UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1764

Statistical processing of Flash X-ray Imaging of protein complexes

ALBERTO PIETRINI

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1651-6214
ISBN 978-91-513-0554-7
urn:nbn:se:uu:diva-372987

Dissertation presented at Uppsala University to be publicly examined in C8:301, BMC, Husargatan 3, Uppsala, Wednesday, 6 March 2019 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Dr Flavio Capotondi (Sincrotrone Trieste S.C.p.A.).

Abstract

Pietrini, A. 2019. Statistical processing of Flash X-ray Imaging of protein complexes. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1764. 88 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0554-7.

Flash X-ray Imaging (FXI) at X-ray Free Electron Lasers (XFELs) is a promising technique that permits the investigation of the 3D structure of molecules without the need for crystallization, by diffracting on single individual sample particles.

In the past few years, some success has been achieved by using FXI on quite large biological complexes (40 nm-1 μm in diameter size). Still, the desired dream-goal of imaging a single individual of a molecule or a protein complex (<15 nm in diameter size) has not been reached yet. The main issue that prevented us from a complete success has been the low signal strength, almost comparable to background noise. That is particularly true for experiments performed at the Coherent X-ray Imaging (CXI) instrument at the Linac Coherent Light Source (LCLS).

In this thesis, we provide a brief review of the CXI instrument (focusing on experiments there performed) and present a statistical method to deal with low signal-to-noise ratios. We take into account a variety of biological particles, showing the benefits of estimating a background model from sample data and using that for processing said data. Moreover, we present the results of some computer simulations in order to explore the limits and potentials of the proposed approach.

Last, we show another method (named COACS) that, being fed with the previous findings from the background model, helps obtaining clearer results in the phase retrieval problem.

Keywords: X-ray, XFEL, single particle, hit-finder, statistical hit-finder, single protein, protein complex, RNA polymerase II, FEL, free-electron laser, FXI, Flash X-ray Imaging, CXI, CSPAD, x-ray imaging, COACS

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

ISSN 1651-6214
ISBN 978-91-513-0554-7
urn:nbn:se:uu:diva-372987 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-372987)

Dedicated to Elide, Paolo and Giacomo

List of papers

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

I **Pietrini, A.** and Nettelblad, C. Artifact reduction in the CSPAD detectors used for LCLS experiments. *J. Synchrotron Rad.* **24**, 1-6 (2017).

II **Pietrini, A.** et al. A statistical approach to detect protein complexes at X-ray Free Electron Laser facilities. *Communications Physics* **1**, 92:1-11 (2018).

III **Pietrini, A.** and Nettelblad, C. Using convex optimization of autocorrelation with constrained support and windowing for improved phase retrieval accuracy. *Optics Express* **26**, 19:24422-24443 (2018).

List of additional papers

IV Daurer, B. J. et al. Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses. *IUCrJ* 4 (3), 251-262 (2017).

V Gorkhover T. et al. Femtosecond X-ray Fourier holography imaging of free-flying nanoparticles. *Nature Photonics* 12(3), 150 (2018).

VI Bielecki J. et al. Electrospray sample injection for single-particle imaging with X-ray lasers. *Science Advances*, submitted (2018).

Contents

Abbreviations

ADU	Analogue-to-Digital Unit
AMO	Atomic, Molecular and Optical Sciences
ASIC	Application-Specific Integrated Circuit
CDI	Coherent Diffractive Imaging
CPU	Central Processing Unit
CSPAD	Coornell-SLAC Pixel-Array Detector
CXI	Coherent X-ray Imaging
CXIDB	Coherent X-ray Imaging Data Base
DESY	Deutsches Elektron Syncrotron
DFT	Discrete Fourier Transform
EMC	Expansion Maximization Compression
ER	Error Reduction
FEL	Free Electron Laser
FLASH	Free Electron Laser in Hamburg
FXI	Flash X-ray Imaging
GDMV	Gas Dynamic Virtual Nozzle
GPU	Graphical Processing Unit
HIO	Hybrid Input Output
IP	Interaction Point
LCLS	Linac Coherent Light Source
LINAC	LINear ACcelerator
OSS	Oversampling Smoothness
RAAR	Relaxed Averaged Alternating Reflections
SASE	Self Amplified Stimulated Emission
SFX	Serial Femtosecond X-ray crystallography
SH	Speckle Healing
SLAC	Stanford Linear ACcelerator
SNR	Signal to Noise Ratio
XFEL	X-ray Free Electron Laser

Part I:
Introduction

Introduction

Explaining and understanding the surrounding reality has been of fundamental importance to all human beings. For a matter of mere survival, we learnt to discern and name things as a trivial way to deal with them. Later on, curiosity and fascination brought us to move toward a deeper understanding and exploration of nature. In particular, we started focusing on the principal investigator of it: the human being itself.

The endeavor of understanding the human machine (and everything related to it) was taken, among others, by the biological sciences. In just a few centuries, we passed from only studying large things (the full body, limbs, organs and tissues) to small ones (cells, bacteria, viruses, proteins, etc.).

Different scientific instruments have helped us in facing different challenges, spanning from the bare eye and optical microscopy, which take advantage of visible light for studying large bulks, to x-ray diffraction techniques — for studying the smallest living components. Visible light permits to achieve a resolution down to 200 nm (and it has been shown that limit can be pushed even to 100 nm [1]), whereas x-rays allow us to push resolution down to the Å scale. As biological samples are formed by atoms having low atomic number, the scattering power of a single particle (i.e. the signal associated to it) is very small and very hard to detect.

One way to make the scattered signal detectable is to amplify it, by crystallizing the sample in highly ordered and repetitive units. Such a technique is called x-ray crystallography and has proven very successful in determining the structure of a large variety of biological complexes (like e.g. the DNA [2][3][4] in 1953 and the hemoglobin in 1960 [5]). In fact, more than 100000 structures have been solved with this method since then.

Unfortunately, not all samples of biological interest can be crystallized. X-ray Free-Electron Lasers overcome the crystallization issues [6][7]. With their maximum brilliance a billion times greater than any other x-ray source and beam pulses of the order of 50 fs, XFELs [8] permit imaging of individual macro-molecular complexes.

The extremely energetic beam almost immediately destroys the sample, but the pulse itself is believed to be shorter than the time of explosion, so that most processes of sample damage occur when the pulse has already passed. The scattered signal is then a continuous diffraction pattern of the unaltered molecule. This process is known as *diffraction before destruction* [9]. A technique called Flash X-ray Imaging (FXI)[10][11][12], under development at the XFELs (especially at the LINAC Coherent Light Source – LCLS [13]), takes

advantage of the diffraction before destruction principle to image individual biological particles.

As the main scope of this method is to image a single instance of the sample, one big issue to solve in order to image and eventually reconstruct the structure of the molecule is how to deal with the experimental conditions including very low signal-to-noise ratio (SNR). The present work shows a new approach to cope with the special case of extremely weak scattered signals of this kind, almost comparable to the background signal.

First, we introduce the concept of x-ray, the different x-ray sources, the Coherent Diffractive Imaging (CDI) and the Flash X-ray Imaging (FXI) techniques (**chapter 1**); then we describe the experimental setup used in our experiments at the Coherent X-ray Imaging (CXI) instrument at the LCLS (**chapter 2**) and present in some detail the CSPAD detectors (**chapter 3**). After that, we discuss the background model and the statistical approach (**chapter 4**) along with its implementation (**chapter 5**). Finally, we outline our new method COACS (Convex Optimization of Autocorrelation with Constrained Support) — that could help in obtaining a better solution for the phase retrieval problem [14][15] (**chapter 6**) — and draw some considerations on this thesis work and possible future prospects (**chapter 7**).

Part II:
X-ray physics and XFEL science

1. X-ray diffraction

X-rays are very energetic photons, typically >100 eV, with wavelengths <10 nm, that can penetrate in depth into matter. Thanks to this peculiarity, they constitute a better probe to see the inner features of objects than visible light or electrons, which have a poor penetration depth [16]. Thus, since their discovery in 1895 by Röntgen [17], they have been widely used in science (medicine, physics, inorganic crystallography, structural biology, etc.), bringing great improvements in each of those fields [18][19].

1.1 X-ray production

Depending on the sources we are using and on the different studies we want to perform, there are different ways to produce x-rays:

- Characteristic x-ray emission
- Bremsstrahlung effect
- Synchrotron radiation
- Self-Amplified Spontaneus Emission (SASE) radiation

1.1.1 Characteristic x-rays

Characteristic x-ray production occurs when an electron collides with an inner core electron of an atom, thus ejecting it and creating a hole. The latter is filled by an outer shell electron, which loses energy in the transition by emitting an x-ray (**Fig. 1.1**).

Characteristic x-rays are commonly produced in x-ray tubes.

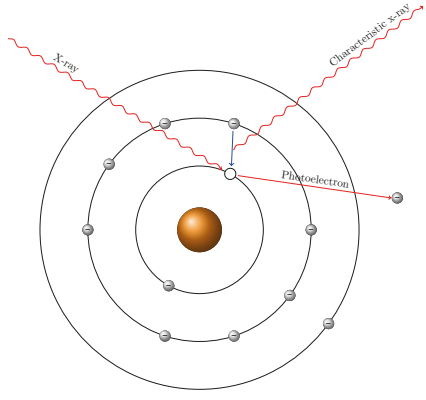

Figure 1.1. Characteristic x-ray emission: an incident x-ray or electron "kicks out" a core electron, thus creating a hole. The hole is replaced by an outer-shell electron, which gives off an x-ray in the process.

1.1.2 Bremsstrahlung effect

From theory, it is known that every time an electron is deflected or decelerated by the electromagnetic field of the nucleus of an atom, it loses energy by emitting photons of an appropriate wavelength (**Fig. 1.2**).

This principle is also the underlying physical process of synchrotron radiation.

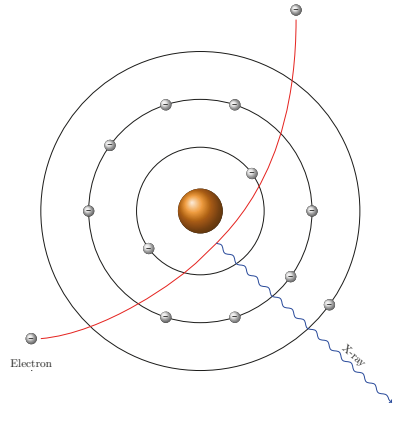

Figure 1.2. The incident electron is "braked" by the nucleus of the atom, thus losing energy emitting an x-ray.

1.1.3 Synchrotron radiation

In a synchrotron, electrons are accelerated to relativistic speeds and then deflected by strong magnetic fields. The deflection causes the emission of non-coherent x-rays. The brilliance is proportional to the number of electrons in each bunch.

By using synchrotrons, x-ray diffraction experiments can rapidly and reliably be performed on crystals — with size in the millimeter scale —, fibers, and powders.

1.1.4 SASE radiation

Nowadays, both circular and linear sources are available to perform a variety of experiments. Both of them takes advantage of electromagnetic fields to accelerate beam particles.

Circular sources (like synchrotrons) take advantage of shorts linear accelerators, that inject a particle beam into them, where it is accelerated and deflected by electromagnetic fields. The beam can then be used for colliding beam experiments or can be extracted to perform experiments on fixed targets.

Instead, linear sources are commonly used for experiments on targets. Typical examples of linear sources are XFELs, which rely on SASE radiation to generate coherent x-rays.

In an XFEL, an electron bunch uniformly distributed (produced by a laser beam hitting a copper plate) is injected at relativistic speed into an undulator (a periodic array structure of dipole magnets, whose magnetic field alternates with a defined wavelength along all its length).

The undulator (**Fig. 1.3**) thus deflects the incoming electrons, which start to emit x-ray photons within a narrow energy band, depending on the undulator strength. The electrons then interact with their own electromagnetic field (the photons themselves): if they are in phase, electrons decelerate; otherwise they gain energy and so accelerate.

This process creates a longitudinal fine structure within the electron bunch, so-called micro-bunching. The distribution of electrons in equidistant micro-bunches is equal to the wavelength of the emitted radiation causing the modulation. At this point, more electrons begin to radiate in phase (i.e. more photons are emitted coherently).

Whereas spontaneous undulator emission is proportional to the number of electrons in the bunch (N), in the case of micro-bunching all electrons radiate in phase, leading to a quasi-coherent emission of radiation, proportional to N^2.

Thus, if we consider that a typical electron bunch injected in the undulator consists of $\sim 10^9$ electrons, the brilliance at an XFEL is a billion times greater than the one at a synchrotron.

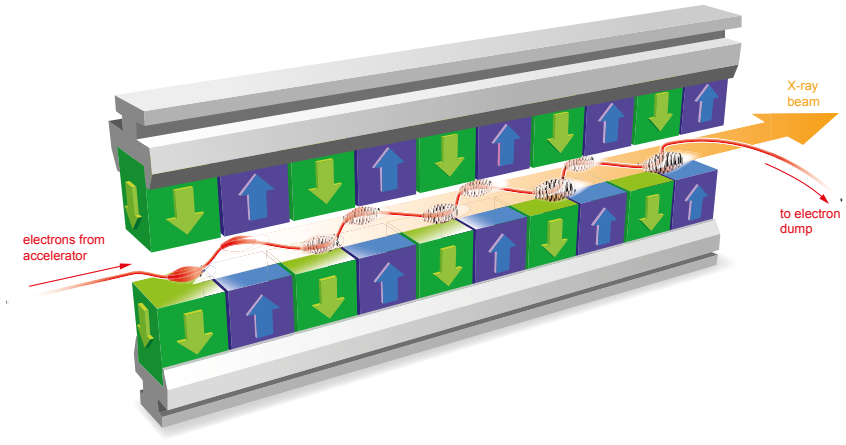

Figure 1.3. Representation of an undulator commonly used at XFELs. The magnetic fields deflect the electrons trajectory, causing electrons giving off x-rays.[Image courtesy of European XFEL].

We need a higher numbers of photons scattered by the sample in order to be able to perform diffraction experiments on smaller objects. As shown, an XFEL can provide this thanks to SASE.

1.2 X-ray crystallography

One of the most known and successful techniques used in biophysics, has been x-ray diffraction on crystals (or x-ray crystallography). A crystal is a very ordered system, constituted by the repetition of an atom or molecule in the 3D space. An ideal crystal is supposed to be infinite in every direction and is equivalent to the so-called *Bravais lattice*. The fundamental repeated unit is called *unit cell* and contains one or more atoms. Its repetition by translation in space originates the entire lattice (or crystal).

To understand how x-ray diffraction from a crystal works, we first explain x-ray radiation and matter interaction[1].

1.2.1 First Born approximation

In real scattering experiments, the total field acting in each point of the scatterer is constituted by the incident field plus the scattered field.

In scattering theory, to simplify calculations, the Born approximation is used. It is a perturbation method that consists in taking the incident field of a

[1]For the scattering theory and the diffraction and crystal theory explained in this chapter see *Coherent x-ray optics* [20] and *Introduction to Solid State Physics* [21].

wave, instead of its total field, as the only field interacting with the scatterer. It holds very well when the scattered field is small compared to the incident field.

Here, we introduce the Born approximation to the first-order (also known as first Born approximation).
Considering the elastic scattering of an incident plane wave, the Schrödinger wave equation is

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\boldsymbol{r})+V(\boldsymbol{r})\psi(\boldsymbol{r})=E\psi(\boldsymbol{r}) \tag{1.1}$$

$$(\nabla^2+k^2)\psi(\boldsymbol{r})=\frac{2m}{\hbar^2}V(\boldsymbol{r})\psi(\boldsymbol{r}) \tag{1.2}$$

where $k^2=\frac{2mE}{\hbar^2}$. Thus, the general solution of the eq. (1.2) can be expressed as follows in terms of Green's function:

$$\psi(\boldsymbol{r})=\phi(\boldsymbol{r})+\frac{2m}{\hbar^2}\int G(\boldsymbol{r}-\boldsymbol{r'})V(\boldsymbol{r'})\psi(\boldsymbol{r'})d\boldsymbol{r'} \tag{1.3}$$

where $\phi(\boldsymbol{r})=e^{i\boldsymbol{k}_0\boldsymbol{r}}$ is the incident plane wave and $G(\boldsymbol{r}-\boldsymbol{r'})$ is obtained by solving the point source equation:

$$(\nabla^2+k^2)G(\boldsymbol{r}-\boldsymbol{r'})=\delta(\boldsymbol{r}-\boldsymbol{r'}) \tag{1.4}$$

Eq. (1.4) has got two solutions:

$$G_+(\boldsymbol{r}-\boldsymbol{r'})=-\frac{1}{4\pi}\frac{e^{ik|\boldsymbol{r}-\boldsymbol{r'}|}}{|\boldsymbol{r}-\boldsymbol{r'}|}\quad\text{and}\quad G_-(\boldsymbol{r}-\boldsymbol{r'})=-\frac{1}{4\pi}\frac{e^{-ik|\boldsymbol{r}-\boldsymbol{r'}|}}{|\boldsymbol{r}-\boldsymbol{r'}|} \tag{1.5}$$

representing respectively the outgoing and incoming spherical waves. As we are interested in the scattered waves, which are outgoing, we use $G_+(\boldsymbol{r}-\boldsymbol{r'})$ as a solution.

We can then write eq. (1.3) as:

$$\psi(\boldsymbol{r})=\phi(\boldsymbol{r})-\frac{m}{2\pi\hbar^2}\int\frac{e^{ik|\boldsymbol{r}-\boldsymbol{r'}|}}{|\boldsymbol{r}-\boldsymbol{r'}|}V(\boldsymbol{r'})\psi(\boldsymbol{r'})d\boldsymbol{r'} \tag{1.6}$$

This is an integral equation and can be solved approximately, by means of a series of iterative approximations known as Born series.

At zero-order, $\psi^0(\boldsymbol{r})=\phi(\boldsymbol{r})$. Substituting this expression into eq. (1.6), renders first-order Born approximation:

$$\psi^1(\boldsymbol{r})=\psi^0(\boldsymbol{r})-\frac{m}{2\pi\hbar^2}\int\frac{e^{ik|\boldsymbol{r}-\boldsymbol{r'}|}}{|\boldsymbol{r}-\boldsymbol{r'}|}V(\boldsymbol{r'})\psi^0(\boldsymbol{r})d\boldsymbol{r'} \tag{1.7}$$

1.2.2 Fraunhofer diffraction

As a special case of the first Born approximation, we consider the case of a scattered field being detected far away from the source (*Fraunhofer diffraction*).

