. Lane, D Zhu, J. Glowia, M Kozina, H. Lemke, et al., "Imaging molecular motion: femtosecond x-ray scattering of an electrocyclic chemical reaction", *Phys. Rev. Lett.* **114**, 255501 (2015).
- <sup>41</sup>L. Landau and E. Lifchitz, *The Classical Theory of Fields* (Pergamon Press, 1971).
- <sup>42</sup>A. H. Compton, "A quantum theory of the scattering of X-rays by light elements", *Phys. Rev.* **21**, 483 (1923).
- <sup>43</sup>O. Klein and Y. Nishina, "The scattering of light by free electrons according to Dirac's new relativistic dynamics", *Nature* **122**, 398–399 (1928).
- <sup>44</sup>R. N. Bracewell and R. N. Bracewell, *The fourier transform and its applications* (McGraw-Hill New York, 1986).
- <sup>45</sup>L. Feigin, D. I. Svergun, et al., *Structure analysis by small-angle x-ray and neutron scattering* (Springer, 1987).
- <sup>46</sup>M. Eckert, "Max von Laue and the discovery of X-ray diffraction in 1912", *Annalen der Physik* **524**, A83–A85 (2012).
- <sup>47</sup>L. Meitner, "Über die entstehung der  $\beta$ -strahl-spektren radioaktiver substanzen", *Zeitschrift für Physik* **9**, 131–144 (1922).
- <sup>48</sup>P. Auger, "Sur les rayons  $\beta$  secondaires produits dans un gaz par des rayons x.", *CR Acad. Sci.(F)* **177**, 169 (1923).
- <sup>49</sup>J. Goodman, *Introduction to Fourier Optics*, 1968.
- <sup>50</sup>G. N. Hounsfield, "A method of and apparatus for examination of a body by radiation such as X-ray or gamma radiation", Patent Specification 1283915 (1972).
- <sup>51</sup>A. M. Cormack and G. N. Hounsfield, "The nobel prize in physiology or medicine 1979 explore perspectives",
- <sup>52</sup>H Rarback, D Shu, S. Feng, H Ade, J Kirz, I McNulty, D. Kern, T. Chang, Y Vladimirsky, N Iskander, et al., "Scanning x-ray microscope with 75-nm resolution", *Rev. Sci. Instrum.* **59**, 52–59 (1988).
- <sup>53</sup>E. Di Fabrizio, F. Romanato, M Gentili, S. Cabrini, B Kaulich, J. Susini, and R Barrett, "High-efficiency multilevel zone plates for keV X-rays", *Nature* **401**, 895–898 (1999).
- <sup>54</sup>W. Chao, P. Fischer, T Tyliczszak, S. Rekawa, E. Anderson, and P. Naulleau, "Real space soft x-ray imaging at 10 nm spatial resolution", *Opt. Express* **20**, 9777–9783 (2012).
- <sup>55</sup>M Lerotic, C Jacobsen, T Schäfer, and S Vogt, "Cluster analysis of soft x-ray spectromicroscopy data", *Ultramicroscopy* **100**, 35–57 (2004).
- <sup>56</sup>G Schmahl and D Rudolph, "High power zone plates as image forming systems for soft x-rays.", *Optik* **29**, 577–585 (1969).

## BIBLIOGRAPHY

---

- <sup>57</sup>M. A. Le Gros, G. McDermott, and C. A. Larabell, "X-ray tomography of whole cells", *Curr. Opin. Struct. Biol.* **15**, 593–600 (2005).
- <sup>58</sup>M. Beck, V. Lučić, F. Förster, W. Baumeister, and O. Medalia, "Snapshots of nuclear pore complexes in action captured by cryo-electron tomography", *Nature* **449**, 611–615 (2007).
- <sup>59</sup>R. Erni, M. D. Rossell, C. Kisielowski, and U. Dahmen, "Atomic-resolution imaging with a sub-50-pm electron probe", *Phys. Rev. Lett.* **102**, 096101 (2009).
- <sup>60</sup>R. Henderson, "Realizing the potential of electron cryo-microscopy", *Q. Rev. Biophys.* **37**, 3–13 (2004).
- <sup>61</sup>J. C. Spence and A. V. Crewe, "Experimental high-resolution electron microscopy", *Phys. Today* **34**, 90 (1981).
- <sup>62</sup>E Suzuki, "High-resolution scanning electron microscopy of immunogold-labelled cells by the use of thin plasma coating of osmium", *J. Microsc.* **208**, 153–157 (2002).
- <sup>63</sup>Z. Chen, M. Odstřil, Y. Jiang, Y. Han, M.-H. Chiu, L.-J. Li, and D. A. Muller, "Mixed-state electron ptychography enables sub-angstrom resolution imaging with picometer precision at low dose", *Nat. Commun.* **11**, 1–10 (2020).
- <sup>64</sup>B. C. Choy, R. J. Cater, F. Mancina, and E. E. Pryor Jr, "A 10-year meta-analysis of membrane protein structural biology: detergents, membrane mimetics, and structure determination techniques", *Biochim. Biophys. Acta - Biomembr.* **1863**, 183533 (2021).
- <sup>65</sup>E. H. Egelman, "The current revolution in cryo-EM", *Biophys. J.* **110**, 1008–1012 (2016).
- <sup>66</sup>E. Callaway, "Revolutionary cryo-EM is taking over structural biology.", *Nature* **578**, 201–202 (2020).
- <sup>67</sup>K. M. Yip, N. Fischer, E. Paknia, A. Chari, and H. Stark, "Atomic-resolution protein structure determination by cryo-EM", *Nature* **587**, 157–161 (2020).
- <sup>68</sup>J. Miao, P. Charalambous, J. Kirz, and D. Sayre, "Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens", *Nature* **400**, 342–344 (1999).
- <sup>69</sup>D. Paganin et al., *Coherent x-ray optics*, 6 (Oxford University Press on Demand, 2006).
- <sup>70</sup>K. A. Nugent, "Coherent methods in the x-ray sciences", *Adv. Phys.* **59**, 1–99 (2010).
- <sup>71</sup>H. N. Chapman and K. A. Nugent, "Coherent lensless x-ray imaging", *Nat. Photonics* **4**, 833–839 (2010).
- <sup>72</sup>I. A. Vartanyants and A. Singer, "Coherence properties of hard x-ray synchrotron sources and x-ray free-electron lasers", *New J. Phys.* **12**, 035004 (2010).
- <sup>73</sup>I. Robinson and R. Harder, "Coherent x-ray diffraction imaging of strain at the nanoscale", *Nat. Mater.* **8**, 291–298 (2009).