If we consider $\boldsymbol{r} \gg \boldsymbol{r}'$ and we call $|\boldsymbol{r} - \boldsymbol{r}'| = R$ we can then write eq. (1.7) as

$$\psi_f(\boldsymbol{r}) = e^{i\boldsymbol{k}_0 \cdot \boldsymbol{r}} - \frac{m}{2\pi\hbar^2} \frac{e^{ikr}}{R} \int e^{-\frac{ik}{R}\boldsymbol{r}\cdot\boldsymbol{r}'} V(\boldsymbol{r}') e^{i\boldsymbol{k}_0 \cdot \boldsymbol{r}'} d\boldsymbol{r}' \tag{1.8}$$

If we introduce the unit vector $\hat{\boldsymbol{r}} \equiv \frac{\boldsymbol{r}}{|\boldsymbol{r}|} \equiv \frac{\boldsymbol{r}}{R}$, eq. (1.8) becomes

$$\psi_f(\boldsymbol{r}) = e^{i\boldsymbol{k}_0 \cdot \boldsymbol{r}} - \frac{m}{2\pi\hbar^2} \frac{e^{ikr}}{R} \int V(\boldsymbol{r}') e^{-i\Delta\boldsymbol{k}\cdot\boldsymbol{r}'} d\boldsymbol{r}' \tag{1.9}$$

where $\Delta\boldsymbol{k} = k\hat{\boldsymbol{r}} - \boldsymbol{k}_0$.

To consider scattering from a perfect crystal, we assume the potential $V(\boldsymbol{r}) = t(\boldsymbol{r})S(\boldsymbol{r})$, where $t(\boldsymbol{r})$ is a function periodic in three dimensions and infinite in extent; $S(\boldsymbol{r})$ is the shape function, which accounts for the finite size of a real crystal. In the next calculations, we are going to assume the latter to be equal to unity.

Since $t(\boldsymbol{r})$ is periodic, given three linearly independent vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$, the following relation holds:

$$t(\boldsymbol{r} + u\boldsymbol{a} + v\boldsymbol{b} + w\boldsymbol{c}) = t(\boldsymbol{r}) \tag{1.10}$$

with u,v and w integers.

All the possible linear combinations of the three vectors form the *direct lattice*.

The periodicity of $t(\boldsymbol{r})$, allows for expansion in Fourier series, as

$$t(\boldsymbol{r}) = \sum_{hkl} t_{hkl} e^{i\boldsymbol{g}_{hkl}\cdot\boldsymbol{r}} \tag{1.11}$$

where $\boldsymbol{g}_{hkl} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$ describes the so-called *reciprocal lattice*; h,k,l are integer values and $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ the basis vectors. t_{hkl} are instead the Fourier coefficients.

Substituting the potential $V(\boldsymbol{r})$ in the argument of the integral in eq. (1.9), with the expression in eq. (1.11), renders

$$\int V(\boldsymbol{r}') e^{-i\Delta\boldsymbol{k}\cdot\boldsymbol{r}'} d\boldsymbol{r}' = \int t(\boldsymbol{r}') e^{-i\Delta\boldsymbol{k}\cdot\boldsymbol{r}'} d\boldsymbol{r}' = \sum_{hkl} t_{hkl} \int e^{-i(\boldsymbol{g}_{hkl} - \Delta\boldsymbol{k})\cdot\boldsymbol{r}'} d\boldsymbol{r}' \tag{1.12}$$

The integral in eq. (1.12) is proportional to a Dirac delta, so we can write eq. (1.12) as

$$\psi_f(\boldsymbol{r}) = e^{i\boldsymbol{k}_0\cdot\boldsymbol{r}} - \frac{4m\pi^2}{\hbar^2}\frac{e^{ikr}}{R}\sum_{hkl} t_{hkl}\,\delta(\boldsymbol{g}_{hkl} - \Delta\boldsymbol{k}) \tag{1.13}$$

From the above equation, it is clear that the scattered waves are observed only when $\boldsymbol{g}_{hkl} - \Delta\boldsymbol{k} = 0$.

The equality

$$\boldsymbol{g}_{hkl} = \Delta\boldsymbol{k} \tag{1.14}$$

is called the *von Laue diffraction condition*. This condition can be shown to be equivalent to the *Bragg's law*. It can be also expressed and visualized in terms of the *Ewald sphere*.

Both of these are explained in the next sections.

1.2.3 Atomic and molecular form factors

The x-ray scattering power for an atom is coming from the electrons. Therefore, the atomic scattering power can be described as:

$$f(\boldsymbol{q}) = \int n(\boldsymbol{\rho})e^{-i\boldsymbol{\rho}\cdot\boldsymbol{q}}d\boldsymbol{\rho} \tag{1.15}$$

where $n(\boldsymbol{\rho})$ is the electron density, depending on the radius $\boldsymbol{\rho}$; $\boldsymbol{q}$ is the scattering vector (**Fig. 1.4**), given by the change in direction between the incident wave and the scattered one $\boldsymbol{q} = \boldsymbol{k}_i - \boldsymbol{k}_s$. We call $f(\boldsymbol{q})$ the atomic scattering factor (or *form factor*).

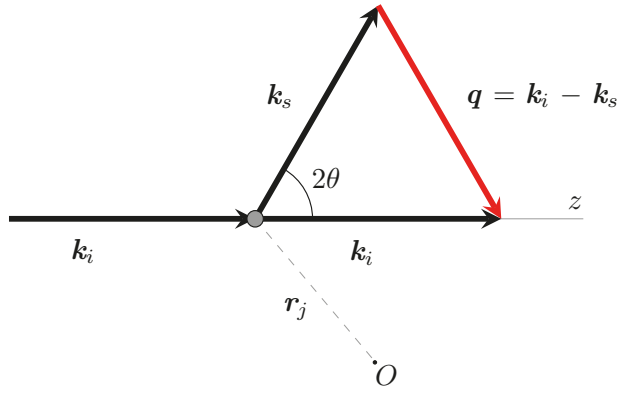

Figure 1.4. Geometrical representation of the scattering vector $\boldsymbol{q}$. $\boldsymbol{k}_i$ and $\boldsymbol{k}_s$ are respectively the incident and scattered wave vectors of a wave interacting with an atom at a point $\boldsymbol{r}_j$ from an arbitrary origin O.

If we now move to the case of a compound object and consider a molecule, the amplitude of the outgoing scattered wave received at an outside point is again proportional to the electron density in the volume considered. The volume element dV contains more atoms in specific positions $\boldsymbol{r}_j$, relative to an arbitrary origin, which interferes with each other. If we consider all the electron density of the different atoms centered in the positions $\boldsymbol{r}_j$, the atoms scatter as point sources, and the resulting amplitude is then described by:

$$F(\boldsymbol{q}) = \sum_j f_j(\boldsymbol{q}) e^{-i\boldsymbol{r}_j \cdot \boldsymbol{q}} \tag{1.16}$$

$F(\boldsymbol{q})$ is called the *molecular form factor*. It represents the amplitude of the scattered wave from a molecule.

When using conventional x-ray sources, the diffracted signal of a single molecule is too low to be detectable or usable. Only by amplifying it, one can investigate the structure of a molecule.

The scattering from a crystal can be calculated considering the scattering from each cell at a position $\boldsymbol{\rho}$ with respect to the origin; and each atom in the cell being at position $\boldsymbol{r}_j$ with respect from the cell origin $\boldsymbol{\rho}$. Eq. (1.15) and (1.16), for a crystal become:

$$A(\boldsymbol{q}) = \sum_{\boldsymbol{\rho}} \sum_j f_j(\boldsymbol{q}) e^{-i(\boldsymbol{\rho}+\boldsymbol{r}_j)\cdot\boldsymbol{q}} = \sum_j f_j(\boldsymbol{q}) e^{-i\boldsymbol{r}_j\cdot\boldsymbol{q}} \sum_{\boldsymbol{\rho}} e^{-i\boldsymbol{\rho}\cdot\boldsymbol{q}} = F(\boldsymbol{q}) \sum_{\boldsymbol{\rho}} e^{-i\boldsymbol{\rho}\cdot\boldsymbol{q}} \tag{1.17}$$

In this case, $F(\boldsymbol{q})$ is called the *geometric structure factor*, representing the amplitude of the scattered wave from a single unit cell.

In a crystal there are billions of unit cells, therefore the signal can be amplified enough to be detectable.

1.2.4 Bragg's law

For plane waves generated by a point source, amplification of the signal (i.e. constructive interference) occurs every time their phase difference is equal to an even number of times π.

The geometrical condition (**Fig. 1.5**) for constructive interference to happen in the case of waves having a wavelength λ, scattered by planes of atoms whose spacing is d, can be found based on *Bragg's law*:

$$2d \sin\theta = n\lambda \tag{1.18}$$

where θ is the angle formed by the incident and diffracted wave with the

atomic planes and n is a positive integer that represents the order of an atomic plane in the crystal.

For a perfect crystal, this law is equivalent to the Laue condition in eq. (1.14).

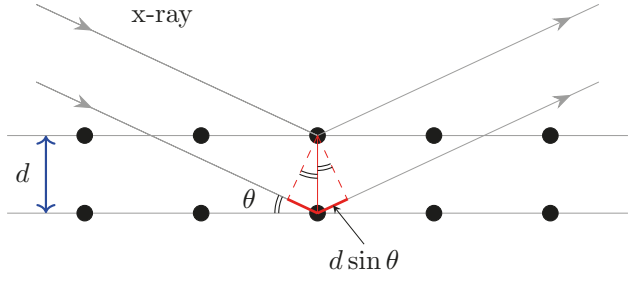

Figure 1.5. Bragg's law describes diffraction from successive parallel atomic planes in a crystal. d is the spacing of those planes; θ is the angle formed by the incoming and outgoing x-rays with respect to the planes; and $d\sin\theta$ is half the optical path difference between the two incoming waves.

1.3 Ewald sphere

The Ewald sphere is a geometrical interpretation of the von Laue condition. All the scattering points in a crystal lattice are defined by such a construction (**Fig. 1.6**). From the figure, we can see that only the points of the reciprocal lattice that lie on the sphere undergo interference and cause diffraction. To cover the scattering from the whole crystal, we can rotate the source, the crystal or change the x-ray wavelength.

The Ewald sphere construction holds analogously also in the case of x-ray diffraction from a single molecule. In the far field regime (small-angle scattering), the 2D diffraction pattern can be approximated with the scattering coming from a portion of the surface of the Ewald sphere. We can sample the whole Ewald sphere by taking snapshots of the sample at different orientations in space. The finer the sampling is, the better we can reconstruct the Fourier object, by orienting all the patterns with respect to each other. From the Fourier object we can the retrieve the 3D electronic structure of the molecule.

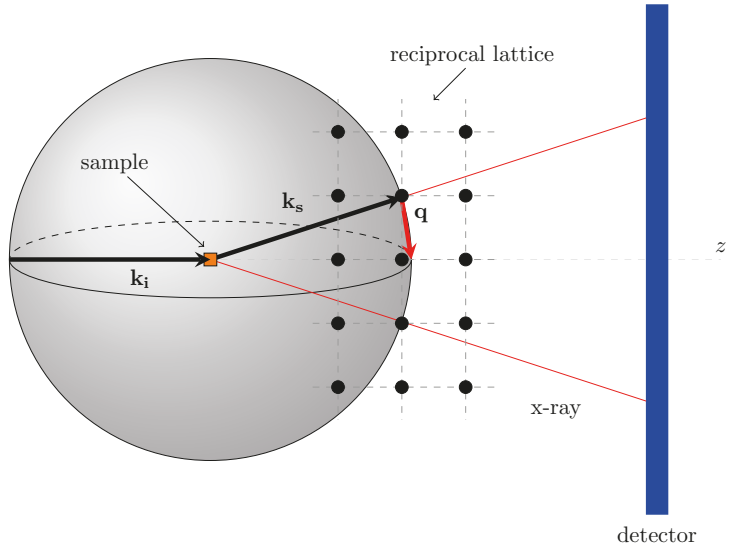

Figure 1.6. Geometrical construction of the Ewald sphere. The Ewald sphere is centered on the origin of the sample and its radius is equal to the modulus of the incoming x-ray (or, equivalently, to the reciprocal of its wavelenght). Only the points of the reciprocal lattice that intersect the sphere cause diffraction.

1.4 X-ray imaging at XFELs

In a crystallographic experiment at a synchrotron, the absorbed x-ray dose, as well as the radiation damage, is distributed over all unit cells, thus limiting the total damage per cell. Only after some time of exposure, enough damage accumulates to become appreciable [22]. Before that time, diffraction is still representative of the original sample.

Instead, in XFEL experiments, the absorbed x-ray dose by a crystal is so high that the sample is irremediably damaged. CDI/FXI techniques permit to overcome radiation damage in XFEL experiments.

1.4.1 CDI – Coherent Diffractive X-ray Imaging

Coherent Diffractive X-ray Imaging (CDXI or simply CDI) is a lensless technique for two- or three-dimensional structure determination at the nanometer

scale of objects difficult to crystallize. It takes advantage of a highly coherent x-ray beam, as non-random interference of the scattered waves is needed to produce a usable diffraction pattern.

On the one hand, as no lenses are used to focus the image resulting from illuminating the object, the final result is aberration-free and is thus limited only by diffraction and dose exposure. On the other hand, the resulting diffraction pattern contains information solely on the magnitude of the scattered waves and all the information on the phases are missing. That means that the 2D or 3D reconstruction in the real space cannot be done directly, but phases need to be retrieved, for example, by using an iterative feedback algorithm [14][15].

1.4.2 FXI – Flash X-ray Imaging

Flash X-ray Imaging is a technique under development at XFELs, which has its roots in CDI. The aim is to perform x-ray diffraction imaging at relevant resolution on individual macromolecular complexes. The high brilliance of the FEL allows enough scattered signal even from a single biological molecule, but the energy deposited on the sample by the beam pulse turns it almost immediately into a plasma. The radiation damage can be outrun by exposing the sample to only a very short x-ray pulse ($\sim$ 50fs), shorter than time scales where any significant movement of the atoms in the molecules happen ($\sim$ 1ps). Thus, the diffraction happens before any other damaging process occurs in the molecule, and the resulting signal still constitutes a 2D snapshot of the unaltered molecule. This main idea underlying the FXI is known as diffraction before destruction principle [9].

1.4.3 SFX – Serial Femtosecond Crystallography

Another successful technique possible at XFELs is SFX [23][24]. It consists in performing CDI on crystals of nanometer sizes (nanocrystals). This technique has been shown to be particularly useful for proteins that are impossible to crystallize in large crystals – e.g. membrane proteins [6] [7].

The first SFX experiment was carried out at LCLS, in December 2009. The sample studied was primarily photosystem I (PSI), which is responsible for converting light energy from the sun to chemical energy in plants, green algae, and cyanobacteria. It is a membrane protein complex constituted by 36 proteins and 381 cofactors[23][25].

This experiment constituted a proof of concept for SFX. Thanks to the thousands of diffraction patterns collected, it allowed for the determination of the PSI structure.

SFX has been proved to overcome radiation damage and to work on crystals with only a few hundred unit cells, thus showing the validity of the diffraction before the destruction principle [23][24][26].

Hence, data can be collected from nanocrystals, which show less long-range disorder than their larger counterparts, making them ideal candidates for the structure determination of challenging proteins. Those features allow for high-resolution determination of proteins structure at XFELs.

2. FXI experiments at the LCLS

Before describing a typical FXI experiment, we present in some details how LCLS works.

Laser pulses — at 120 Hz — in the ultraviolet wavelength regime travel to an injector "gun" and strike the surface of a copper plate. The plate releases electrons, which are accelerated in a 1 km section of the 3 km SLAC linear accelerator.

Accelerated electrons enter the LCLS Undulator Hall, where undulators force them to give off a coherent X-ray beam, as described in section 1.1.4.

The electrons, no longer needed, are discarded and the X-ray laser pulses are delivered to six specialized experimental stations.

Thanks to the features we are going to describe in the following section, and thanks to CSPAD technology deployment [27] [28] [29] [30], the CXI instrument, which was the original instrument designed for FXI experiments, is (under ideal circumstances) a better alternative to the Atomic, Molecular and Optical science (AMO) instrument [31] — that instead deploys pnCCD technology [32], operates in the energy range 480 eV – 2 keV and provides a beam focal spot of $\geq 1\,\mu m$. As of yet, the AMO instrument has in practice been more successful for imaging biological particles.

2.1 The CXI instrument

The Coherent X-ray Imaging (CXI) [33] instrument consists of a series of tools especially suited to perform coherent diffractive imaging experiments thanks to the near complete transverse coherence of the LCLS beam, by using hard x-rays in a vacuum sample environment. It is also able to perform Serial Femtosecond Crystallography (SFX) [23] measurements.

The CXI end station is equipped with a variety of tools and devices in order to make it possible to use multiple techniques such as x-ray emission spectroscopy, back-scattering, small and wide angle scattering, ion and electron time of flight spectroscopy. A pump laser system is also available for time-resolved experiments in the femtosecond time scale.

The CXI instrument is available for any scientific field requiring use of the LCLS beam and is especially suitable for any forward scattering experiment which may benefit from a vacuum sample environment, including structural biology, material science, materials in extreme conditions, atomic molecular

and optical physics, chemistry, soft condensed matter and high field x-ray science.

CXI operates primarily in the 5 keV – 11 keV range. Samples can be introduced to the x-ray beam either fixed on targets or using a particle injector that can deliver samples in an aerosolized jet to the beam. The experiments reported in the present work make use of the latter delivery method. Two vacuum chambers are available at the CXI instrument: one which provides an x-ray beam focus spot of 1 μm and the other which provides an x-ray beam focus spot of 100 nm. The capability of providing so narrow foci allows the concentration of the beam strength in a narrow spot, whereas the short wavelengths ensure a good resolution.

Those features make CXI ideal for imaging small samples.

Two imaging detectors are mounted at CXI, both making use of CSPAD technology: a CSPAD-2.3M (designed for wide-angle/high-resolution experiments) and a CSPAD-140k (designed for small-angle/low-resolution experiments) [27][28][29]. Both detectors are placed on a movable stage, so they can be easily moved along the incident x-ray beam direction to satisfy the different needs of the users. A dumping system (*beamstop*) is put in between the two detectors, in order to prevent any pixel damage due to the incident direct laser beam. In **Fig. 2.1** a schematic experimental setup of the CXI instrument is shown.