- <sup>74</sup>P. Kirkpatrick and A. V. Baez, "Formation of optical images by x-rays", *JOSA* **38**, 766–774 (1948).
- <sup>75</sup>A. Snigirev, V. Kohn, I. Snigireva, and B. Lengeler, "A compound refractive lens for focusing high-energy x-rays", *Nature* **384**, 49–51 (1996).
- <sup>76</sup>Q. Shen, I. Bazarov, and P. Thibault, "Diffractive imaging of nonperiodic materials with future coherent x-ray sources", *J. Synchrotron Radiat.* **11**, 432–438 (2004).
- <sup>77</sup>X. Huang, J. Nelson, J. Kirz, E. Lima, S. Marchesini, H. Miao, A. M. Neiman, D. Shapiro, J. Steinbrener, A. Stewart, et al., "Soft x-ray diffraction microscopy of a frozen hydrated yeast cell", *Phys. Rev. Lett.* **103**, 198101 (2009).
- <sup>78</sup>M. Rose, T. Senkbeil, A. R. von Gundlach, S. Stuhr, C. Rumancev, D. Dzhigaev, I. Besedin, P. Skopintsev, L. Loetgering, J. Viefhaus, et al., "Quantitative ptychographic bio-imaging in the water window", *Opt. Express* **26**, 1237–1254 (2018).
- <sup>79</sup>D. Dzhigaev, *Characterization of nanowires by coherent x-ray diffractive imaging and ptychography*, tech. rep. (Deutsches Elektronen-Synchrotron, 2017).
- <sup>80</sup>J. Carnis, L. Gao, S. Labat, Y. Y. Kim, J. P. Hofmann, S. J. Leake, T. U. Schüllli, E. J. Hensen, O. Thomas, and M.-I. Richard, "Towards a quantitative determination of strain in Bragg Coherent X-ray Diffraction Imaging: artefacts and sign convention in reconstructions", *Sci. Rep.* **9**, 1–13 (2019).
- <sup>81</sup>I. Vartanyants and I. Robinson, "Partial coherence effects on the imaging of small crystals using coherent x-ray diffraction", *J. Phys. Condens. Matter* **13**, 10593 (2001).
- <sup>82</sup>M. A. Pfeifer, G. J. Williams, I. A. Vartanyants, R. Harder, and I. K. Robinson, "Three-dimensional mapping of a deformation field inside a nanocrystal", *Nature* **442**, 63–66 (2006).
- <sup>83</sup>Y. Y. Kim, T. F. Keller, T. J. Goncalves, M. Abuin, H. Runge, L. Gelisio, J. Carnis, V. Vonk, P. N. Plessow, I. A. Vartanyants, et al., "Single alloy nanoparticle x-ray imaging during a catalytic reaction", *Sci. Adv.* **7**, eabh0757 (2021).
- <sup>84</sup>I. Vartanyants and O. Yefanov, *Coherent x-ray diffraction imaging of nanostructures*, tech. rep. (FS-Photon Science, 2015).
- <sup>85</sup>A. Ulvestad, A. Singer, H.-M. Cho, J. N. Clark, R. Harder, J. Maser, Y. S. Meng, and O. G. Shpyrko, "Single particle nanomechanics in operando batteries via lensless strain mapping", *Nano Lett.* **14**, 5123–5127 (2014).
- <sup>86</sup>M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light* (Elsevier, 2013).
- <sup>87</sup>S. Marchesini, "Phase retrieval and saddle-point optimization", *JOSA A* **24**, 3289–3296 (2007).

## BIBLIOGRAPHY

---

- <sup>88</sup>S. Marchesini, "Invited article: a unified evaluation of iterative projection algorithms for phase retrieval", *Rev. Sci. Instrum.* **78**, 011301 (2007).
- <sup>89</sup>H. Nyquist, "Certain topics in telegraph transmission theory", *Transactions of the American Institute of Electrical Engineers* **47**, 617–644 (1928).
- <sup>90</sup>C. E. Shannon, "Communication in the presence of noise", *Proceedings of the IRE* **37**, 10–21 (1949).
- <sup>91</sup>V. A. Kotelnikov, "On the transmission capacity of the 'ether' and of cables in electrical communications", in *Proceedings of the first all-union conference on the technological reconstruction of the communications sector and the development of low-current engineering* (Citeseer, 1933).
- <sup>92</sup>R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures", *Optik* **35**, 237–246 (1972).
- <sup>93</sup>J. R. Fienup, "Phase retrieval algorithms: a comparison", *Appl. Opt.* **21**, 2758–2769 (1982).
- <sup>94</sup>R. Bates, "Fourier phase problems are uniquely solvable in more than one dimension. I: underlying theory", *Optik (Stuttgart)* **61**, 5 (1982).
- <sup>95</sup>J. Seldin and J. Fienup, "Numerical investigation of the uniqueness of phase retrieval", *JOSA A* **7**, 412–427 (1990).
- <sup>96</sup>M. V. Klivanov, "On the recovery of a 2-d function from the modulus of its fourier transform", *J. Math. Anal. Appl.* **323**, 818–843 (2006).
- <sup>97</sup>R. Bates, "Uniqueness of solutions to two-dimensional fourier phase problems for localized and positive images", *Comput. Gr. Image Process.* **25**, 205–217 (1984).
- <sup>98</sup>J. R. Fienup, "Reconstruction of an object from the modulus of its fourier transform", *Opt. Lett.* **3**, 27–29 (1978).
- <sup>99</sup>J. R. Fienup, "Phase retrieval with continuous version of hybrid input-output", in *Frontiers in optics* (2003).
- <sup>100</sup>S. Marchesini, H. He, H. N. Chapman, S. P. Hau-Riege, A. Noy, M. R. Howells, U. Weierstall, and J. C. Spence, "X-ray image reconstruction from a diffraction pattern alone", *Phys. Rev. B* **68**, 140101 (2003).
- <sup>101</sup>P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. Bucksbaum, R. Coffee, F.-J. Decker, et al., "First lasing and operation of an ångstrom-wavelength free-electron laser", *Nat. Photonics* **4**, 641–647 (2010).
- <sup>102</sup>T. Ishikawa, H. Aoyagi, T. Asaka, Y. Asano, N. Azumi, T. Bizen, H. Ego, K. Fukami, T. Fukui, Y. Furukawa, et al., "A compact x-ray free-electron laser emitting in the sub-ångström region", *Nat. Photonics* **6**, 540–544 (2012).