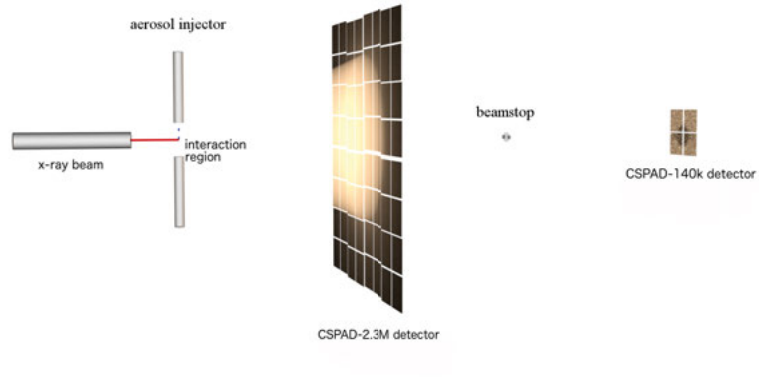

Figure 2.1. Typical experimental setup for flash x-ray imaging at the CXI end station at LCLS. The sample is injected in aerosol form by an injector. When the aerosol reaches the vacuum chamber, it is hit by the beam pulse in the interaction region. Due to the severe ionization from the x-ray pulse, the sample turns into a plasma and starts exploding. As the atoms in the sample have significant movement on time scales of $10^{-13} - 10^{-12}$s, which is longer than the beam pulse $10^{-15} - 10^{-14}$s, diffraction before destruction is considered feasible: the scattered photons are recorded in the front (high resolution) detector and in the back (low resolution) detector. A beamstop is interposed between the two detectors to avoid the full intensity of the direct beam hitting and damaging the center of the CSPAD-140k detector.

2.1.1 Sample delivery

Sample is delivered into the x-ray beam using an aerosol injector. It is present in a volatile buffer solution (ammonium acetate, in the case of RNA polymerase II) and then introduced into the injector via a gas dynamic virtual (GDMV) nozzle [34]. After this, the aerosol stream passes through a skimmer and relaxation chamber and is finally narrowly focused into the interaction point (IP) by an aerodynamic lens system [35] [36].

Even though the particle beam can be regulated by tuning gas and liquid flow and skimmer pressure, the event of a sample particle being in the x-ray focus (the interaction region) when a pulse hits is stochastic. Furthermore, the orientation of the particle at that moment is also random. Too high a concentration will result in hits of aggregates and multiple particles in the focus at the same time, while on the other hand a low concentration will lead to low hit rates, i.e. a low fraction of shots containing any sample diffraction at all.

2.2 The hit-finding problem

Due to the very nature of sample delivery, particle hits — and consequently sample diffraction pattern — are purely stochastic. That means that we can collect a variety of 2D snapshots, but only few representative of injected material, non-buffer diffraction (what we call *sample hits*). As we are interested only in those, we need a tool to discern them from background (the *hit-finding* problem). This tool is what we call a *hit-finder*. The state of art of hit-finders and of the hit-finding problem will be described in more details in **chapter 4**.

Part III:
Project

3. Artifact reduction in the CSPAD detectors used for LCLS experiments

Our motivation for the work was the analysis of data collected in a low-flux low signal-to-noise ratio regime. The data were collected in May 2013, during an attempt to image the RNA polymerase II — the first protein complex ever studied at an XFEL — at the CXI instrument at the LCLS.

Even after the first offsets removal, those data showed the presence of a spatially non-uniform artifact. Here we describe this artifact and show how to reduce it.

3.1 CSPAD detectors

The CXI instrument is equipped with two distinct CSPADs (Cornell-SLAC pixel Pixel Array Detector): the CSPAD-140k and the CSPAD-2.3M. Each detector is composed of multiple units of the same fundamental component: a CSPAD-2x1 module (388×185 pixels). Every CSPAD-2x1 module consists of 2 ASICs (Application-Specific Integration Circuit), having an independent readout electronic. Each ASIC is exactly half a CSPAD-2x1 module. Every CSPAD pixel – $110\,\mu m^2 \times 110\,\mu m^2$ in size – consists of an analog memory cell, a counting register and a comparator. The readout is based on the number of counted ticks before a distributed reference voltage matches the level within the cell [28][30]. The raw readout of the number of ticks is in analogue-to-digital units (ADUs) and it is proportional to the total energy of the photons hitting the pixel.

The CSPAD-2.3M (**Fig. 3.1**) consists of 64 ASICs and is the closest to the interaction point. It provides high-resolution/wide-angle scattering information. The CSPAD-140k, on the other hand, is located further downstream, typically of the order of 2 m from the interaction point. It consists of 2 CSPAD-2x1 modules and provides low-resolution/small-angle scattering features.

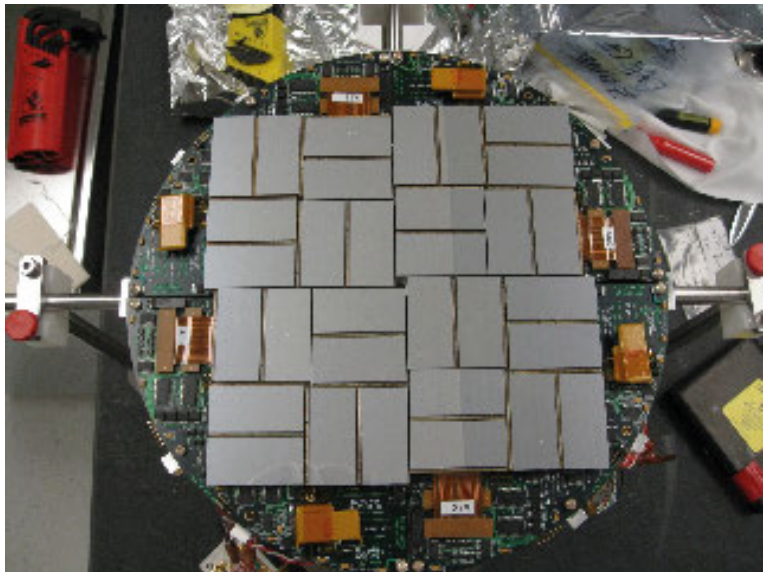

Figure 3.1. The CSPAD-2.3M consists of 32 CSPAD-2x1 modules – each one in turn consisting of 2 ASICs. Every ASIC contains 185×194 pixels ($110\,\mu m^2 \times 110\,\mu m^2$ each). [Image courtesy of SLAC/LCLS Detector Group].

3.2 Data processing

The raw data signals are stored in the proprietary XTC file format, which can be read and manipulated using PSANA, a specific environment developed at the LCLS [37]. It offers a number of calibration and correction steps necessary to perform data analysis:

- Pedestals
- Common mode correction per ASIC
- Gain determination

3.2.1 Pedestals

During experiments, we collect runs with the x-ray beam off (called *dark runs* in jargon), in order to account for detector background signal levels. We estimate the presence of noise by taking the mode of all data in such runs, per pixel. Thus, the most frequent ADU value within this set corresponds to the zero-photon peak (no photons) in a specific pixel (**Fig. 3.2**).

When collecting full beamline background or actual sample diffraction patterns, this typical dark frame is subtracted from those. The standard approach is to treat the dark frame from the latest dark run as a common subtraction term for subsequent sample runs.

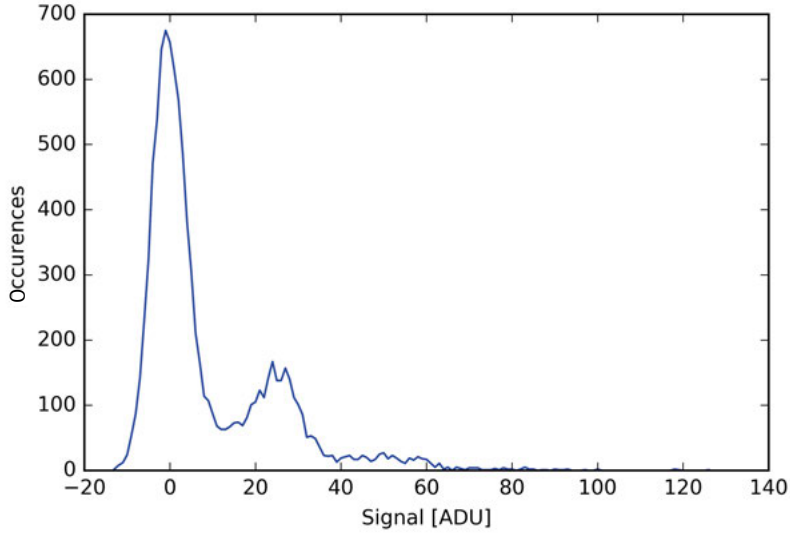

Figure 3.2. ADU histogram for a single pixel of the CSPAD-140k in L730 beamtime data (run 399), recorded with ongoing injection of sample. The zero-photon peak is clearly dominating.

3.2.2 Common mode correction per-ASIC

Common mode subtraction is needed in order to cancel out frame-wise offsets in the detector setup. This step is done ASIC-wise (instead of applying it on the overall frame) because the detector consists of an assembly of different ASICs, which have independent electronic read-out. There is support for several common mode correction settings in PSANA, and the current recommended approach can be found in the official documentation [38]. In our processing, we define the common mode as the median of all the values registered in the ASIC for a certain frame.

The common mode processing step ensures that the signal for each pixel becomes centered around 0 for zero-photon events. The choice of subtracting the median was preceded by evaluating the mode (most frequent) value and the mean value. The mode value proved inefficient, since it is more sensitive to noise, and, in the case of a strong event, tends to the 1-photon peak value (as a larger number of photons are revealed).

The mean, on the other hand, is prone to strong variations and tends towards the 1-photon peak markedly, even if we cap which values we take into account.

Instead, the median was chosen as it is more statistically robust, remaining unaffected by the presence of larger or lower values in a distribution.

3.2.3 Gain determination

Following the common mode correction, conventional PSANA identifies the zero- and one-photon peak values, fitting them with two separate Gaussian functions. The noise level and the overall quality of the signal can be judged based on the sharpness of the zero-photon peak and on the separation — in terms of standard deviations — of the two peaks. The more precise the discrimination of the one-photon peak, the better is the conversion in number of photons of the ADU data.

3.3 Detector artifact

Despite these corrections, when we tried to perform size determination on the patterns using fitting of spherical models, we had unexpected results. We noticed that some patterns seemed consistent with a very small particle diffracting off-center. The patterns presented an increasing gradient towards the edge, putatively representing the outermost part of a central speckle centered outside the detector.

Looking at many other patterns, we found out that a part of the right CSPAD-2x1 module was biased. By subtracting the expected background from the strongest hits, averaging the result and down-sampling it, we revealed the presence of a framewise column artifact in L730 data (left picture in **Fig. 3.3**). This artifact was spatially non-uniform, varying frame by frame. By applying an additional correction step we managed to correct for it (right-hand panel in **Fig. 3.3**) reducing the noise level (**Fig. 3.5**).

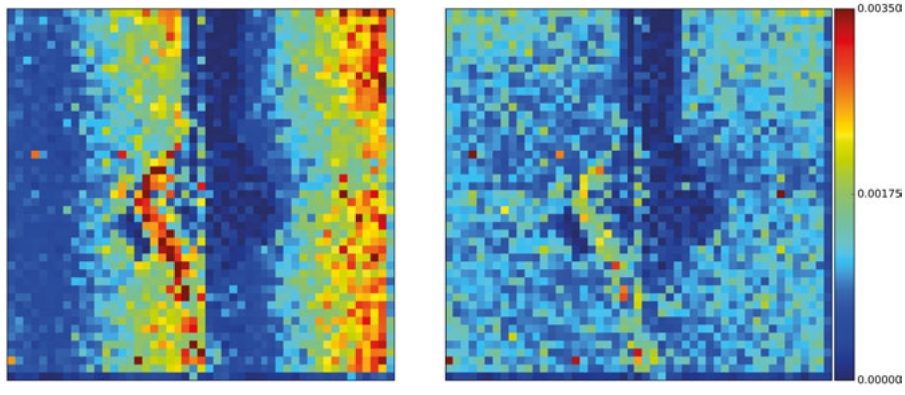

Figure 3.3. The two patterns represent a downsampled (8×8) average over over 605 dark frames. A strong artifact is visible (band in the right CSPAD-2x1 module in the left-hand panel) along with its correction (right-hand panel). Data used here come from the L730 beamtime, specifically run 398.

3.3.1 Common mode correction per-column per-ASIC

In **Paper I** we analyzed the detector noise histograms for L730 data (left plot; blue histogram in **Fig. 3.5**) and noticed a broad distribution in photon count space ($\mu = 244.10$ and $\sigma = 57.60$).

We also identified a spatially non-uniform artifact in (**Fig. 3.4**), by averaging over the sum of 2932 outliers (i.e. putative sample hits, if no artifact was expected, percentile 90–95 in photon count to avoid true hits) in the distribution of specific sample runs from L730. We first thought that the artifact was present only in sample runs, and so that it was probably related to some issues with lit pixels. Then, we realized that this artifact was also present in the dark data (**Fig. 3.3**), meaning a systematic bias was intrinsic in the CSPAD detector and not due to its interaction with incoming x-ray photons.

The structure of the artifact in both these cases suggested a *per column* offset. Those considerations led us to introduce an additional per column, per ASIC common mode correction.

Our suggested per-ASIC per-column correction consists of taking the median value of each ASIC's column (pixels from top to bottom in **Fig. 3.4**) and subtract it from the values in the given column.

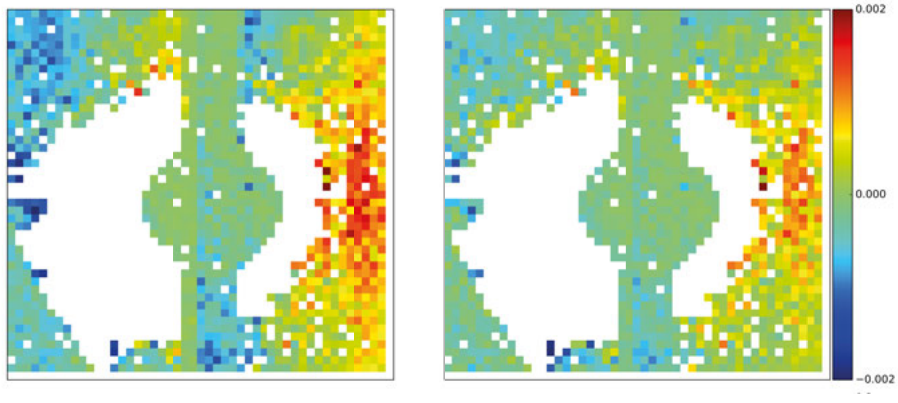

Figure 3.4. The 2 patterns represent a downsampled (8×8) average over 2932 frames. They are representative of the sample runs, where a normalized expected background is subtracted from each frame used for the average; pixels with too high photon counts have been masked out. A strong artifact is reported (band in the right CSPAD-2x1 module in the left image) along with its correction (the respective on the right image).

The outcome of the proposed correction can be seen in **Fig. 3.5**, both on L730 and L867 data. In the L730 data, the photon count distribution after correcting for the artifact in the dark frames is more well defined and the right tail is reduced a lot, as we would expect for a run with few or no actual photon events (plot to the left, red histogram).

There were appreciable improvements also in the L867 data, even though not so pronounced (plot to the right, red histogram).

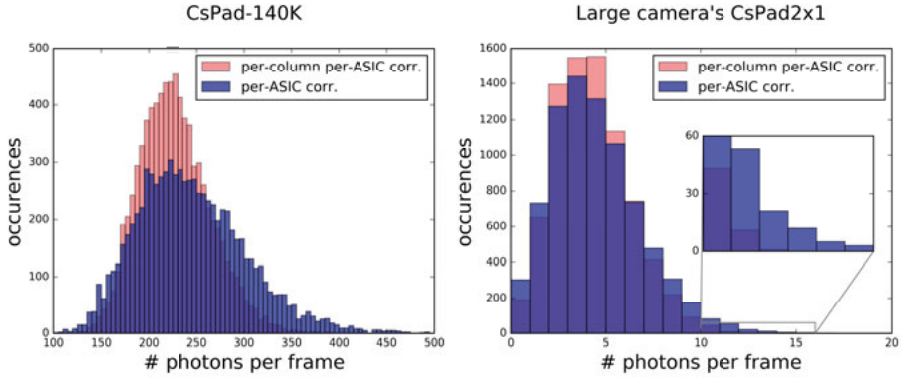

Figure 3.5. Detector noise histograms for data collected during L730 beamtime on the CSPAD-140k detector (to the left) and for data collected during L867 beamtime on one CSPAD-2x1 module in the CSPAD-2.3M (to the right). For both plots, the x-axis shows the number of photons in each dark frame (representing the detector noise); on the y-axis the number of occurrences of each value in that dataset is found. The correction for a per-ASIC offset is displayed in blue; the one for per-column per-ASIC offset is depicted in red. The inset on the right figure shows a larger tail for the per-ASIC offset correction.

In **Fig. 3.6**, we reported the ratio between the standard deviation for the corrected ADU values using a per-ASIC offset and our proposed approach with an additional per-column common mode subtraction. We can see a clear noise level decrease by employing the per-column per-ASIC common mode correction, since this ratio is above 1 for the vast majority of pixels ($> 99.4\%$ for both the detectors). Using the per-ASIC correction, the average per-pixel ADU standard deviations are 2.6252 and 3.9291, whereas with the per-column correction we obtained 2.5942 and 3.8536.

Despite an improvement of only 1.9% and 1.2% (L730 and L867) in the width of the zero-photon peak, there was a larger difference in false-positive photon reduction (**Table 3.1**).

This is not due to reducing overall photon detection power. The average offset imposed over all pixels over all frames was less than 0.0001, implying that the average threshold for photon detection went unchanged and was thus no more restrictive in the per-column offset correction mode.

The reductions in false-positive photon counts can also be represented as a decrease in the false-positive photon detection rate. Remarkably, we notice that the relative reduction of the standard deviation in the latter case is up to 35% (**Table 3.1**).