- <sup>103</sup>A. Aquila, M. S. Hunter, R. B. Doak, R. A. Kirian, P. Fromme, T. A. White, J. Andreasson, D. Arnlund, S. Bajt, T. R. Barends, et al., “Time-resolved protein nanocrystallography using an x-ray free-electron laser”, *Opt. Express* **20**, 2706–2716 (2012).
- <sup>104</sup>S. Boutet, L. Lomb, G. J. Williams, T. R. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. DePonte, J. Steinbrener, R. L. Shoeman, et al., “High-resolution protein structure determination by serial femtosecond crystallography”, *Science* **337**, 362–364 (2012).
- <sup>105</sup>T. A. White, R. A. Kirian, A. V. Martin, A. Aquila, K. Nass, A. Barty, and H. N. Chapman, “CrystFEL: a software suite for snapshot serial crystallography”, *J. Appl. Crystallogr.* **45**, 335–341 (2012).
- <sup>106</sup>J. Miao, T. Ishikawa, B. Johnson, E. H. Anderson, B. Lai, and K. O. Hodgson, “High resolution 3D x-ray diffraction microscopy”, *Phys. Rev. Lett.* **89**, 088303 (2002).
- <sup>107</sup>K. Gaffney and H. N. Chapman, “Imaging atomic structure and dynamics with ultrafast x-ray scattering”, *Science* **316**, 1444–1448 (2007).
- <sup>108</sup>O. Y. Gorobtsov, U. Lorenz, N. M. Kabachnik, and I. A. Vartanyants, “Theoretical study of electronic damage in single-particle imaging experiments at x-ray free-electron lasers for pulse durations from 0.1 to 10 fs”, *Phys. Rev. E* **91**, 062712 (2015).
- <sup>109</sup>H. Yang and Z. Rao, “Structural biology of SARS-CoV-2 and implications for therapeutic development”, *Nat. Rev. Microbiol.* **19**, 685–700 (2021).
- <sup>110</sup>H. N. Chapman, A. Barty, M. J. Bogan, S. Boutet, M. Frank, S. P. Hau-Riege, S. Marchesini, B. W. Woods, S. Bajt, W. H. Benner, et al., “Femtosecond diffractive imaging with a soft-X-ray free-electron laser”, *Nat. Phys.* **2**, 839–843 (2006).
- <sup>111</sup>S. Herrmann, S. Boutet, B. Duda, D. Fritz, G. Haller, P. Hart, R. Herbst, C. Kenney, H. Lemke, M. Messerschmidt, et al., “CSPAD-140k: A versatile detector for LCLS experiments”, *Nucl. Instrum. Methods. Phys. Res. A* **718**, 550–553 (2013).
- <sup>112</sup>N. Meidinger, R. Andritschke, R. Hartmann, S. Herrmann, P. Holl, G. Lutz, and L. Strüder, “PnCCD for photon detection from near-infrared to X-rays”, *Nucl. Instrum. Methods. Phys. Res. A* **565**, 251–257 (2006).
- <sup>113</sup>B. Henrich, J. Becker, R. Dinapoli, P. Goettlicher, H. Graafsma, H. Hirsemann, R. Klanner, H. Krueger, R. Mazzocco, A. Mozzanica, et al., “The adaptive gain integrating pixel detector AGIPD a detector for the european XFEL”, *Nucl. Instrum. Methods. Phys. Res. A* **633**, S11–S14 (2011).
- <sup>114</sup>A. Aquila, A. Barty, C. Bostedt, S. Boutet, G. Carini, D. DePonte, P. Drell, S. Doniach, K. Downing, T. Earnest, et al., “The Linac Coherent Light Source single particle imaging road map”, *Struct. Dyn.* **2**, 041701 (2015).