	Average per-pixel std. dev. [ADU]		False-positive photon reduction [photons × frame^{-1}]		False-positive detection rate [photons × pixel^{-1} × frame^{-1}]	
	L730	L867	L730	L867	L730	L867
Correction per-ASIC	3.9291	2.6252	$\mu = 244.10$ $\sigma = 57.60$	$\mu = 3.99$ $\sigma = 2.35$	$\mu = 1.70 \times 10^{-3}$ $\sigma = 4.00 \times 10^{-3}$	$\mu = 5.57 \times 10^{-5}$ $\sigma = 3.28 \times 10^{-5}$
Correction per-column per-ASIC	3.8536	2.5942	$\mu = 226.42$ $\sigma = 37.33$	$\mu = 3.81$ $\sigma = 1.99$	$\mu = 1.58 \times 10^{-3}$ $\sigma = 2.60 \times 10^{-4}$	$\mu = 5.32 \times 10^{-5}$ $\sigma = 2.78 \times 10^{-5}$
Relative improvement	1.2%	1.9%	-	-	$\mu_{impr.} = 7.06\%$ $\sigma_{impr.} = 35.00\%$	$\mu_{impr.} = 4.49\%$ $\sigma_{impr.} = 15.24\%$

Table 3.1. *Average per-pixel standard deviations, false-positive photon reduction and false-positive detection are reported for the two different corrections — per-ASIC and per-column per-ASIC common mode — and for the two experiments — L730 and L867.*

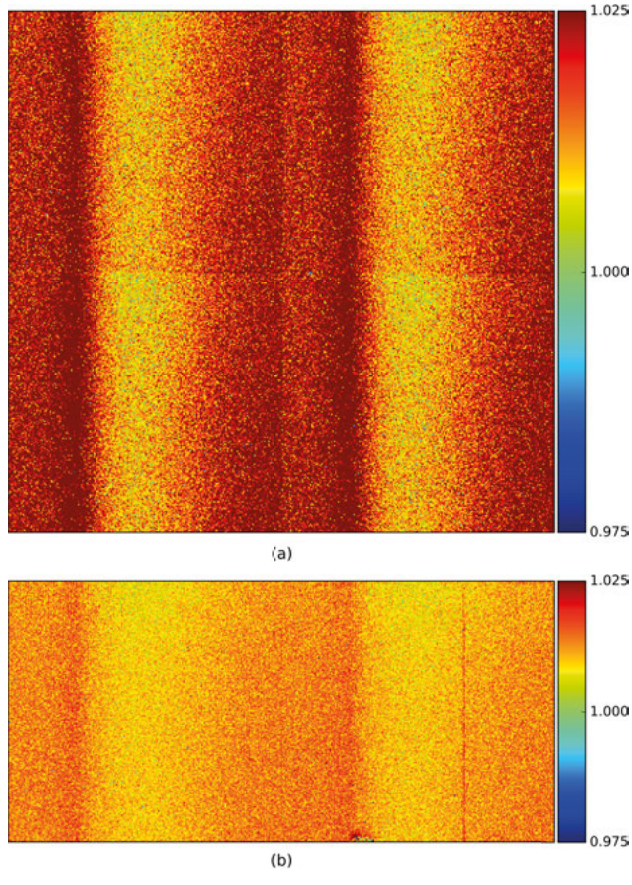

Figure 3.6. Comparison between per-ASIC and per-column per-ASIC offset correction. Value reported is the ratio between the standard deviation of corrected ADU values for the same dark run using the two methods. The vast majority of values are above 1, meaning that the per-ASIC correction generates more noisy data than per-column per-ASIC approach. (a) shows the comparison for CSPAD-140k (L730) whereas (b) shows the same for a single CSPAD-2x1 module in the CSPAD-2.3M (L867).

4. A statistical approach to detect protein complexes at X-ray Free Electron Laser facilities

In order to identify diffraction from individual macromolecular complexes when noise and scattered sample signal are comparable, we need a robust and highly sensitive hit-identification method. Current hit-finding methods, such as using arbitrary thresholds in terms of the number of lit pixels in down-sampled detector images [10][11] are still too coarse to be effective in the case of smaller biological particles. Hence, we developed a more sophisticated method (**Paper II**), which is based on photon statistics and — by modeling the background — is able to discern the latter from scattered sample signal. Our approach is essential even taking into consideration the improvements recently had at the CXI instrument with background reduction [39].

The background was reduced by introducing apodized apertures into the beam. However, this also resulted in a substantial reduction of photon flux on the sample and thus in a similar reduction of photons scattered by the sample (i.e. the actual signal usable for any data analysis).

Therefore, as the method proposed permits to know the underlying expected background of a specific pattern, we can work with a higher background level, but still gather a larger amount of sample signal, that is needed to fully reconstruct the 3D structure of the particle [40].

4.1 Statistical hit-finder implementation

Here we describe the foundations of this approach and illustrate how it works.

4.1.1 Background model

We built our model on two assumptions: i) the photon count in detector pixel k of frame i follows a Poisson distribution defined by the rate parameter λ_{ki} (whose observation is constituted by the number of photons n_{ki} — **Fig. 4.1**); ii) the detector pixels are considered to be independent from one another, once the rate parameters have been determined.

The first assumption, in a low photon emission regime like the one we are working with, constitutes a well known physical process [41]; the second one

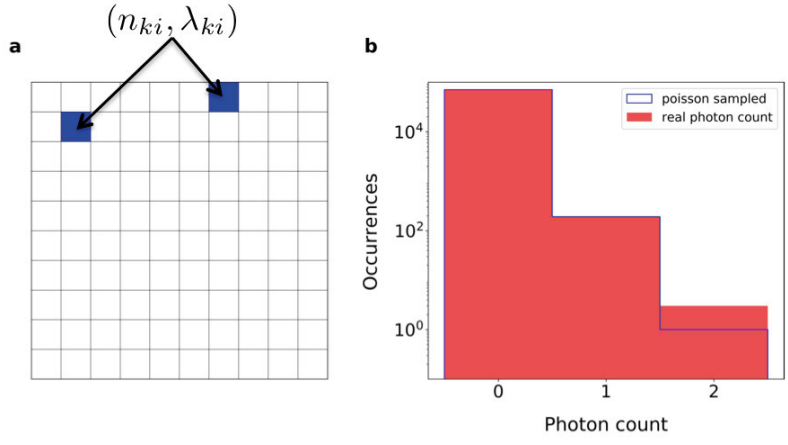

Figure 4.1. Model structure based on independent pixels adhering to Poisson statistics. a) a pixel-based detector and how each pixel in a background frame is provided with the observed photon count n_{ki} and its mean photon count λ_{ki}; b) histogram plot (semi-logscale) showing the distribution of a certain pixel in the background run collected during the June 2015 experiment, having a mean photon count equal to 0.022 (red), compared with a simulated Poisson distribution (blue empty histogram), based on the same mean photon count and the same pulse energies per event as in the real case. The agreement is evident (in the case of the 2-photon count, the histograms differ only for two occurrences), thus proving the validity of the assumption.

depends on the detector design and it is reasonable after proper pre-processing of the data for the CSPAD detectors used at the CXI end station. The values n_{ki} and λ_{ki} are determined experimentally from the data. A specific filtering step will mask out pixels where we fail to predict n_{ki} properly given our estimate of λ_{ki}.

In order to estimate the rate parameters, first the raw signal in each pattern is corrected, the gain corresponding to 1-photon is retrieved and the pattern itself is photon-converted — i.e. photon counts n_{ki} are obtained — as described in **chapter 3**. Then, as a preliminary step, the derived photon distribution is used to remove all events with very high or very low signal (as explained later in section 4.2), so that λ_{ki} can be calculated without being strongly influenced by outliers. The formula used is:

$$\lambda_{ki} = \frac{\Phi_i \sum_j n_{kj}}{\sum_j \Phi_j} \tag{4.1}$$

where $\Phi_{i(j)}$ represents the expected number of photons per frame (or *expected photon count*). This parameter was introduced to take into account the varying beam pulse energy, specific of each event. The beam pulse energy, which is specific to each event, is expected to be linearly proportional to the number of scattered photons. Instead, we found a non-linear relationship between the

recorded energy and the number of detected photons. We used a third degree polynomial to fit the trend in order to obtain a proper expected photon count for each event. This non-linearity was probably due to a calibration issue in the pulse energy detection system, hampering its linear performance across the full range of energies in our dataset.

After calculating the rate parameters, we can determine which detector pixels are working correctly or not. To do that, we start by assuming as a "good" pixel one meeting our assumption i). The set of rate parameters $\{\lambda_{k\cdot}\}$ for a pixel k is sampled according to Poisson statistics and is then compared with the set of observed photon counts $\{n_{k\cdot}\}$ for that pixel. A scalar product metric (ranging from 0 to 1) of the normalized vectors obtained from the two sets is then used as discriminator: the closer it is to 1, the more the pixel fulfill our null hypothesis.

This last step gives the advantage of automatically masking out "bad" pixels, instead of excluding them manually [10][11]. It deals with a broad set of cases: malfunctioning pixels, saturated pixels, areas of the detectors where unstable background is revealed and other possible accidents that make a pixel deviate from ordinary behaviour.

4.1.2 Score definition

Using the information above, we can calculate a single score for event i. We compare each experimental observation (i.e. the diffraction pattern photon-converted) with its expected background (i.e. the mean rate parameters for that specific pattern), by computing a log-likelihood ratio [42]:

$$s_i = \sum_k n_{ki} \log \frac{n_{ki}}{\lambda_{ki}} \tag{4.2}$$

4.1.3 Threshold definition

Under our previous stated assumptions i) and ii), we can use the central limit theorem to assume our s_i score distribution to be normal. Due to non-ideality in the data, we found it impossible to derive reasonable test thresholds based on the ideal behavior of the underlying distribution properties.

Instead, we noted that the distribution of such scores for a stable background as a function of the expected photon count for each event gives a linear relationship (**Fig. 4.2a**, upper plot). By fitting the log-likelihood scores, we obtained the average expected scores for all events in the dataset. Those fitted values (red line in the upper plot of **Fig. 4.2a**) can be seen as the log-likelihood scores of the expected background in each event. Then, after subtracting these values from our scores (bottom plot in **Fig. 4.2a**), the resulting distribution is independent of the pulse energy. For this new transformed distribution, we can express our hit threshold as $\mu + 4\sigma$ (μ being the mean of the new distribution

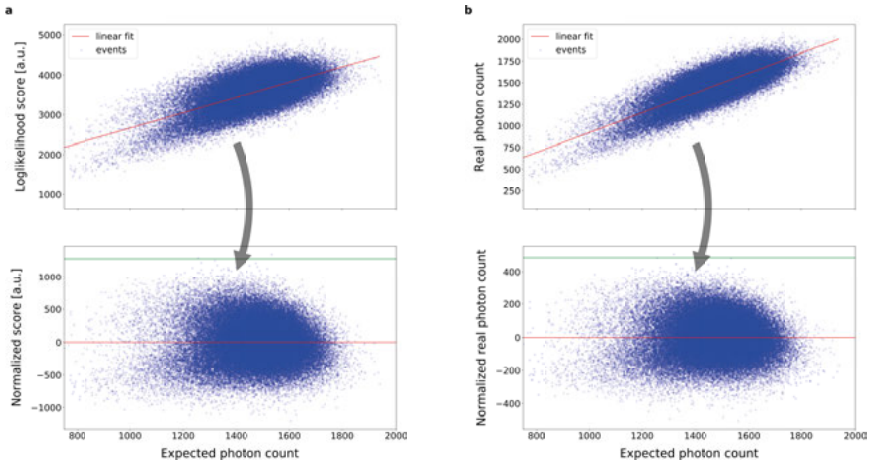

Figure 4.2. a) and b) show respectively the original log-likelihood scores and real photon counts and their transformed counterparts for a background run (from the L867 experiment, June 2015). Blue circles represent all events in the dataset, whereas the red line is a linear fit of the log-likelihood scores (a) and real photon count (b) to expected photon counts. The lower plots are obtained by subtracting the offset from the data. The resulting normalized distribution is — for both a) and b) — roughly Gaussian, centered at $\mu = 0$, making a hit-finding threshold of $\mu + 4\sigma$ straightforward to define (green line). While overall distributions look similar, the score-based method in fact provides tighter bounds.

and σ the corresponding standard deviation). The choice of a 4σ threshold ensures a theoretical false positive rate of 3.16×10^{-5}. That theoretical value is consistent with the experimental one found for background data from a June 2015 experiment [39]: 2.83×10^{-5} (**Fig. 4.3**).

4.2 Model refinement

During experiments, significant changes are frequently made between experimental runs, without recording new background data, so that characterizing μ and σ is not always trivial. To account for those variations, we can instead use events in the sample runs that are unlikely to be hits. We call them *preliminary misses*.

4.2.1 Preliminary misses

Preliminary misses are defined on the basis of the photon count distribution inside individual chronologically ordered bins of multiple events. In order to calculate stable estimates of the means $\mu_{\{\text{bin}\}}$ and standard deviations $\sigma_{\{\text{bin}\}}$, the bins are constructed so that each contains at least 100 events.

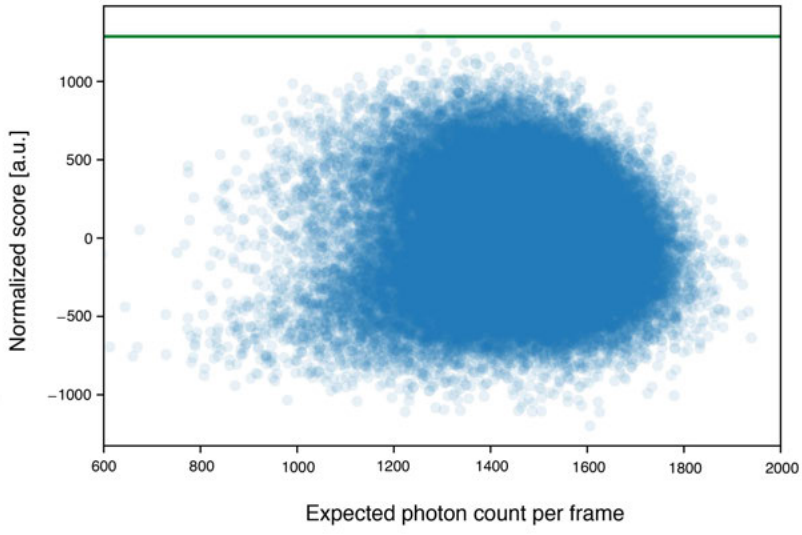

Figure 4.3. The plot of normalized score vs expected photon counts for background — run 156 L867 — shows 2 outliers over 70613 events (giving a false positive rate of $\sim 0.0028\%$). Blue circles represent the events filtered for pulse energy>1 and photon count>400; the green line represents the 4σ threshold.

The photon counts in each bin are fitted to the pulse energy values to satisfy a third order polynomial relation. These fitted values (representing the expected photon counts of the events) are subtracted from the original photon counts, in order to make the distribution energy independent.

Then, the mean and standard deviation of the bin ($\mu_{\{\text{bin}\}}$ and $\sigma_{\{\text{bin}\}}$) are calculated and the elements of the bin are selected in the interval ($\mu_{\{\text{bin}\}} - 4\sigma_{\{\text{bin}\}}$, $\mu_{\{\text{bin}\}} + 4\sigma_{\{\text{bin}\}}$), calculated with the method of moments [43], which works well as long as the true background distribution is Gaussian. This operation is performed iteratively, reworking the fit of all parameters based on the current set of shots within the range, as long as the bin contains more than 100 elements or until $\sigma_{\{\text{bin}\}}$ does not change anymore.

The preliminary misses based on this photon count criteria, as well as the constraints (on pulse energy and photon count) to exclude blank outliers per bin, are then combined to form the total set of preliminary misses. These are then used as the background events in our approach.

As long as the background is Gaussian and the data is dominated by non-hits (experimental hit-rates up to 10%), the method, even if coarse, is still reliable.

This preliminary filtering is particularly important because it ensures that the background statistics is not affected by very strong or very weak events when calculating the rate parameters (λ_{ki}). In such a way, we avoid to reduce the detection power in those areas of the detector where the background signal is very clean. It also means that the image we get when subtracting the background should be representative of actual scattering.

Moreover, this filter allows an online mode [44] for our hit-finding methodology, by permitting a real time adjustment of the background model and consequently of the threshold, while data are being collected.

4.2.2 Identifying a relevant background

Ideally, in the case of perfect background stability, a background run could be used as the basis for the background model for the following sample runs, and a threshold could be defined on it. That may be true for the first few sample runs collected immediately after the background run. For sample runs collected later in time, we found that using a previously acquired background is detrimental. The background changes both due to continuous drift and active changes done in order to tune the experimental parameters being executed during and between runs.

In the case of a stable background — where we assume a variations $< 10\%$ — we can deal with sample runs as they were a single one, and apply the approach proposed. Instead, if the background varies a lot during run acquisition, each run must be treated separately.

Usually, typical background does not vary very much within a single run. But, if it were ever the case, one could still consider to split the run itself in smaller chunks containing a reasonable number of events, or to discard background outlier events in a preliminary filtering.

4.3 RNA polymerase II as an application

The proposed approach was applied to diffraction data of RNA polymerase II and the results achieved were compared with a hit-finder based on time-of-flight (ToF) detector ion spectra (see following section), in order to test its reliability in detecting hits. RNA polymerase II is an enzyme involved in DNA transcription [45]. It is the first protein complex ever injected at an XFEL. The beam focal spot size was nominally 100 nm [46] and the photon energy 6 keV. The sample buffer used consisted of water and ammonium acetate; the sample itself was labeled with gold spherical nanoparticles to increase its scattering power.

4.3.1 Time of Flight detector (ToF)

We used a ToF detector as an independent hit-finder to validate our proof of concept method. A Multi-Channel Plate (MCP) detector was deployed, placed at a distance of $\sim$50 cm from the interaction region. It was used in *drift mode* (i.e. ions were not accelerated by a potential field in the direction of the detector applied across the interaction point). Thus, the recorded flight times reflect only the kinetic energy gained by the ions from the explosion of the sample particle. In our analysis a ToF event is considered to come from sample if more than one single-proton (2 mV) signal is detected (see **Fig. 4.6** for examples of typical ToF traces).