## BIBLIOGRAPHY

---

- <sup>115</sup>A. Munke, J. Andreasson, A. Aquila, S. Awel, K. Ayyer, A. Barty, R. J. Bean, P. Berntsen, J. Bielecki, S. Boutet, et al., “Coherent diffraction of single Rice Dwarf virus particles using hard X-rays at the Linac Coherent Light Source”, *Sci. Data* **3**, 1–12 (2016).
- <sup>116</sup>H. K. Reddy, C. H. Yoon, A. Aquila, S. Awel, K. Ayyer, A. Barty, P. Berntsen, J. Bielecki, S. Bobkov, M. Bucher, et al., “Coherent soft X-ray diffraction imaging of coliphage PR772 at the Linac Coherent light source”, *Sci. Data* **4**, 1–9 (2017).
- <sup>117</sup>Y. Shi, K. Yin, X. Tai, H. DeMirici, A. Hosseinizadeh, B. G. Hogue, H. Li, A. Ourmazd, P. Schwander, I. A. Vartanyants, et al., “Evaluation of the performance of classification algorithms for XFEL single-particle imaging data”, *IUCrJ* **6**, 331–340 (2019).
- <sup>118</sup>A. Ignatenko, D. Assalauova, S. A. Bobkov, L. Gelisio, A. B. Teslyuk, V. A. Ilyin, and I. A. Vartanyants, “Classification of diffraction patterns in single particle imaging experiments performed at x-ray free-electron lasers using a convolutional neural network”, *Mach. learn.: Sci. Technol.* **2**, 025014 (2021).
- <sup>119</sup>D. Assalauova, A. Ignatenko, F. Isensee, D. Trofimova, and I. A. Vartanyants, “Classification of diffraction patterns using a convolutional neural network in single-particle-imaging experiments performed at x-ray free-electron lasers”, *J. Appl. Crystallogr.* **55** (2022).
- <sup>120</sup>M. R. Howells, T. Beetz, H. N. Chapman, C Cui, J. Holton, C. Jacobsen, J Kirz, E. Lima, S. Marchesini, H. Miao, et al., “An assessment of the resolution limitation due to radiation-damage in x-ray diffraction microscopy”, *J. Electron Spectrosc. Relat. Phenom.* **170**, 4–12 (2009).
- <sup>121</sup>A. Rose, “A unified approach to the performance of photographic film, television pickup tubes, and the human eye”, *J. Soc. Motion Pict. Eng.* **47**, 273–294 (1946).
- <sup>122</sup>M. J. Bogan, W. H. Benner, S. Boutet, U. Rohner, M. Frank, A. Barty, M. M. Seibert, F. Maia, S. Marchesini, S. Bajt, et al., “Single particle X-ray diffractive imaging”, *Nano Lett.* **8**, 310–316 (2008).
- <sup>123</sup>M. M. Seibert, T. Ekeberg, F. R. Maia, M. Svenda, J. Andreasson, O. Jönsson, D. Odić, B. Iwan, A. Rocker, D. Westphal, et al., “Single mimivirus particles intercepted and imaged with an X-ray laser”, *Nature* **470**, 78–81 (2011).
- <sup>124</sup>M. F. Hantke, D. Hasse, F. R. Maia, T. Ekeberg, K. John, M. Svenda, N. D. Loh, A. V. Martin, N. Timneanu, D. S. Larsson, et al., “High-throughput imaging of heterogeneous cell organelles with an x-ray laser”, *Nat. Photonics* **8**, 943–949 (2014).
- <sup>125</sup>G. Van Der Schot, M. Svenda, F. R. Maia, M. Hantke, D. P. DePonte, M. M. Seibert, A. Aquila, J. Schulz, R. Kirian, M. Liang, et al., “Imaging single cells in a beam of live cyanobacteria with an X-ray laser”, *Nat. Commun.* **6**, 1–9 (2015).



- <sup>126</sup>U Lorenz, N. Kabachnik, E Weckert, and I. Vartanyants, “Impact of ultrafast electronic damage in single-particle x-ray imaging experiments”, *Phys. Rev. E* **86**, 051911 (2012).
- <sup>127</sup>A. Allahgholi, J. Becker, A. Delfs, R. Dinapoli, P. Goettlicher, D. Greiffenberg, B. Henrich, H. Hirsemann, M. Kuhn, R. Klanner, et al., “The adaptive gain integrating pixel detector at the european xfel”, *J. Synchrotron Radiat.* **26**, 74–82 (2019).
- <sup>128</sup>U. Weierstall, R. Doak, J. Spence, D Starodub, D Shapiro, P Kennedy, J Warner, G. Hembree, P. Fromme, and H. Chapman, “Droplet streams for serial crystallography of proteins”, *Exp. Fluids* **44**, 675–689 (2008).
- <sup>129</sup>M. Wilm and M. Mann, “Analytical properties of the nanoelectrospray ion source”, *Anal. Chem.* **68**, 1–8 (1996).
- <sup>130</sup>R. Nazari, S. Zaare, R. C. Alvarez, K. Karpos, T. Engelman, C. Madsen, G. Nelson, J. C. Spence, U. Weierstall, R. J. Adrian, et al., “3D printing of gas-dynamic virtual nozzles and optical characterization of high-speed microjets”, *Opt. Express* **28**, 21749–21765 (2020).
- <sup>131</sup>T. Ekeberg, D. Assalauova, J. Bielecki, R. Boll, B. J. Daurer, L. A. Eichacker, L. E. Franken, D. E. Galli, L. Gelisio, L. Gumprecht, et al., “Observation of a single protein by ultrafast x-ray diffraction”, *bioRxiv* (2022).
- <sup>132</sup>I. V. Lundholm, J. A. Sellberg, T. Ekeberg, M. F. Hantke, K. Okamoto, G. van der Schot, J. Andreasson, A. Barty, J. Bielecki, P. Bruza, et al., “Considerations for three-dimensional image reconstruction from experimental data in coherent diffractive imaging”, *IUCrJ* **5**, 531–541 (2018).
- <sup>133</sup>V. Elser, “Phase retrieval by iterated projections”, *JOSA A* **20**, 40–55 (2003).
- <sup>134</sup>J. Clark, X Huang, R Harder, and I. Robinson, “High-resolution three-dimensional partially coherent diffraction imaging”, *Nat. Commun.* **3**, 1–6 (2012).
- <sup>135</sup>S. Bobkov, A. Teslyuk, T. Baymukhametov, E. Pichkur, Y. M. Chesnokov, D Assalauova, A. Poyda, A. Novikov, S. Zolotarev, K. Ikonnikova, et al., “Advances in modern information technologies for data analysis in cryo-em and xfel experiments”, *Crystallogr. Rep.* **65**, 1081–1092 (2020).
- <sup>136</sup>N.-T. D. Loh and V. Elser, “Reconstruction algorithm for single-particle diffraction imaging experiments”, *Phys. Rev. E* **80**, 026705 (2009).
- <sup>137</sup>K. Ayyer, T.-Y. Lan, V. Elser, and N. D. Loh, “Dragonfly: an implementation of the expand-maximize-compress algorithm for single-particle imaging”, *J. Appl. Crystallogr.* **49**, 1320–1335 (2016).
- <sup>138</sup>E. Sobolev, S. Zolotarev, K. Giewekemeyer, J. Bielecki, K. Okamoto, H. K. Reddy, J. Andreasson, K. Ayyer, I. Barak, S. Bari, et al., “Megahertz single-particle imaging at the european xfel”, *Commun. Phys.* **3**, 1–11 (2020).

- <sup>139</sup>D. Merkel et al., “Docker: lightweight linux containers for consistent development and deployment”, *Linux j.* **2014**, 2 (2014).
- <sup>140</sup>*SPI analysis pipeline tools*, [https://gitlab.com/spi\\_xfel](https://gitlab.com/spi_xfel).
- <sup>141</sup>I. T. Jolliffe, *Principal component analysis for special types of data* (Springer, 2002).
- <sup>142</sup>Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning”, *Nature* **521**, 436–444 (2015).
- <sup>143</sup>J. MacQueen et al., “Some methods for classification and analysis of multivariate observations”, in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, 14 (1967), pp. 281–297.
- <sup>144</sup>T. Yoshidome, T. Oroguchi, M. Nakasako, and M. Ikeguchi, “Classification of projection images of proteins with structural polymorphism by manifold: a simulation study for x-ray free-electron laser diffraction imaging”, *Phys. Rev. E* **92**, 032710 (2015).
- <sup>145</sup>C. H. Yoon, P. Schwander, C. Abergel, I. Andersson, J. Andreasson, A. Aquila, S. Bajt, M. Barthelmess, A. Barty, M. J. Bogan, et al., “Unsupervised classification of single-particle x-ray diffraction snapshots by spectral clustering”, *Opt. Express* **19**, 16542–16549 (2011).
- <sup>146</sup>E. R. Cruz-Chú, A. Hosseinizadeh, G. Mashayekhi, R. Fung, A. Ourmazd, and P. Schwander, “Selecting XFEL single-particle snapshots by geometric machine learning”, *Struct. Dyn.* **8**, 014701 (2021).
- <sup>147</sup>S. Bobkov, A. Teslyuk, R. Kurta, O. Y. Gorobtsov, O. Yefanov, V. Ilyin, R. Senin, and I. Vartanyants, “Sorting algorithms for single-particle imaging experiments at x-ray free-electron lasers”, *J. Synchrotron Radiat.* **22**, 1345–1352 (2015).
- <sup>148</sup>H. Steinhaus et al., “Sur la division des corps matériels en parties”, *Bull. Acad. Polon. Sci* **1**, 801 (1956).
- <sup>149</sup>G. Hamerly and C. Elkan, “Alternatives to the k-means algorithm that find better clusterings”, in *Proceedings of the eleventh international conference on Information and knowledge management* (2002), pp. 600–607.
- <sup>150</sup>A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *J. R. Stat. Soc. Ser. B Methodol.* **39**, 1–22 (1977).
- <sup>151</sup>S. H. Scheres, M. Valle, R. Nuñez, C. O. Sorzano, R. Marabini, G. T. Herman, and J.-M. Carazo, “Maximum-likelihood multi-reference refinement for electron microscopy images”, *J. Mol. Biol.* **348**, 139–149 (2005).
- <sup>152</sup>N. D. Loh, “A minimal view of single-particle imaging with X-ray lasers”, *Philos. Trans. R. Soc. B: Biol. Sci.* **369**, 20130328 (2014).
- <sup>153</sup>A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, *Adv. Neural Inf. Process. Syst.* **25** (2012).

- <sup>154</sup>C. Szegedy, A. Toshev, and D. Erhan, “Deep neural networks for object detection”, *Adv. Neural Inf. Process. Syst.* **26** (2013).
- <sup>155</sup>J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation”, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3431–3440.
- <sup>156</sup>X. Yang, M. Kahnt, D. Brückner, A. Schropp, Y. Fam, J. Becher, J.-D. Grunwaldt, T. L. Sheppard, and C. G. Schroer, “Tomographic reconstruction with a generative adversarial network”, *J. Synchrotron Radiat.* **27**, 486–493 (2020).
- <sup>157</sup>L. Wu, P. Juhas, S. Yoo, and I. Robinson, “Complex imaging of phase domains by deep neural networks”, *IUCrJ* **8**, 12–21 (2021).
- <sup>158</sup>L. Wu, S. Yoo, A. F. Suzana, T. A. Assefa, J. Diao, R. J. Harder, W. Cha, and I. K. Robinson, “Three-dimensional coherent x-ray diffraction imaging via deep convolutional neural networks”, *Npj Comput. Mater.* **7**, 1–8 (2021).
- <sup>159</sup>J. Zimmermann, B. Langbehn, R. Cucini, M. Di Fraia, P. Finetti, A. C. LaForge, T. Nishiyama, Y. Ovcharenko, P. Piseri, O. Plekan, et al., “Deep neural networks for classifying complex features in diffraction images”, *Phys. Rev. E* **99**, 063309 (2019).
- <sup>160</sup>I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).
- <sup>161</sup>M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge”, *Int. J. Comput. Vis.* **88**, 303–338 (2010).
- <sup>162</sup>O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., “Imagenet large scale visual recognition challenge”, *Int. J. Comput. Vis.* **115**, 211–252 (2015).
- <sup>163</sup>S. Ruder, “An overview of multi-task learning in deep neural networks”, arXiv preprint arXiv:1706.05098 (2017).
- <sup>164</sup>P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch SGD: training Imagenet in 1 hour”, arXiv preprint arXiv:1706.02677 (2017).
- <sup>165</sup>H. N. Chapman, A. Barty, S. Marchesini, A. Noy, S. P. Hau-Riege, C. Cui, M. R. Howells, R. Rosen, H. He, J. C. Spence, et al., “High-resolution ab initio three-dimensional x-ray diffraction microscopy”, *JOSA A* **23**, 1179–1200 (2006).
- <sup>166</sup>G. Harauz and M. van Heel, “Exact filters for general geometry three dimensional reconstruction.”, *Optik.* **73**, 146–156 (1986).
- <sup>167</sup>M. Van Heel and M. Schatz, “Fourier shell correlation threshold criteria”, *J. Struct. Biol.* **151**, 250–262 (2005).