4.3.2 Hit-identification done with the statistical hit-finder

Fig. 4.4a illustrates how to perform hit-identification with our statistical approach. The density plot represents all the events collected, where the densest region is darker — and describes the background — whereas less dense regions are lighter. A threshold (green line) is defined as described in section 4.1.3, and the subset of events identified as hits by our approach (blue circle outlines) are reported. The red circles represent the hits found via the independent time-of-flight ion hit finder (see section 4.3.1).

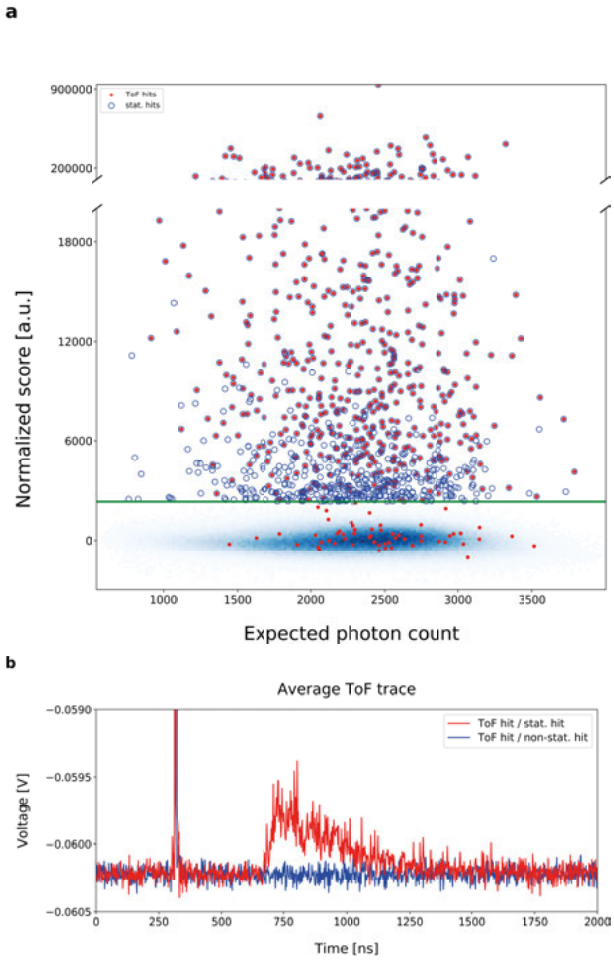

Figure 4.4. Summary of comparison of statistical hit-finder against ion-based hit identification using a time-of-flight detector. (a) shows a density plot (darker blue implying denser regions) of the normalized score distribution of all shots. The green line is the threshold defined by our statistical hit-finder, and whatever is above is considered a hit (blue circle outlines). ToF hits are also shown (red). (b) shows the average of 56 ToF non-blank traces that are not statistical hits (blue line) and a corresponding number of ToF hits that are statistical hits as well (red line). In the latter, proton peak is visible around 800 ns.

When a 4σ outlier threshold is applied, background events are clearly separated from hits. There were 1,165 hits of varying strength over a total of 402,296 non-blank events considered over 25 runs (418,153 events in total). Furthermore, 828 hits were identified using the ToF detector.

Thus, by noticing that most hits are shared between the ToF hit-finder (red circles) and our statistical model (blue circle outlines), we can state that our hits are not spurious. The total fraction of ToF hits that are above our defined threshold is 94% (771 ToF hits); the remaining 57 belong to background. For

these shots, the proton peak is also visually absent when considering the average integrated ToF trace as well as the individual traces.

It seems likely that all or most of these are in fact false positives by the ToF hitfinding algorithm. We observe that 57 events out of 402,296 amounts to $\sim 0.013\%$, which is consistent with the expected false positive rate for the ToF method of approximately $\sim 0.01\%$ previously reported [47]. It is worth noting the distribution of these non-matching ToF hits follows the overall distribution of recorded events in terms of the loglikelihood scores (**Fig. 4.5**). If they were instead very weak hits not picked up by our approach, one would still expect them to cluster between the main background "cloud" and the threshold.

It is not surprising that the statistical hit-finder recovered a higher number of total hits, since the ToF ion detector experimental geometry covered a small portion of the total solid angle. Therefore, only a limited fraction of the ions emanating from sample explosions could be picked up.

Unfortunately, we could not unambiguously attribute any of the hits found with the two different methods to a single RNA polymerase II complex. However, when no injection was performed, or only buffer was injected, the corresponding hit count was very low. This indicates that the hits detected are arising specifically from the sample solution.

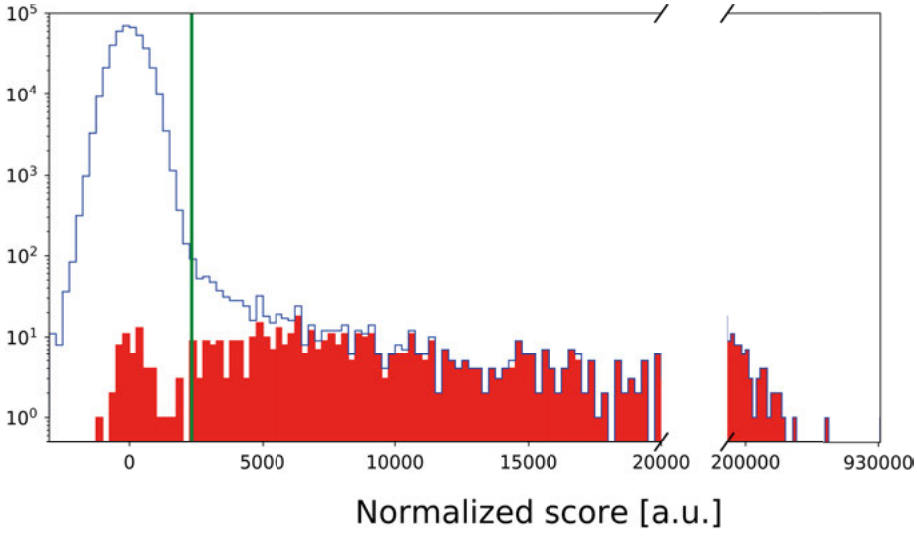

Figure 4.5. Histogram of all events (blue unfilled) and of the ToF hits (red filled). The green vertical line represents the threshold according to our statistical hit-finder. ToF hits below our threshold roughly match the overall background distribution, indicating that most of those are false positives. The y-axis is consistently showing the number of counts, but the bin size is different in order to guarantee a better visibility of the second part.

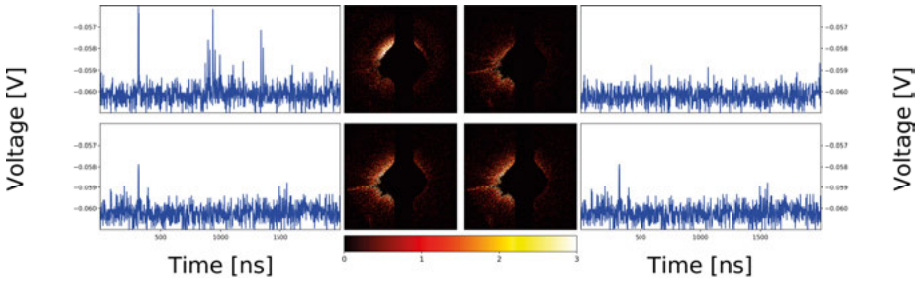

Figure 4.6. Four representative events, with recorded x-ray photons as well as ToF traces. The top-left event is a hit (as evident from the proton peak in the 500 - 1000 ns window); the top-right represents an event revealed as hit by the ToF but not by our statistical hit-finder; viceversa, the bottom-left is an event revealed as hit by our statistical hit-finder but not by the ToF; the bottom-right represents a background event for both hit-finders.

4.4 Statistical hit-finder efficiency

We also tested our statistical approach on other datasets collected at the CXI end station during experiments in April 2014 and April 2016. The experimental setup was the same as described for the RNA polymerase II experiment; the CSPAD detectors shared the same revision (v. 1.6) as on May 2013.

To reduce the total amount of incident photon flux (and so the scattered background), more aggressive aperturing of the beam was applied, reducing the background scattering significantly, but also decreasing the scattered signal from the samples.

Besides, we also performed computer simulations on protein-like hits, to show the reliability of our approach.

4.4.1 Results on larger biological particles: Omono River virus and bacteriophage PR772

Omono River virus (OmRV) and PR772 are two icosahedral viruses, larger in size than RNA polymerase II: the first is 40 nm in diameter and the second is 70 nm in diameter. OmRV was injected as described in a previous work [10]. In the PR772 dataset analyzed, the specific run used was collected while the injection system was flushed with water, creating a slow elution of remaining sample particles towards the end.

We found 870 hits for OmRV and 460 hits for PR772 (**Fig. 4.7a,b**), meaning respectively 4.29% and 0.28% hit-rate.

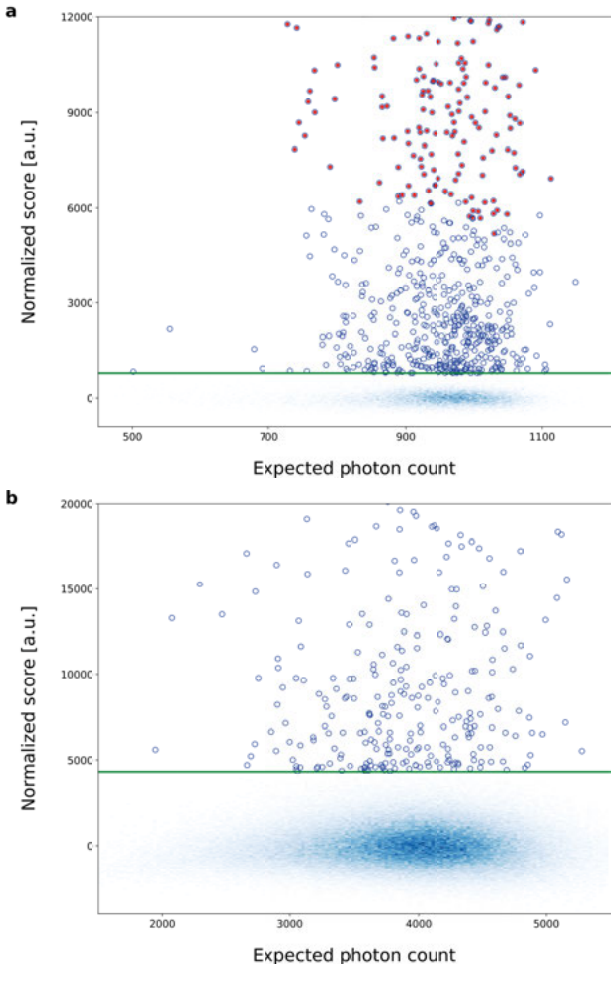

Figure 4.7. Results from the statistical hit-finder applied to OmRV (a) and PR772 data (b). Both plots show the density (darker blue implying higher density) of the normalized score distribution of all shots in the dataset. The green line is the threshold defined by our statistical hit-finder; the blue circle outlines are the identified hits (870 and 460 hits, respectively). In addition (a) shows the 421 hits identified in a previous reported analysis. All those hits were also identified by our hit finder.

Fig. 4.8a,b show icosahedral patterns for some of these hits, with representative single particle shots for the OmRV virus and bacteriophage PR772, respectively. The snapshots show particles of different sizes, as the size distribution was quite broad for both viruses with the injection system used [10][48].

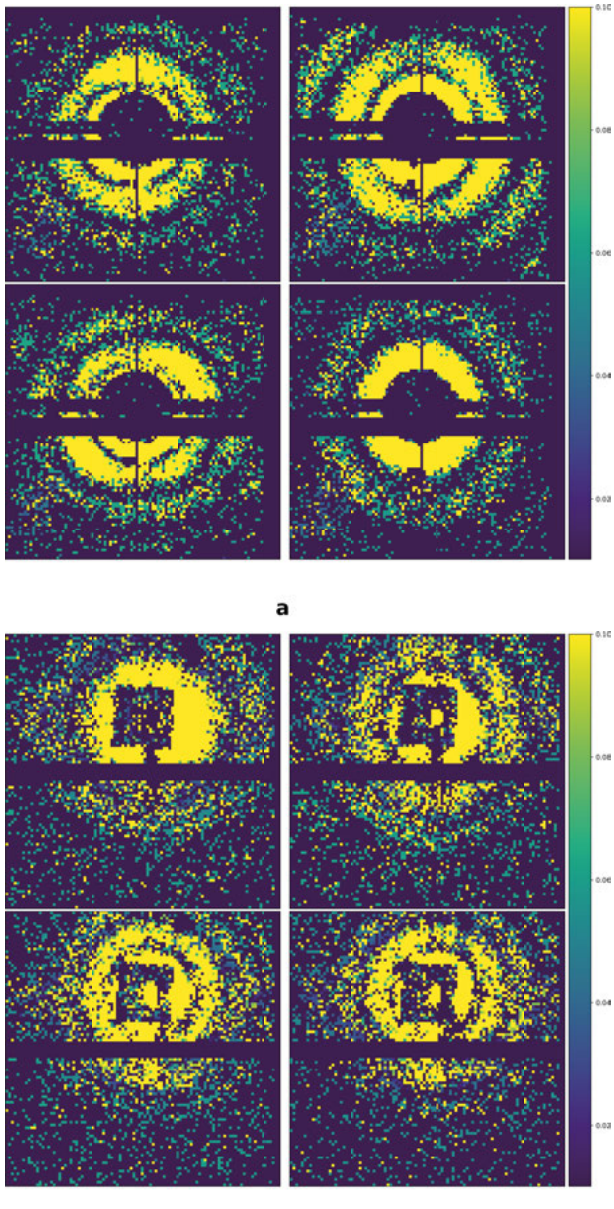

Figure 4.8. (a) and (b) show respectively 4 single OmRV and PR772 hits (down-sampled at 4×4 pixels). These patterns represent hits of different sizes, as the size distribution for those viruses is quite broad using the injection methods in place at the time.

In the previously published analysis on the OmRV dataset [10], 421 hits were found in the same run we analyzed. Our set of 870 hits was a strict super-set, thus including all the 421 hits previously identified. Moreover, we show

that some of the additional hits are clearly representative of single Omono River virus particles, albeit weaker (**Fig. 4.9**).

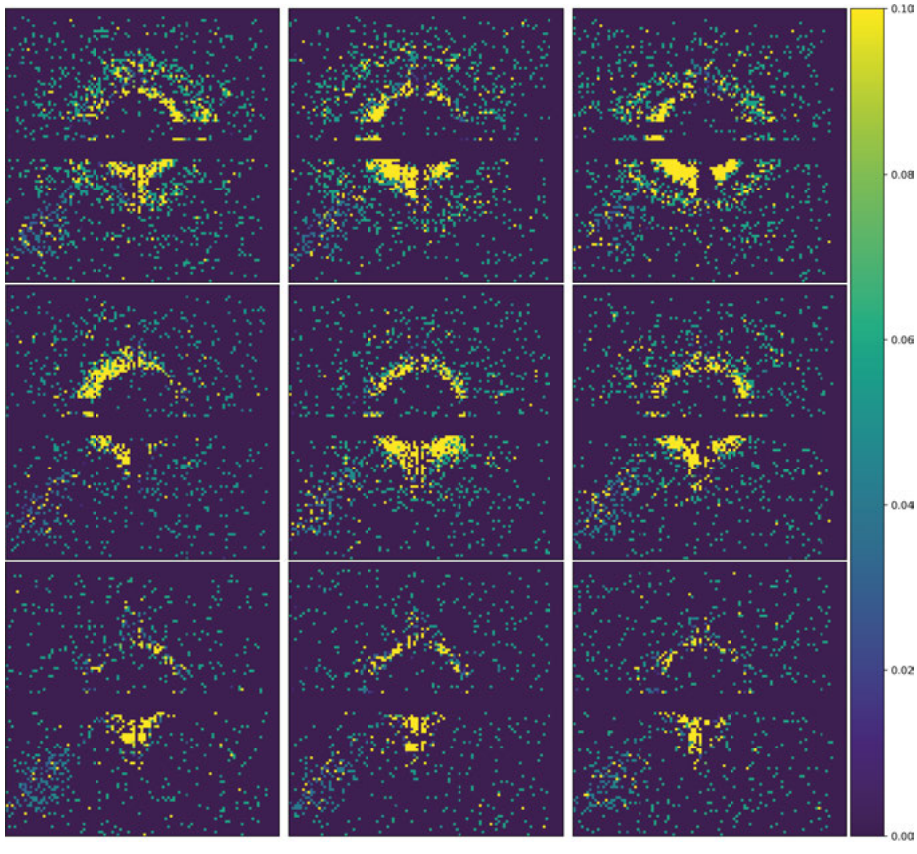

Figure 4.9. Downsampled (4×4 pixels) patterns belonging to the OmRV dataset. The first 2 rows show events identified as hits by our hit-finder — in the score range 4000-6000 (first row) and in the range 2000-4000 (second row) — that are not identified by a simpler hit-identification scheme (such as the one used by the Cheetah package); the third row shows 3 blank events (belonging to background), to be used as a reference to the eye for discerning hits/misses. In particular, the first row shows that our hit-finder can find particle hits from a single OmRV virus that were excluded by a standard hit-finder.

4.4.2 Protein hits simulated on top of true experimental background

Computer simulations of focus-centered and off-center hits of spherical particles with protein-like scattering power (diameter size 8, 13, and 40 nm) were performed. The results using our statistical hit-finding method were compared

against a more traditional approach based on overall photon counts, with the threshold defined as in **Fig. 4.2b**, are shown in **Fig. 4.10**.

The ratio of correctly identified hits given varying photon beam intensities is reported. The focus-centered spherical hits of protein material were simulated using Condor online [49]. A range of intensities (4.46×10^9– 4.46×10^{18} photons $\times$ pulse$^{-1} \times \mu$m^{-2}) has been explored for the three sizes. These were superimposed on a background run recently collected on the CSPAD-140k detector (from the June 2015 experiment, L867). We simulated particles hit by an off-center Lorentzian and Gaussian beam, to better represent that in a real FXI experiment hits are rarely perfectly focused relative to the x-ray pulse.

Examining **Fig. 4.10**, we can see that perfect efficiency for particle hits perfectly in focus by a tophat beam (straight and dash-dotted lines), for 40 nm particles, is reached at 10^{10} photons $\times$ pulse$^{-1} \times \mu$m^{-2}; for 13 nm particles, 50% and 100% correct hit-identification are reached at 5×10^{11} and 10^{12} photons $\times$ pulse$^{-1} \times \mu$m^{-2}, respectively.

Instead, in the case of non-centered hits, we see that the efficiency of the approach at 10^{12} photons $\times$ pulse$^{-1} \times \mu$m^{-2} is lower than 50% for all the cases.

By increasing beam intensity to $> 10^{13}$ photons $\times$ pulse$^{-1} \times \mu$m^{-2}, the goal of revealing hits of diameter size $\sim$13 nm or smaller is doable ($> 50\%$ of correct identifications with our statistical approach).

A comparison of the two hit-finding methods shows that for hits of 13 nm the statistical one gives a 50% recovery rate at half of the intensity: i.e. the statistical hit-finder and the simpler pure photon count hit-finder give a similar result respectively at 9.73×10^{13} and 2.12×10^{14} photons $\times$ pulse$^{-1} \times \mu$m^{-2}.

At the intensity where the statistical hit-finder recovers 50% of hits, SNR $= -7.86$ dB (for the median shot). On the other hand, where the pure photon count hit-finder recovers 50% of hits, SNR $= -5.27$ dB. That means that, even though the intensities needed to reliably detect an 8 nm particle are currently unachievable at the CXI instrument [10], our statistical approach can still operate with less photon flux while achieving an equivalent identification rate.

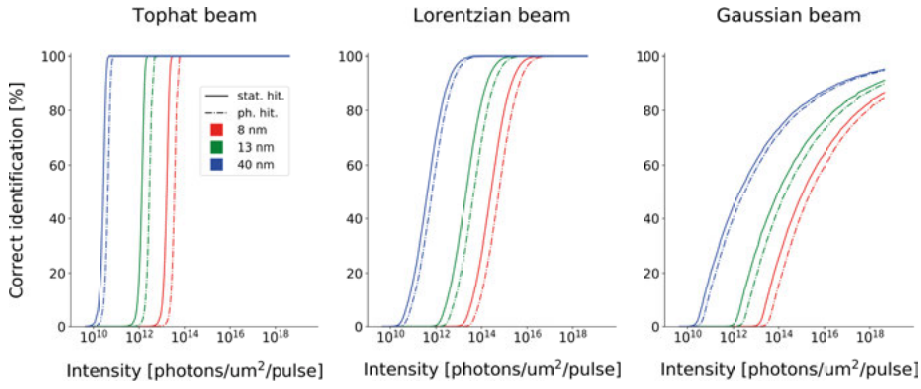

Figure 4.10. Hit-finding performance of simulated hits superimposed on real background, for varying pulse intensities, particle sizes, focal properties, and hit-finding methods. Three representative spherical particle sizes (8, 13, 40 nm) were simulated. We present our hit-finding method and a simpler scheme using our derived pixel mask and the total expected photon count given the pulse energy. Hit-finding was also performed in three distinct beam settings: particles hit perfectly on focus by a tophat beam (on the left) and the more realistic cases of a Lorentzian and a Gaussian beam hitting the particle (center and right). In all the plots, normalized scores are shown as a solid line, whereas photon count based detection is dash-dotted. For the most interesting cases of a 13 nm particle hit by a Lorentzian and Gaussian beam, we found a 50% recovery respectively at intensities of 2.10×10^{13} photons $\times$ pulse$^{-1} \times \mu$m^{-2} at and 9.73×10^{13} photons $\times$ pulse$^{-1} \times \mu$m^{-2} with our statistical hit-finder; at 2.12×10^{14} photons $\times$ pulse$^{-1} \times \mu$m^{-2} with the simpler photon count version.

5. Statistical hit-finder software

In order to develop the method presented in the previous chapter and the underlying algorithms, we needed to write some code and test it. Besides, dealing with the amount of data we collected implied the need to implement a distributed parallel approach.

Here we explain how we analyzed the data collected at the CXI end station, focusing on he actual processing steps, their order, and practical implementation.

5.1 Computational environment

Before diving into the algorithms implementation, we show the framework used for the data analysis. The data analysis reported was implemented in Python and run on a cluster private to the Uppsala LMB group (having a total of 54 compute nodes — 24 logical CPUs with Hyperthreading per node — with 64 GB of memory each). To handle an amount of data spanning from 10-120 GB (meaning between 8000 and 200000 events) just for a single run, a variable number of nodes (1-8) and worker processes (10-50) was used in parallel, taking advantage of the *mpi4py* and the *h5py* modules. The latter was built with MPI support (instructions in `http://docs.h5py.org/en/latest/mpi.html`) in order for the workers to read/write from a same HDF5 file simultaneously. As some steps of the algorithm need to process and store all the data in the datasets, we chose to read/write the partial results to HDF5 files, even though it slows down job's execution, so to have control over the final results of each step and to avoid to exceed the amount of memory available. The SLURM workload manager [50] was used to manage jobs running on the cluster.

5.2 Preprocessing

The very first step is to preprocess the data into a generally accessible form. By using the Cheetah software [51] (or, in some cases, Hummingbird [44]) we can convert the raw data, collected by the detectors and stored in the XTC format, into the CXIDB format [52] (a specific HDF5-based schema for Coherent X-ray Imaging data). In our analysis, most preprocessing steps in Cheetah (or Hummingbird) were disabled and we carried out our own, as explained in the following sections.

5.3 A step-by-step script

The software performs all the necessary steps described in **Paper I** and **Paper II** to obtain the non-normalized loglikelihood scores (section 4.1.2). The code is available at `https://github.com/albpi/Statisticalhitfinder/`.

Those steps are executed by the main function of the script (see code below) and are described in the following seven subsections. The normalized values are obtained in a later step. Its execution can be found in a Juppyter notebook in the same repository.

The six main steps can be run either altogether or separately (if one of the steps has already been executed before) by commenting out one of them.

```
def main():
    """If one of the steps has already been executed and there
        is no need to re-run it, just comment it out"""
    #dark_mode()
    #common_mode()
    #gain()
    #photon_count()
    #lambda_values()
    #poissmask()
    photon_count()
    lambda_values()
    baglivo_score()
```

A simple configuration file, which reads the name of the experiment and the names of the raw data files to read and process, is read at the beginning by all the workers.

```
# FILE TO GENERATE .h5 in the main_analysis.py

[exp_details]
exp = cxi73013
dark_runnr = 3
sample_runnr = 78
```

Sample data (including raw frames, photon counts and hit-identification scores) have been deposited into the CXIDB repository (entry 78).

5.3.1 Pedestals

The offset given by environmental and detector noise is subtracted from the raw data. The mode value of each pixel for 5000 events in a dark run is taken.

5.3.2 Common mode correction per-column per-ASIC

As the detector is constituted by independent ASICs [27] [28], a first per-ASIC offset, calculated as the median value of each ASIC, is subtracted from all the pixels of the specific ASIC. On top of that, as we found an artifact with a column structure (see **chapter 3**) even after that per-ASIC subtraction, we added a subtraction of a per-column offset (calculated as explained in section 3.1.2).

To perform this step (and step 4, 5, 7) each worker operates *framewise* on the dataset (**Fig. 5.1**, in cyan). That means that if we have 10000 events — each event constituted by a 2D matrix of 370×388 integers — the program has to store in memory $370 \times 388 \times 10000$ elements and has to work on each of the 10000 2D matrices. In order to avoid memory overload, we split the task among multiple workers. For instance, if we are using 10 workers the script is going to divide the workload so that each worker operates on a smaller chunk of $370 \times 388 \times 1000$ elements.

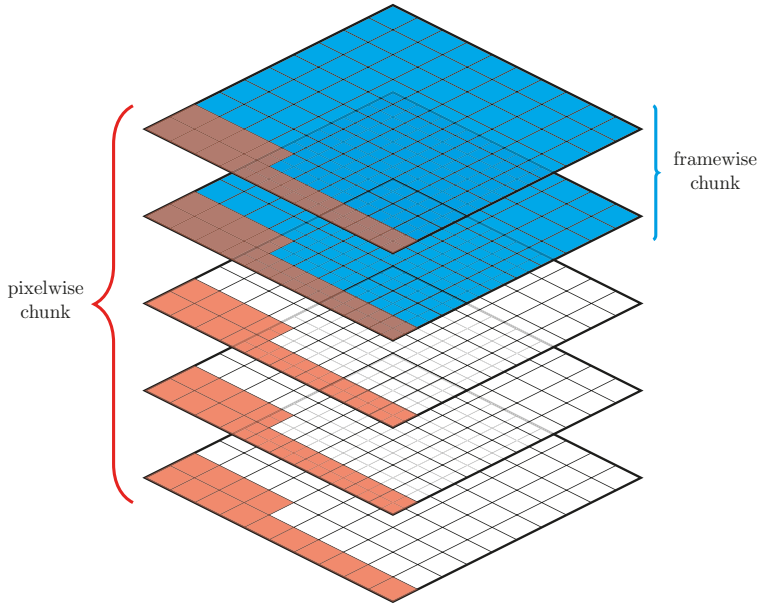

Figure 5.1. Schematic representation of a framewise and pixelwise "chuncking". In the first case, just a subset of all frames in a run is given to compute to each worker (cyan); in the latter instead, a same subset of all detector pixels — each containing all the values stored in a run — is given to each worker (red).

5.3.3 Detector gain calculation

The histograms of the values recorded by each pixel are collected, and the 1-photon ADU gains are retrieved, by fitting each histogram with Gaussian curves.

To perform this step (along with step 6), each worker operates instead on all the values recorded by a same subset of pixels (*pixelwise*), by reading what stored in the HDF5 file created in step 2 (**Fig. 5.1**, in red). Continuing on the previous example, each worker operates on a subset of 14356 pixels, with data from each of the 10000 frames for each pixel.

5.3.4 Photon conversion

The values obtained in step 2 are divided by the gains and rounded to the closest integer. Occasional negative values are set to 0.

5.3.5 Mean photon count calculation

The mean rate parameters are calculated as explained in section 4.1.1.

5.3.6 Pixel mask

Based on the data from steps 4 and 5, and assuming i) in section 4.1.1, we mask out the “bad” pixels.

After this step, steps 4 and 5 are re-run to refine the background model excluding the "bad" pixels from the calculations.

5.3.7 Log-likelihood scores

The log-likelihood scores are computed based on eq. (4.2) and stored for each event.

5.4 Computational time

Obtaining the log-likelihood scores (step 7) for the RNA polymerase II sample runs (418153 frames in total — 370×388 pixels per frame) takes about 20000 seconds on 80 CPUs. This level of performance is adequate for online as well as off-shift operation during a beamtime, given proper adaptations. For consistent online operation, however, a reasonable pixel mask and a detector gainmap for the specific beam energy are required. The mean photon counts per pixel per frame can be estimated online from incoming samples.

To further speed up the data analysis, we could use a compiled language (such as C/C++) and/or try to run the script here described on GPUs. That will substantially reduce the computational time currently required.

6. COACS – Convex Optimization of Autocorrelation with Constrained Support

To achieve the ultimate goal of imaging an object, we need to know its scattered wave function, that can be expressed as:

$$f(x) = |f(x)|e^{i\theta(x)} \tag{6.1}$$

The only physical observables that we can measure are the amplitudes of the object in the Fourier space, which are related to eq. (6.1) via:

$$F(u) = |F(u)|e^{i\phi(u)} = \mathscr{F}\{f(x)\} \tag{6.2}$$

As the detectors record only the intensities $I = |F(u)|^2$, to be able to reconstruct the object, we need to find the phases of the wave function in eq. (6.2) — or equivalently of eq. (6.1). That is called the *phase retrieval* problem and needs to be solved.

6.1 Phase retrieval

The most commonly used phasing algorithms are the so called alternating projection algorithms. We will now examine one of the simplest and look at its main features.

6.1.1 Error Reduction

The error reduction (ER) algorithm consists of four steps: 1) Fourier transform of an estimate of the area occupied by the object (also called *support*); 2) replace the current Fourier intensities with the intensities recorded to obtain a better estimate of this transform; 3) inverse Fourier transform this new transform in real space; 4) replace the modulus of the computed image with the object modulus (measured or known *a priori*) to form a new estimate of the object [14].

$$G_k(u) = |G_k(u)|e^{i\hat{\phi}_k(u)} = \mathscr{F}\{g_k(x)\} \tag{6.3}$$

$$G'_k(u) = |F(u)|e^{i\hat{\phi}_k(u)} \tag{6.4}$$

$$g'_k(x) = |g'_k(x)|e^{i\hat{\theta}'_k(x)} = \mathscr{F}^{-1}\{G'_k(u)\} \tag{6.5}$$

$$g_{k+1}(x) = |f(x)|e^{i\hat{\theta}_{k+1}(x)} = |f(x)|e^{i\hat{\theta}'_k(x)} \tag{6.6}$$

where g_k, $\hat{\theta}_k$, G'_k, $\hat{\phi}_k$, are estimates of f, θ, F and ϕ.

This algorithm works for any problem in which some constraints are known both in the Fourier and real space domain. It is an iterative algorithm that aims to find the optimum solution for the computed Fourier transform which satisfies the Fourier-domain constraints (or viceversa, for the computed image which satisfies the object-domain constraint).

The convergence of the algorithm is quite slow and can be proved by computing the squared error [14]:

$$E^2_{F_k} = \sum_u \frac{(|G_k(u)| - |F(u)|)^2}{N^2} \tag{6.7}$$

6.1.2 Non-convex problems

The ER algorithm (as well as all the other iterative methods for phasing [15]) consists in solving a non-convex optimization problem (**Fig. 6.1**). Thus, the algorithm, when looking for the global minimum, can stagnate in local minima instead of the global minimum, so not returning the optimal solution. The hybrid input-output (HIO), Relaxed Averaged Alternating Reflections (RAAR) and other methods try to avoid stagnation by escaping the local minima and continuing the search for the global. However, that does not guarantee the global minimum is found.

6.2 COACS

We here present a method that promises to improve phase retrieval results: COACS (Convex Optimization of Autocorrelation with Constrained Support).

All the alternating projection algorithms assume that the intensities used in the Fourier constraint are exact. This is not true, as the intensities that we see on the detector (discrete and finite) reflect just an approximation of the amplitudes of the object wave function (which is instead infinite and continuous). The proposed model considers the probabilistic nature of the relationship between the observed individual photons and the underlying intensities. The model can be used to calculate the most likely intensities for a certain diffraction pattern, possibly taking into account the background model developed in section 4.1. Guessing the most correct intensities and using it as Fourier constraint for one of the phase retrieval algorithms is especially important in the case of sparse data.

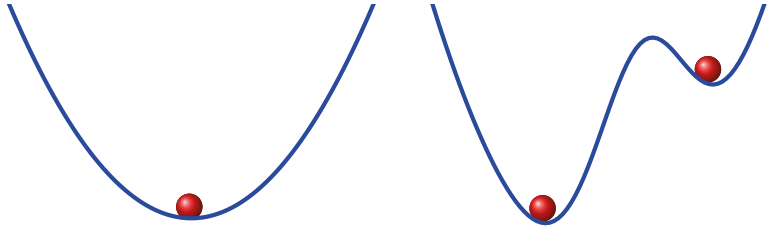

Figure 6.1. A convex problem presents only a global minimum and so a single possible solution for an optimization problem. On the other hand, a non-convex problem has multiple minima. In the case of iterative algorithms, that can cause stagnation of the solution in a local minimum and makes it more difficult to reach the global minimum (and so the real solution).

6.2.1 COACS theory

If we consider the ideal conditions of a plane incident wave, a thin object to be imaged, far field regime and low-angle scattering, we can write the original scattered wave $\mathbf{X}$ of our real-space image $\mathbf{P}$ as:

$$\mathbf{X} = \mathscr{F}\{\mathbf{P}\} \tag{6.8}$$

where here $\mathscr{F}$ is the discrete Fourier transform and $\mathbf{P}$ and $\mathbf{X}$ are the 2D center-cropped discretization of their continuous counterparts. The scattered wave $\mathbf{X}$ and its real-space image $\mathbf{P}$ have been cropped to account for the finite extension and resolution of the detector.

If we then assume a detector with uniform quantum efficiency r, we can then write a Poisson sampled diffraction pattern as:

$$\mathbf{B}_{ij} = \text{Poisson}(r\mathbf{X}_{ij}\overline{\mathbf{X}}_{ij}) = \text{Poisson}(r|\mathbf{X}_{ij}|^2) \tag{6.9}$$

This equation expresses a constraint on the resulting pattern (the *Fourier constraint*).

Since in FXI the object is imaged in isolation, it will have a compact support $\mathbf{S}$, so that $\mathbf{P}_{\mathbf{S}^C} = 0$ (where $\mathbf{S}^C$ is the complementary of the support $\mathbf{S}$). Thus,

we have an additional constraint, this time on the object (*support constraint*) that can be expressed as:

$$\mathbf{P} \odot (1 - \mathbf{S}) = 0 \tag{6.10}$$

where $\odot$ represents the elementwise Hadamard product.

The two constraints — eq. (6.9) and eq. (6.10) — can then be written as a non-linear equation system:

$$\begin{cases} r\mathbf{X} \odot \overline{\mathbf{X}} = \mathbf{B} \\ \mathbf{P} \odot (1 - \mathbf{S}) = 0 \end{cases} \tag{6.11}$$

If we call $\mathbf{Y} = \mathbf{X} * \overline{\mathbf{X}}$ (where $*$ is the convolution) the Patterson function of $\mathbf{X}$, $\hat{\mathbf{P}} = \mathbf{P} * \mathbf{P}$ and $\hat{\mathbf{S}} = \mathbf{1}_{(\mathbf{S}*\mathbf{S})}$ the support of the autocorrelation (i.e. the autocorrelation of the original support $\mathbf{S}$), we can rewrite eq. (6.11) in terms of the autocorrelation function as:

$$\begin{cases} r\mathbf{Y} = \mathbf{B} \\ \hat{\mathbf{P}} \odot (1 - \hat{\mathbf{S}}) = 0 \end{cases} \tag{6.12}$$

Eq. (6.12), along with a maximum likelihood interpretation of the true intensities for $\mathbf{B}$, are the core ideas of COACS.