## BIBLIOGRAPHY

---

- <sup>168</sup>P. Thibault and A. Menzel, “Reconstructing state mixtures from diffraction measurements”, *Nature* **494**, 68–71 (2013).
- <sup>169</sup>J. D. Bozek, “AMO instrumentation for the LCLS X-ray FEL”, *Eur. Phys. J.: Spec. Top.* **169**, 129–132 (2009).
- <sup>170</sup>K. R. Ferguson, M. Bucher, J. D. Bozek, S. Carron, J.-C. Castagna, R. Coffee, G. I. Curiel, M. Holmes, J. Krzywinski, M. Messerschmidt, et al., “The atomic, molecular and optical science instrument at the Linac Coherent Light Source”, *J. Synchrotron Radiat.* **22**, 492–497 (2015).
- <sup>171</sup>T. Osipov, C. Bostedt, J.-C. Castagna, K. R. Ferguson, M. Bucher, S. C. Montero, M. L. Swiggers, R. Obaid, D. Rolles, A. Rudenko, et al., “The LAMP instrument at the linac coherent light source free-electron laser”, *Rev. Sci. Instrum.* **89**, 035112 (2018).
- <sup>172</sup>H. Li, R. Nazari, B. Abbey, R. Alvarez, A. Aquila, K. Ayyer, A. Barty, P. Berntsen, J. Bielecki, A. Pietrini, et al., “Diffraction data from aerosolized coliphage PR772 virus particles imaged with the linac coherent light source”, *Sci. Data* **7**, 1–12 (2020).
- <sup>173</sup>H. K. Reddy, M. Carroni, J. Hajdu, and M. Svenda, “Electron cryo-microscopy of bacteriophage PR772 reveals the elusive vertex complex and the capsid architecture”, *Elife* **8**, e48496 (2019).
- <sup>174</sup>D. DePonte, U. Weierstall, K. Schmidt, J. Warner, D. Starodub, J. Spence, and R. Doak, “Gas dynamic virtual nozzle for generation of microscopic droplet streams”, *J. Phys. D Appl. Phys.* **41**, 195505 (2008).
- <sup>175</sup>U. Weierstall, J. Spence, and R. Doak, “Injector for scattering measurements on fully solvated biospecies”, *Rev. Sci. Instrum.* **83**, 035108 (2012).
- <sup>176</sup>W. H. Benner, M. J. Bogan, U. Rohner, S. Boutet, B. Woods, and M. Frank, “Non-destructive characterization and alignment of aerodynamically focused particle beams using single particle charge detection”, *J. Aerosol Sci.* **39**, 917–928 (2008).
- <sup>177</sup>L. Strüder, S. Epp, D. Rolles, R. Hartmann, P. Holl, G. Lutz, H. Soltau, R. Eckart, C. Reich, K. Heinzinger, 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”, *Nucl. Instrum. Methods Phys. Res. A: Accel. Spectrom. Detect. Assoc. Equip.* **614**, 483–496 (2010).
- <sup>178</sup>AMO SPI experimental data repository, <https://www.cxidb.org/id-156.html>.
- <sup>179</sup>D. Damiani, M. Dubrovin, I. Gaponenko, W. Kroeger, T. Lane, A. Mitra, C. O’Grady, A. Salnikov, A. Sanchez-Gonzalez, D. Schneider, et al., “Linac Coherent Light Source data analysis using psana”, *J. Appl. Crystallogr.* **49**, 672–679 (2016).
- <sup>180</sup>J. R. Fienup, “Phase retrieval algorithms: a personal tour”, *Appl. Opt.* **52**, 45–56 (2013).

- <sup>181</sup>S. Mäntynen, L.-R. Sundberg, H. M. Oksanen, and M. M. Poranen, “Half a century of research on membrane-containing bacteriophages: bringing new concepts to modern virology”, *Viruses* **11**, 76 (2019).
- <sup>182</sup>B. Peralta, D. Gil-Carton, D. Castaño-Díez, A. Bertin, C. Boulogne, H. M. Oksanen, D. H. Bamford, and N. G. Abrescia, “Mechanism of membranous tunnelling nanotube formation in viral genome delivery”, *PLoS Biol.* **11**, e1001667 (2013).
- <sup>183</sup>I. Santos-Pérez, H. M. Oksanen, D. H. Bamford, F. M. Goñi, D. Reguera, and N. G. Abrescia, “Membrane-assisted viral dna ejection”, *Biochim. Biophys. Acta Gen. Subj.* **1861**, 664–672 (2017).
- <sup>184</sup>A. Hosseinizadeh, G. Mashayekhi, J. Copperman, P. Schwander, A. Dashti, R. Sepehr, R. Fung, M. Schmidt, C. H. Yoon, B. G. Hogue, et al., “Conformational landscape of a virus by single-particle x-ray scattering”, *Nat. Methods* **14**, 877–881 (2017).
- <sup>185</sup>W Decking, S Abeghyan, P Abramian, A Abramsky, A Aguirre, C Albrecht, P Alou, M Altarelli, P Altmann, K Amyan, et al., “A MHz-repetition-rate hard X-ray free-electron laser driven by a superconducting linear accelerator”, *Nat. Photonics* **14**, 391–397 (2020).
- <sup>186</sup>A. P. Mancuso, A. Aquila, L. Batchelor, R. J. Bean, J. Bielecki, G. Borchers, K. Doerner, K. Giewekemeyer, R. Graceffa, O. D. Kelsey, et al., “The single particles, clusters and biomolecules and serial femtosecond crystallography instrument of the European XFEL: initial installation”, *J. Synchrotron Radiat.* **26**, 660–676 (2019).
- <sup>187</sup>P Abbamonte, F. Abild-Pedersen, P Adams, M Ahmed, F Albert, R. A. Mori, P Anfinrud, A Aquila, M Armstrong, J Arthur, et al., *New science opportunities enabled by lcls-ii x-ray lasers*, tech. rep. (2015).
- <sup>188</sup>K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks”, in *European conference on computer vision* (Springer, 2016), pp. 630–645.
- <sup>189</sup>B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network”, *arXiv preprint arXiv:1505.00853* (2015).
- <sup>190</sup>S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, in *International conference on machine learning* (PMLR, 2015), pp. 448–456.
- <sup>191</sup>D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization”, *arXiv preprint arXiv:1412.6980* (2014).
- <sup>192</sup>L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs”, *IEEE PAMI* **40**, 834–848 (2017).