6.2.2 Apodization

The discrete Fourier transform used in eq. (6.8) assumes that the underlying object is periodic. The angles to which the Fourier approximation of diffraction is valid are finite, but the mathematical model itself is based on the continuous, unbounded Fourier transform, that is infinite in space. **Fig. 6.2** shows the effects of discretizing this idealized diffraction pattern of a simulated test particle. Small, but non-negligible artifacts at the edge of the image are present when the sampled diffraction signal back into the object space. In order to eliminate these artifacts and their resulting violation of the compact support assumption in phase retrieval methods, we extend eq. (6.8) with the Hann window $\mathbf{W}$ for our square diffraction pattern with side L.

$$\mathbf{W}_{ij} = 0.25(1 - \cos 2\pi \frac{i}{L})(1 - \cos 2\pi \frac{j}{L}) \quad 0 \leq i < L, 0 \leq j < L \tag{6.13}$$

$$\mathbf{X} = \mathscr{F}(\mathbf{P} \odot \mathbf{W}) \tag{6.14}$$

Thanks to this windowing the high-frequency content smoothly goes to zero, corresponding to the lack of further high-frequency information beyond the edge of the detector.

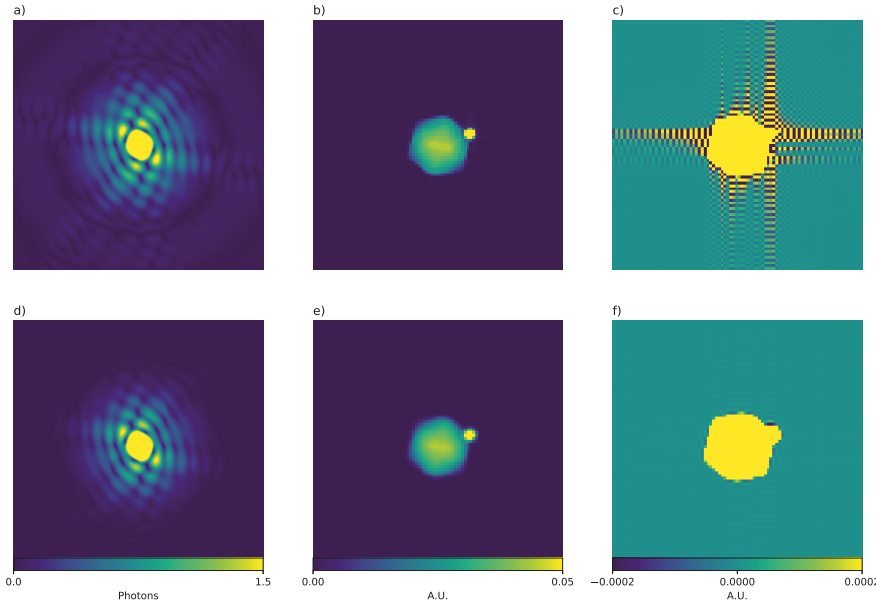

Figure 6.2. Simulated particle with and without Hann window apodization. a) Original simulated diffraction pattern with high-density sphere. b) Central region of discrete Fourier transform of a), showing the particle. c) The image in b) with limited range to showcase artifacts outside of the object outline. d) The simulated pattern with Hann window applied, resulting in lower intensity away from the center. e) Discrete Fourier transform of pattern after apodization. Slight blur visible. f) The image in e) with limited range. Artifacts found in c) mostly absent.

6.3 Benefits of applying COACS healing

A qualitative comparison of the results possible using COACS, speckle healing [53] and oversampling smoothness (OSS) [54] are given in **Fig. 6.3**, for 10 simulated patterns. For each unique random pattern with approximately 10000 photons outside the beamstop, HIO with COACS, OSS with COACS, HIO without COACS, OSS without COACS, and Speckle Healing (SH) are shown, as well as two cases without apodization. The high-density spherical feature on the edge is clearly visible in all COACS-healed reconstructions, and the contours of the icosahedron projection are always evident. Even with the additional processing introduced by OSS and SH, the non-COACS healed versions are not able to resolve this non-symmetrical feature. Furthermore, the edges of the icosahedron are less clearly defined and the noise outside it stronger. Without apodization, the correct features are less readily identified in the COACS case. While SH is supposed to implement a constraint that is equivalent to the one in COACS, the end result is better than pure HIO, but far from HIO + COACS in terms of overall visual quality.

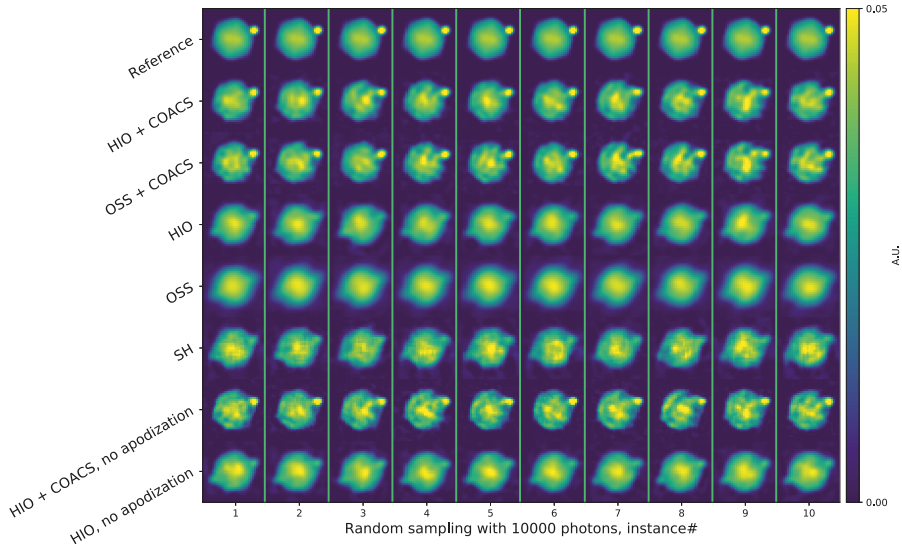

Figure 6.3. 10 phased reconstructions of sparse patterns based on the simulated particle, each with approximately 10000 photons, with different phase retrieval schemes compared against the true noiseless reference image. The methods include two cases using our COACS pre-processing, together with the two phasing methods Hybrid Input-Output and Oversampling Smoothness. The same methods were tried without COACS as well, plus Speckle Healing, since the additional constraint in that method is in theory equivalent to COACS. The final two cases verify the behavior when the apodizing Hann window is not applied. Each picture is an aligned average of the 10 best individual phasings out of 100 replicates for each combination of phasing method and sampled image.

In **Fig. 6.4** we show the R factor calculated at various radii (in pixels). This radial R factor has sometimes been referred to as the R Factor Transfer Function (RFTF [55]). The non-COACS methods are not able to correctly recover the intensities in the central missing data region.

Since the missing data region is a square with side 25, the inability to resolve it is most visible up to a radius of 12.5. This aforementioned inability to resolve the low-angle features correctly jeopardize the whole structure of the reconstruction.

In the overall minimum for the true signal at 90 pixels, due to the shape of the high-density spherical feature in the image, the direct application of COACS gives inferior results due to the presence of many more speckles in that pattern. Although the relative error is higher at this point, the absolute error is small nonetheless, since the true signal is close to 0.

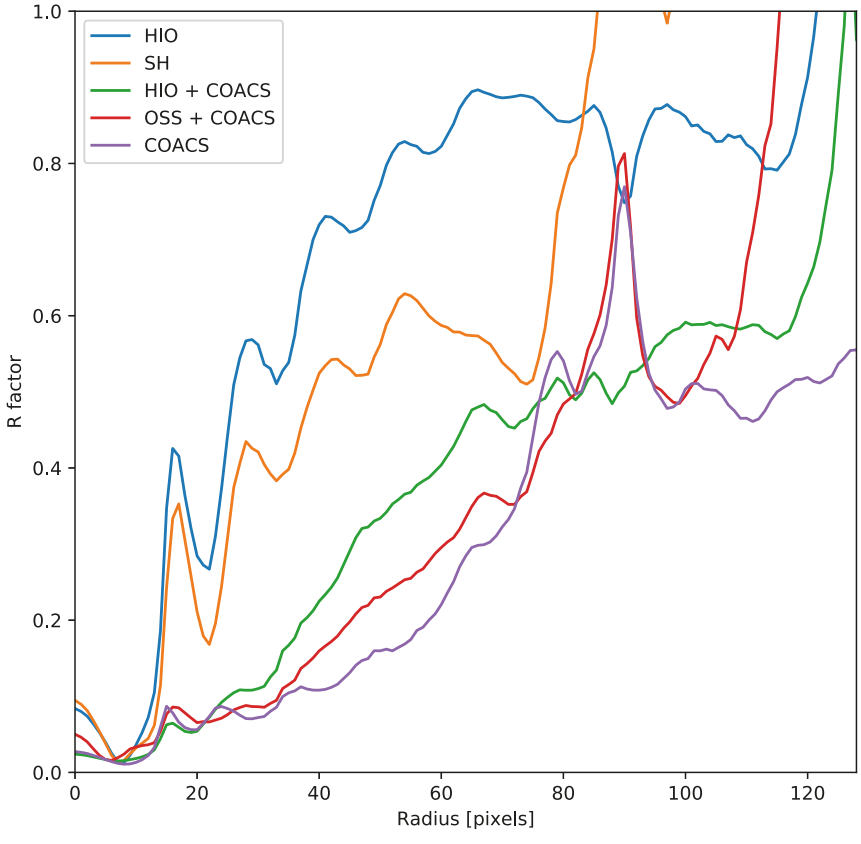

Figure 6.4. R factors (normalized relative L1 error) for various radius shells in pixels. Curves are averages over the individually computed results for all 50 simulated particles. Comparison between results based on average phasing of 10 best reconstructions of original pattern, average phasing of 10 best reconstructions of COACS-healed pattern, and using the structural factors from the COACS pattern directly. COACS-healing reduces phasing errors, but the phasing step is still a significant contributor to errors. OSS is competitive with HIO overall, but fails at reproducing correct signal at the lowest frequencies, which also contributes to the higher MSE error for that method. Speckle Healing, which theoretically implements the same autocorrelation constraint as COACS, improves results over pure HIO, but does not come near HIO + COACS. COACS peak at around 90 pixels is due to the R factor being a relative error metric. This is the location of a minimum due to the shape of the small spherical feature. Hence, absolute errors of the same magnitude are amplified.

7. Summary

Achievements

In this work we focused in modelling the beamline background, in order to make the dream of FXI on single protein-complexes possible. We specifically analyzed data collected during FXI experiments performed at the CXI instrument (at LCLS), which is the end station that has the most promising features to face the challenge.

In **Paper I** we reported the presence of a spatially non-uniform artifact of the CSPADs and we outlined a pipeline to reduce it, without losing detection power.

This paper was a prelude to **Paper II**, where we introduced a novel statistical approach to identify protein complexes at XFELs. To achieve this, we analyzed data collected during an FXI experiment on the RNA polymerase II: the first protein complex ever injected at an XFEL. We developed a hit-finding methodology, first tailored on the aforementioned dataset and then generalized in order to work on different datasets.

We first proved the reliability of the method proposed by comparing it with an independent hit-finding scheme, based on ion spectra collected by a ToF.

Then, we tested it on datasets concerning larger specimens (namely the OmRV and the PR772, two icosahedral viruses). Herein, in the case of OmRV, we showed we could find a doubled quantity of hits compared to simpler hit-finding schemes. Moreover, we also showed we could identify actual sample hits (clearly icosahedron edged) that were not identified previously.

Additionally, we corroborated our experimental results with computer simulations on protein hits. Those highlighted the superiority of the statistical hit-finder proposed over previously used hit-finders, and showed the feasibility of FXI on particle smaller than 40 nm (particularly noticeable is the case of 13 nm).

Finally, in (**Paper III**), we developed an algorithm (COACS) that aims to improve phase retrieval results.

It considers the probabilistic nature of the intensities on a detector and calculates the best (most probable) intensities for a given diffraction pattern. This is achieved by relying on the background model introduced in **Paper II**.

The “true” intensities obtained in this way are then used to feed alternating projection algorithms (e.g. ER). This resulted in an improved resolution of the phased 2D patterns and in a lowered error metrics.

Discussion

The entire project work was based on CXI experiments. As mentioned in **chapter 2**, the CXI instrument has all the features needed to realize the dream of performing FXI on single protein complexes (and single molecules).

In spite of this, experiments conducted at the AMO instrument have so far given better results (for particles greater or equal to 40 nm) compared to the similar experiments at the CXI Instrument. Why is that?

To explain this, it is crucial to focus on the pros and cons of the two.

First, the narrower beam focus spot we used at CXI — nominal values of 100 nm compared to 1 μm available at AMO —, theoretically allows stronger in-focus hits of the particles. In practice, however, the most clear result is a lower hit-rate.

Second, the shorter wavelengths (higher energies) available are optimal to provide higher resolution, but they also give a lower cross section. Besides, for an identical pulse energy, the number of photons is lower if each photon is more energetic. These factors combine to give a lower scattered signal.

Last but not least, even though CSPAD technology guarantees independent readouts per pixel (in contrast to AMO pnCCD detectors), it also provides worse gain-separation (ADU signal between 0 and 1 photon peak), which means more noise in the recorded patterns.

All those features that make the CXI instrument an ideal candidate for FXI on single particles, also prevented us from a complete success.

Future outlook

The XFEL community is rapidly growing, and, since the first free electron soft x-ray laser (FLASH at DESY, Hamburg, 2005), many other XFELs came out all over the world and are still coming (LCLS, United States; FERMI, Italy; SACLA, Japan; SwissFEL, Switzerland; European XFEL, Germany; LCLS-II, United States, etc.).

The European XFEL in particular has, as its most interesting feature, a very high repetition rate (nominally 27000 pulses/second). If the acquisition rate of CSPADs used at the LCLS (120 patterns/second) is compared with the one of AGIPDs [56] used at European XFEL (3500 patterns/second), that increases the number of patterns (and sample hits) collected by almost a factor 30.

In an hypothetical FXI experiment on the RNA polymerase II (or a somewhat larger sample, like ribosomes), in only one day of data collection at the European XFEL, we would have enough single-particle hits to retrieve its 3D electronic density.

As a matter of fact, a single protein imaging beamtime is approved for June 2019 at the European XFEL.

Given the high repetition rate of the new sources it is becoming cumbersome

to store data for every single pulse. Data reduction strategies are being implemented in those facilities.

An online mode that triggers acquisition is deployed to discard blank (non sample hit) events.

For instance, in the case of AGIPDs (at the European XFEL) a ToF detector scheme is being tried out as trigger. That is used as a first way of filtering.

Our proposed statistical hit-finder could be also used as a second filtering step, if the ToF detector settings are tuned to avoid hits at the cost of more false positives.

The background model underlying our hit-finding approach can also be used in phase retrieval (as already mentioned) and in the EMC algorithm, to reconstruct 3D sample structures from a number of individual 2D snapshots.

Sammanfattning på svenska

Studiet av biologiska molekyler och mekanismer har inneburit stora framsteg från mänskligheten.

Louis Pasteurs upptäckter rörande infektionssjukdomar och mikrobdriven fermentering under 1800-talet var att glänta på dörren till den lilla biologins värld. Upptäckterna under 1900-talet omfattade bland annat strukturerna för DNA och hemoglobin. De senaste decennierna har bilden blivit mycker mer komplett, med en ökande förståelse av bland annat enzymer och antikroppar, som 2018 års Nobelpris i kemi till Frances H. Arnold, George P. Smith och Sir Gregory P. Winter.

Att förstå smittämnens struktur (svampar, bakterier, virus o.s.v.) och celler (den grundläggande enheten för allt liv) har gjort det möjligt att bättre förstå livets beteende och variation. Det i sin tur har gjort det möjligt att ta fram läkemedel, vacciner och behandlingar för att förhindra, lindra och behandla sjukdomar. Men den här kunskapen har inte bara varit viktig inom medicin, men också för tekniska tillämpningar. Exempelvis har insikter från hur fotosyntesen använder solljus för att tillverka socker i växter och bakterier utgjort grunden för effektiva solceller som imiterar samma process.

Men om det är så viktigt att förstå livets struktur, ända ned på molekylnivå, hur har vi egentligen kunnat se alla de här sakerna, som kan vara mycket mindre än 1 mm? Och dessutom se alla detaljer? Vilken sorts teknik, vilken sorts instrument används då?

När den moderna biologin inleddes var det optiska mikroskopet det viktigaste instrumentet. Mikroskopet uppfanns på 1600-talet av nederländaren Anton van Leeuwenhoek. Plötsligt gick det att se olika prover i mycket större än förstoring än vad som gick med bara en lupp. Med detta steg kunde man börja upptäcka mikroorganismer och celler.

Trots alla upptäckter som möjliggjordes med mikroskopet hade även det optiska mikroskopet begränsningar. Även den bästa mikroskoplins kommer alltid att orsaka så kallad aberration (d.v.s. att ljuset sprids över ett litet område i stället för att perfekt fokuseras i en enda punkt). Den maximala upplösningen, möjligheten att skilja närliggande detaljer åt, när man använder ett mikroskop med vanliga linser, är omkring 100 nm. En nanometer äe en miljarddels meter. En miljarddel kan vara svår att riktigt föreställa sig, men om en meter är hela sträckan mellan Uppsala och Stockholm vore en nanometer inte tjockare än ett hårstrå!

I ett mikroskop studeras provet med synligt ljus. Synligt ljus är en form av elektromagnetisk strålning, med våglängder i intervallet 360 - 800 nm. Om

man vill urskilja mindre detaljer än 100 nm måste något annat användas för att studera provet.

Med grundläggande vågfysik kan man förstå att med en kortare våglängd blir det möjligt att uppnå högre upplösning – och se mindre detaljer.

Därför inleddes en ny revolution 1895 när Wilhelm Röntgen upptäckte den strålning som fått hans namn (som visade sig ha våglängd ≤ 10 nm). Det var grunden för ett antal metoder som nu är vanliga i fysik, biofysik, strukturbiologi, kemi och medicin.

I början skapade man röntgenstrålning med så kallade röntgenrör. Det var enkla katodstrålerör som kunde omvandla elektricitet till strålning. Efter några årtionden kom en ny avgörande typ av röntgenkällor, synkrotroner. Där används en stor ring som vrider en injicierad elektronstråle. Från den kommer så kallad icke-koherent strålning. Det betyder att de olika fotonerna i strålningen inte är i fas. Synkrotroner kan ge relativt många fotoner per ytenhet per sekund, jämfört med röntgenrör.