## BIBLIOGRAPHY

---

- <sup>193</sup>F Isensee, P Jäger, J Wasserthal, D Zimmerer, J Petersen, S Kohl, J Schock, A Klein, T RoSS, S Wirkert, et al., “Batchgenerators—a python framework for data augmentation”, Zenodo <https://doi.org/10.5281/zenodo.3632567> (2020).
- <sup>194</sup>T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout”, arXiv preprint arXiv:1708.04552 (2017).
- <sup>195</sup>CNN *classification in SPI*, [https://gitlab.hzdr.de/hi-dkfz/applied-computer-vision-lab/collaborations/desy\\_2021\\_singleparticleimaging\\_cnn](https://gitlab.hzdr.de/hi-dkfz/applied-computer-vision-lab/collaborations/desy_2021_singleparticleimaging_cnn).
- <sup>196</sup>D. Assalauova, A. Ignatenko, F. Isensee, D. Trofimova, and I. A. Vartanyants, *Data repository for the article: “Classification of diffraction patterns using a convolutional neural network in single-particle-imaging experiments performed at X-ray free-electron lasers”*, <https://doi.org/10.5281/zenodo.6451444>, Apr. 2022.
- <sup>197</sup>X. Zhang, R. Jia, H. Shen, M. Wang, Z. Yin, and A. Cheng, “Structures and functions of the envelope glycoprotein in flavivirus infections”, *Viruses* **9**, 338 (2017).
- <sup>198</sup>M. G. Rossmann, “The study of virus structure and function: a personal history”, *Phys. Scr.* **89**, 098005 (2014).
- <sup>199</sup>J. L. Neira, “Nuclear magnetic resonance spectroscopy to study virus structure”, *Structure and Physics of Viruses*, 145–176 (2013).
- <sup>200</sup>H.-S. Kang, C.-K. Min, H. Heo, C. Kim, H. Yang, G. Kim, I. Nam, S. Y. Baek, H.-J. Choi, G. Mun, et al., “Hard X-ray free-electron laser with femtosecond-scale timing jitter”, *Nat. Photonics* **11**, 708–713 (2017).
- <sup>201</sup>I. A. Vartanyants, A. Singer, A. P. Mancuso, O. M. Yefanov, A Sakdinawat, Y. Liu, E Bang, G. J. Williams, G. Cadenazzi, B. Abbey, et al., “Coherence properties of individual femtosecond pulses of an x-ray free-electron laser”, *Phys. Rev. Lett.* **107**, 144801 (2011).
- <sup>202</sup>A Singer, F Sorgenfrei, A. P. Mancuso, N Gerasimova, O. Yefanov, J Gulden, T. Gorniak, T. Senkbeil, A Sakdinawat, Y Liu, et al., “Spatial and temporal coherence properties of single free-electron laser pulses”, *Opt. Express* **20**, 17480–17495 (2012).
- <sup>203</sup>C. Gutt, P Wochner, B Fischer, H Conrad, M Castro-Colin, S Lee, F Lehmkuhler, I Steinke, M Sprung, W Roseker, et al., “Single shot spatial and temporal coherence properties of the SLAC linac coherent light source in the hard X-ray regime”, *Phys. Rev. Lett.* **108**, 024801 (2012).
- <sup>204</sup>J. Miao, T. Ishikawa, I. K. Robinson, and M. M. Murnane, “Beyond crystallography: diffractive imaging using coherent x-ray light sources”, *Science* **348**, 530–535 (2015).
- <sup>205</sup>D. Assalauova and I. Vartanyants, “The structure of tick-borne encephalitis virus determined at x-ray free-electron lasers. simulations”, (submitted).

- <sup>206</sup>N. J. Barrows, R. K. Campos, K.-C. Liao, K. R. Prasanth, R. Soto-Acosta, S.-C. Yeh, G. Schott-Lerner, J. Pompon, O. M. Sessions, S. S. Bradrick, et al., “Biochemistry and molecular biology of flaviviruses”, *Chem. Rev.* **118**, 4448–4482 (2018).
- <sup>207</sup>T. C. Pierson and M. S. Diamond, “The continued threat of emerging flaviviruses”, *Nat. Microbiol.* **5**, 796–812 (2020).
- <sup>208</sup>T. Fuzik, P. Formanova, D. Ruzek, K. Yoshii, M. Niedrig, and P. Plevka, “Structure of tick-borne encephalitis virus and its neutralization by a monoclonal antibody”, *Nat. Commun.* **9**, 1–11 (2018).
- <sup>209</sup>E. Pichkur, V. Samygina, A. Ivanova, A. Y. Fedotov, A. Ivanov, E. Khvatov, A. Ishmukhmetov, and M. Vorovich, “Preliminary structural study of inactivated yellow fever virus”, *Crystallogr. Rep.* **65**, 915–921 (2020).
- <sup>210</sup>*Protein Data Bank*, <https://www.rcsb.org/>.
- <sup>211</sup>F. Long, M. Doyle, E. Fernandez, A. S. Miller, T. Klose, M. Sevvana, A. Bryan, E. Davidson, B. J. Doranz, R. J. Kuhn, et al., “Structural basis of a potent human monoclonal antibody against Zika virus targeting a quaternary epitope”, *Proceedings of the National Academy of Sciences* **116**, 1591–1596 (2019).
- <sup>212</sup>F. A. Rey, K. Stiasny, M.-C. Vaney, M. Dellarole, and F. X. Heinz, “The bright and the dark side of human antibody responses to flaviviruses: lessons for vaccine design”, *EMBO Rep.* **19**, 206–224 (2018).
- <sup>213</sup>*TBEV structure 5O6A*, <https://www.rcsb.org/structure/5o6a>.
- <sup>214</sup>*TBEV structure 5O6V*, <https://www.rcsb.org/structure/5o6v>.
- <sup>215</sup>K. Ayyer, P. L. Xavier, J. Bielecki, Z. Shen, B. J. Daurer, A. K. Samanta, S. Awel, R. Bean, A. Barty, M. Bergemann, et al., “3D diffractive imaging of nanoparticle ensembles using an x-ray laser”, *Optica* **8**, 15–23 (2021).

# Acknowledgments

First, I would like to thank Prof. Edgar Weckert for giving me an opportunity to be a part of the Coherent X-ray Diffractive Imaging group. I am grateful to Prof. Andreas Stierle and Prof. Ivan Vartaniants for the supervision of this thesis. I also want to thank Dr. Michael Sprung for being my mentor at DESY. I would like to express my gratitude to Dr. Fabian Isensee, Prof. Viacheslav Ilyin, Dr. Andrew Aquila, and the SPI consortium for great scientific collaboration.

I would like to thank my dearest friends from my home: Roman Sidorets, Nikita Sidorets, Shamil Musin, Yury Gavrilov, Artem Lebedskoy-Tambiev, Vladislav Buzdin, Maxim Milov for being real family to me and for the most memorable time in my life. There was no day in Germany when I was not thinking about you and I hope we will meet all together soon enough.

I want to thank Alexandr Vlasyuk, Nikita Udalov, Anton Kostin, Roman Zharenkov, Dmitry Degtev, Ilya Golokolenov, Andrey Kurdyuk, Vladimir Krasnobaev, Alexandr Kamin-sky, Semyon Nesterov, Denis Yagodkin, Vechyaslav Matyunin, David Aznaurov, Ilya Gukov, Javid Javadzade, Nikolay Lebedev, Fyodor Chervinsky, Nikolay Stepanov, Sergey Lemzyakov, Alexandr Svetogorov, Stepan Solodnev, Alexandr Penko, Ivan Kislitsyn, Pavel Semenenko for crazy days and sleepless nights, and for the enormous amount of positive energy and laughs during the university times and beyond. I would like to thank my oldest friends Yulia Nevalennaya and Elizaveta Shcherbakova. I wish we could meet more often and I can not wait to see you again.

I would like to thank Pavel Lytaev for being my flatmate since the beginning of my life in Germany. Thanks a lot for taking care of me all these years. I also thank Maria Lytaeva for being an amazing person with a big heart. And thanks to Dmitry Zabelsky for joining our flat-team and spending a fun time together.

I am thankful to the former and current members of the CXDI group at DESY and also invited scientists: Sergey Bobkov, Jerome Carnis, Dmitry Dzhigaev, Luca Gelisio, Oleg Gorobtsov, Gerard Hinsley, Alexandr Ignatenko, Ruslan Khubbutdinov, Young Yong Kim, Ruslan Kurta, Dmitry Lapkin, Sergey Lazarev, Nastasia Mukharamova, Kuan Hoon Ngoi, Max Rose, Shweta Singh, Anatoly Shabalin, Anton Teslyuk, Bihan Wang, Olexandr Yefanov, Ivan Zaluzhnyy for interesting discussions and productive group meetings. I want to ex-



press my special thanks to Dmitry Lapkin for your intelligence and readiness to answer all my questions, Sergey Bobkov for your willingness and patience while explaining algorithms to me, Alexandr Ignatenko for your guidance through the world of neural networks, Ruslan Khubbutdinov for being one of the smartest people I met, Young Yong Kim for your support in performing phase retrieval tasks, Max Rose for your guidance at the beginning of my PhD. Special thanks to the people who read my thesis and listened to my practice defense presentations, I appreciate your valuable input and advice.

I am specifically grateful to Dmitry Lapkin, Nastasia Mukharamova, Ruslan Khubbutdinov, Anatoly Shabalin, Sergei Riabchuk, Semyon Goncharov, Andrey Fedulin, Ivan Sobolev, Anton Sokolov, Maria Naumova, Sergiu Levcenko, Taisiia Cheremnykh, and Radik Batraev for travelling around Europe, for watching a lot of movies, and for having a great time together. I would like to separately thank Shabalin-Mukharamova Enterprise for your emotional support during the hard time. I am also thankful to Tej Varma Yenupuri, Philipp Pagel, Katharina Roeper, Frederik Bartelmann, Aleksandra Tolstikova, Aram Kalaydzhan, Svitozar Serkez, Yaryna Serkez, Lev Serkez, Alexandr Shakhov, Rustam Rysov for your companionship.

And, of course, I thank my parents for supporting and encouraging me from the beginning of my life; my dear sisters, my nieces, and my nephew for making me feel so welcomed and loved. And finally, I want to thank Ivan Maryshev for making my life brighter.