Både röntgenrör och synkrotroner har användsf ör att studera 3D-strukturerna för biologiska komplex. Hur då? Vi måste först förstå hur röntgenstrålarna och atomerna växelverkar och vad det är vi sedan kan mäta med våra instrument. Sedan måste vi förstå hur det går till att från de här mätvärdena få fram provets elektrontäthet (3D-struktur).

När något prov träffas av röntgenstrålar kommer en del av strålarna att spridas åt olika håll. Det beror främst på hur strålarna växelverkar med det elektronmoln som rör sig i provets atomer.

Hur mycket röntgenstrålarna sprids från provet (oavsett om det är en atom, en molekyl, en cell eller något annat) benämns som provets spridningsstyrka. En enskild atom har alltså en viss spridningsstyrka. En hel molekyls spridningsstyrka är en summa av styrkan för de enskilda atomerna.

Olika detektorer kan "se" röntgenstrålningen som har spridits från provet. Antingen mäts strålningen bara som ett värde, eller så är den så avancerad att den kan räkna det exakta antalet fotoner som har träffat olika delar av detektorn. I grunden använder detektorerna samma teknik som vanliga digitalkameror, men för röntgen i stället för synligt ljus. Men det går ändå inte att uppfatta signalen från en enda atom. Ännu längre tillbaka användes vanliga röntgenplåtar eller film som detektorer, precis som i medicinska röntgenundersökningar på den tiden. Då gick det inte att se den spridda signalen från en enskild molekyl, eller från en atom. Men hur var det då möjliga att upptäcka strukturen på DNA från röntgenbilder, till exempel? Vi måste förstärka signalen.

Ett sätt att få en starkare signal är att ha strukturer som upprepar sig i samma mösnter – kristaller. Kristaller användes tidigt och är fortfarande oerhört viktiga. Om ett villkor som kallas Braggs lag är uppfyllt kommer den signal som sprids från någon typ av prov (med atomer eller molekyler) att motsvara att bidraget från varje del i kristallen läggs ihop. Eftersom bilden på detektorn faktiskt är signalens intensitet i stället för ampltiud blir signalen

från molekylen faktiskt proportionell med kvadraten på antalet atomer (eller molekyler) i kristallen.

Om man har en kristall med en miljon molekyler kommer den då att förstärkas en biljon gånger!

Sedan 2006 finns det även en nby typ av röntgenkällor, röntgenfrielektronlasrar (eller X-ray Free Electron Lasers, XFELs). Med en XFEL kan man rikta en koherent och mycket stark röntgenpuls mot provet, en miljard gånger starkare än från en synkrotron, under ett kort ögonblick, bara någon tusendels miljarddel av en sekund.

En så stark puls kommer att spränga hela provet nästan direkt. Men provets atomer hinner inte röra sig särskilt mycket på en tusendaels miljarddels sekund. Röntgenpulsen är helt enkelt så kort att den spridningsbild man får föreställer provet innan det hann sprängas. Detta kallas spridning före tillintetgörande, eller "diffraction before destruction".

De detektorer som används tar plana, tvådimensionella bilder. De prov vi vill avbilda är i 3D. För att kunna göra en 3D-avbildning skulle man behöva ta flera bilder av samma prov i olika vinklar.

För att göra det kan man flytta röntgenkällan. Den metoden var vanlig med röntgenrör. Det går också att rotera provet, en teknik som används både med röntgenrör och synkrotroner. En sista möjlighet är att ha flera exemplar av provet och exponera varje exemplar i en slumpmässig vridning. Det är denna metod som vi har använ vid röntgenlasrar under så kallad ögonblicksröntgenavbildning (Flash X-ray Imaging, FXI).

I samtliga fall representerar bilden på detektorn den spridda signalens intensiteter. Det är inte samma sak som en direkt bild av provet. Spridningsbilden kan kopplas till det ursprungliga provet genom den matematiska metoden fouriertransformation. Det går att göra fouriertransformationen baklänges och på så vis få fram elektrontätheten för provet, det vill säga provets struktur. Men vissa delar av signalen måste återskapas med beräkningsmetoder för att det ska bli möjligt.

Vid ett FXI-experiment vid en röntgenlaser (främst Linac Coherent Light Source, LCLS, nära Stanford i USA), har man lyckat srekonstruera en 2D-bild och ibland 3D-strukturer för biologiska partiklar större än 40 nm. Det har varit svårare för mindre partiklar. Det beror främst på att biologiska makromolekl̈ära komplex har en relativt låg spridningsstyrka. Ett intressant prov har varit enzymkomplexet RNA-polymeras II, omkring 13 nm i diameter.

Signalen från ett sådant prov är bara marginellt starkare än bakgrundsbruset. Därför blir det svårt att skilja korrekta bilder av prov från rent brus.

I den här avhandlingen har syftet varit att komma ett steg närmare att utföra ögonblicksröntgenavbildning av enskilda proteinkomplex. Jag visar att det går att minska en specifik brusstörning som påverkade detektorerna i våra experiment (artikel I). Vidare studerar jag avbildning av RNA-polymeras II, det första proteinkomplex som har injicerats vid en röntgelaser. Jag visar att det går att skapa en modell för bakgrundsbruset och använda den modellen för

att lättare automatiskt separera bakgrundsbilder från bilder av provet, även när signal-brusförhållandet (SNR) gör det mycket svårt. Jag visar också att vår metod är mer generellt relevant genom att använda den på insamlade data från andra prover. Genom separata datorsimuleringar visar jag närmare hur mycket bättre vår metod fungerar och vilka begränsningar den kan ha (artikel II). Jag redogör också för den programvara som jag har utvecklat för att utföra dessa analyser. Slutligen presenterar jag i samarbete en metod som kan använda en bakgrundsmodell och en ny beskrivning av rekonstruktionsproblemet för att skapa bättre 2D-rekonstruktioner av provet utifrån intensiteterna på detektorn (artikel III).

Jag är övertygad om att dessa nya insikter och metoder, tillsammans med nya röntgenlasrar som kan avge många fler pulser per sekund, upp till 27000 Hz vid den europeiska XFEL som nu finns i Hamburg, kan möjlilggöra den första 3D-rekonstruktionen genom röntgenavbildning av enskilda proteinkomplex.

Author contributions

Paper I
I performed all data analysis, created all the figures and wrote the manuscript.

Paper II
I contributed in developing the statistical hit-finding methodology, implemented the source code, and performed all data analysis and simulations. I also created all the figures and wrote the manuscript.

Paper III
I performed the data analysis whose results where used for developing the COACS method. I also simulated the test particle used as benchmark in the paper.

Acknowledgements

This work was possible only thanks to a team effort made in collaboration with a large international network of researchers and institutions. So, I would like to thank every and each person who took part in it!

In particular, I would like to thank my supervisor **Carl Nettelblad** for his endless patience, his availability and readiness to solve any issue or doubt I had. I am really thankful for the time you have devoted to me and for all the things I learned from you.

Special thanks go to **Janos Hajdu** and **Inger Andersson** for accepting me in the Laboratory of Molecular Biophysics group and for their support over these years.

Thanks to **Filipe Maia**, who was always there, ready to give me a lot of good advice and to help me with my coding troubles; thanks to **Benedikt Daurer** and **Tomas Ekeberg** for being always available if I needed advice and for letting me discover bouldering; thanks to **Nicusor Timneaunu** for clarifying the more technical and experimental aspects of the work. I would also like to thank **Laura H. Gunn** for cheering me up many times, for solving my doubts in English and for being a friend.

Thanks to all the computer guys for the help and interesting discussions: **Gijs van Der Schot**,**Max Hantke**, **Ida Lundholm**, **Jing Liu** and **Jonas Sellberg**. Thanks to all the experimental guys for developing, building and operating the experimental setup at the LCLS: **Johan Bielecki**, **Daniel Westphal**, **Alessandro Zani**, **Federico Benzi**, **Marvin Seibert**, **Kerstin Mühlig**, **Olena Kulyk** and **Jakob Andreasson**. Many thanks to the biologists, who provided the sample for the experiments: **Hemanth Kumar**, **Kenta Okamoto**, **Anna Munke**, **Anna Larsson**, **Gunilla Carlsson**, **Dirk Hasse**, **Karin Valegård**, **Martin Svenda** and **Margareta Ingelman**.

I am deeply indebted to all the aforementioned people and to all the current and former members of the lab, that with a word, a

discussion or simply their presence, made me feel — even once – better. Thank you all for the wonderful time spent together, inside or outside the corridor's lab, for the awesome ski trips and for making me grow up as a scientist and a person!

Moreover, I would like to thank all the friends I met here in Uppsala: **Marco C.** for entertaining me during lunch times, for the good advice given to me, for both the serious and silly discussions and for being always able to make me laugh; **Matteo C.** for being a lunch companion as well and for a lot of good bouldering/climbing tips that made me level up; **Marco R.** for the endless discussions on everything, for the breakfasts/lunches/dinners/fikas had together, for being a brilliant and witty person, a good neighbor and a true friend; **Simone**, for a lot of good advice, for cheering me up, being supportive and for calling and waking me up at 8 o' clock in the morning of Christmas day (damn you! don't you ever dare again!), but, especially, for being an amazing travel companion (what country is the next?); **Ilaria**, for being such an amazing person and talented biologist, and for making me discover the breathtaking, magic and mysterious landscapes, towns and villages of Basilicata, an Italian region that, until 2016, was signed on my geographic map as "hic sunt leones"; **Alice**, for being so "Èlis" (seriously, what are the odds that one from Ospedalicchio meets another one from Foligno in Uppsala?); **Olivia** for her patience in trying to teach me Swedish, for being always kind and for owning a perfect mix of Swedish and Italian soul; **Irene**, for being my "ad honorem" flatmate. Thanks to all the other marvellous people of the Uppsalingo and "Mafia Uppsala" gang, for always cheering me up and being fun: **Armin**, **Elisa**, **Hedda**, **Claudio**, **Giulia**, **Jõao**, **Fabio**, **Karl** and **Nora**.

A special thank goes to the Italian friends gang, who made (and still make) me feel at home: **Tony** and **Paola**, for being good friends and flatmates since the days in Flogsta; **Chiara**, without whom no days in Flogsta would have been possible; **Valentina**, for being a wonderful person, always lively, cheerful and positive, and for lot of good advice; **Francesca** and **Diego**, again for a lot of good advice, for being such splendid people, and for giving birth to **Niccolò** (and now **Mattia**); **Beppe**, **Matteo B.**, **Luca**, **Riccardo** e **Francesca**, **Arianna** and **Siggi**, for being always awesome!

Un ringraziamento speciale va anche ai miei amici e familiari in Italia, che, da anni, hanno troppo spesso l'onere (e raramente l'onore) di avermi accanto.

Come non ringraziare voi, **Silvia** e **Francesca**? Siete state sempre al mio fianco, amiche e sorelle, sia quando ero ancora a Perugia, sia ora che sono qui, nelle fredde lande svedesi. Avete sempre su(o)pportato ogni mia scelta e ogni mia cavolata. Le vostre parole di incoraggiamento, i vostri consigli — e ancor di più le vostre critiche — sono state, sono, e sempre saranno per me uno sprone al miglioramento. (Dopo questa captatio benevolentiae, vedete di venirmi a trovare qui in Svezia, ché è ora!).

Un altro grazie immenso va a **Giulia**, **Martina**, **Betta** e **Leo**, che mi sopportano dai tempi del liceo e che da allora mi sono sempre stati accanto. Anche voi dispensatori di molti buoni consigli e altrettante critiche, ieri come oggi, mi siete indispensabili, con la vostra presenza di spirito e con il vostro affetto. Grazie davvero!

Voglio ringraziare le mie zie e i miei zii, i miei cugini (**Gabriella** e **Peppe**, **Rosanna** e **Benito**, **Federico**, **Roberto** ed **Emanuele**), mio fratello **Giacomo** e **Nancy**, per essere sempre disponibili e incoraggianti.

Infine, la mia eterna gratitudine va a tutti i miei nonni, per aver fortemente contribuito a plasmare la persona che sono diventato. Sono loro debitore di tutto ciò che di positivo vi è in me (per il negativo, opera mia!).

Sono e sarò sempre riconoscente a nonno **Nazzareno**, che non c'è più, e a nonna **Franca**, per tutto quello che mi hanno insegnato e per l'affetto che mi hanno dato; e a nonno **Nando**, che, solitamente burbero e arcigno, mi riserva sempre un sorriso quando torno a casa in Italia.

A nonna **Maria**, in particolare, devo la mia intera carriera scolastica. Senza di te, sarei, ahimé, fermo alla materna! Ricordo ancora quando mi aiutavi con il dettato, con le operazioni, e, in generale, con i compiti alle elementari. Ma, soprattutto, ricordo le innumerevoli sveglie ("Albe', ma s'è ancora a letto? Alzete, ch'ènno le 8!") per farmi andare a prendere l'autobus che mi avrebbe condotto a scuola (alla medie e alle superiori). Invariabilmente, ogni mattina, le tue 8 erano sempre le 7!

E come dimenticare gli anni dell'università, finanziati a suon di cinquantoni per esame passato? (Son valsi più dei rimborsi e delle borse di studio!).

Per questo, e tanto altro ancora, mille volte grazie!

Un altro e ultimo ringraziamento (per il quale non basterebbero le pagine di questa tesi) è d'uopo: a babbo e mamma, cui devo tutto.

References

[1] E. G. van Putten, D. Akbulut, J. Bertolotti, W. L. Vos, A. Lagendijk, and A. P. Mosk. Scattering Lens Resolves Sub-100 nm Structures with Visible Light. *Physical Review Letters*, 106(193905), 2011.

[2] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

[3] R. E. Franklin and R. G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171:740–741, 1953.

[4] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson. Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738–740, 1953.

[5] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North. Structure of hæmoglobin: a three-dimensional fourier synthesis at 5.5-Å resolution, obtained by x-ray analysis. *Nature*, 185(4711):752–777, 1960.

[6] L. C. Johansson, A. B. Wöhri, G. Katona, S. Engström, and R. Neutze. Membrane protein crystallization from lipidic phases. *Current opinion in structural biology*, 19(4):372–378, 2009.

[7] E. P. Carpenter, K. Beis, and A. D. Cameron. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, 18:581–586, 2008.

[8] B. W. J. McNeil and N. R. Thompson. X-ray free-electron lasers. *Nature Photonics*, 4:814–821, 2010.

[9] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu. Potential for biomolecular imaging with femtosecond x-ray pulses. *Nature*, 406:416–422, 2000.

[10] B. J. Daurer, K. Okamoto, J. Bielecki, F. R. N. C. Maia, K. Mühlig, M. M. Seibert, M. F. Hantke, C. Nettelblad, W. H. Benner, M. Svenda, N. Tîmneanu, T. Ekeberg, N. Duane Loh, A. Pietrini, A. Zani, A. D. Rath, D. Westphal, R. A. Kirian, S. Awel, M. O. Wiedorn, G. van Der Schot, G. H. Carlsson, D. Hasse, J. A. Sellberg, A. Barty, J. Andreasson, S. Boutet, G. Williams, J. Koglin, I. Andersson, J. Hajdu, and D. S. D. Larsson. Experimental strategies for imaging bioparticles with femtosecond hard x-ray pulses. *IUCrJ*, 4:251–262, 2017.

[11] M. F. Hantke, D. Hasse, F. R. N. C. Maia, T. Ekeberg, K. John, M. Svenda, N. Duane Loh, A. V. Martin, N. Timneanu, D. S. D. Larsson, G. van der Schot, G. H. Carlsson, M. Ingelman, J. Andreasson, D. Westphal, M. Liang, F. Stellato, D. P. DePonte, R. Hartmann, N. Kimmel, R. A. Kirian, M. M. Seibert, K. Mühlig, S. Schorb, K. Ferguson, C. Bostedt, S. Carron, J. D. Bozek, D. Rolles, A. Rudenko, S. Epp, H. N. Chapman, A. Barty, J. Hajdu, and I. Andersson. High-throughput imaging of heterogeneous cell organelles with an x-ray laser. *Nature Photonics*, 8:943–949, 2014.

[12] G. van der Schot, M. Svenda, F. R. N. C. Maia, M. F. Hantke, D. P. DePonte, M. M. Seibert, A. Aquila, J. Schulz, R. Kirian, M. Liang, F. Stellato, B. Iwan, J. Andreasson, N. Timneanu, D. Westphal, F. N. Almeida, D. Odic, D. Hasse, G. H. Carlsson, D. S. D. Larsson, A. Barty, A. V. Martin, S. Schorb, C. Bostedt, J. D. Bozek, D. Rolles, A. Rudenko, S. Epp, L. Foucar, B. Rudek, R. Hartmann, N. Kimmel, P. Holl, L. Englert, N. Duane Loh, H. N. Chapman, I. Andersson, J. Hajdu, and T. Ekeberg. Imaging single cells in a beam of live cyanobacteria with an x-ray laser. *Nature Communications*, 6(5704), 2015.

[13] P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. Bucksbaum, R. Coffee, F.-J. Decker, Y. Ding, D. Dowell, S. Edstrom, A. Fisher, J. Frisch, S. Gilevich, J. Hastings, G. Hays, P. Hering, Z. Huang, R. Iverson, H. Loos, M. Messerschmidt, A. Miahnahri, S. Moeller, H.-D. Nuhn, G. Pile, D. Ratner, J. Rzepiela, D. Schultz, T. Smith, P. Stefan, H. Tompkins, J. Turner, J. Welch, W. White, J. Wu, G. Yocky, and J. Galayda. First lasing and operation of an ångstrom-wavelength free-electron laser. *Nature Photonics*, 4:641–647, 2010.

[14] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21:2758–2769, 1982.

[15] S. Marchesini. A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments*, 78(011301), 2007.

[16] S. Stolika, J. A. Delgado, A. Pérez, and L. Anasagasti. Measurement of the penetration depths of red and near infrared light in human “ex vivo” tissues. *Journal of photochemistry and photobiology B: Biology*, 57:90–93, 2000.

[17] W. C. Röntgen. On a new kind of rays. *Nature*, 53:274–276, 1896.

[18] H. S. Allen. X-rays and their applications. *Nature*, 127:356–358, 1931.

[19] Y. Shi. A glimpse of structural biology through x-ray crystallography. *Cell*, 159:995–1014, 2014.

[20] D. Paganin. *Coherent X-ray Optics*. Oxford Series on Synchrotron radiation. 2006.

[21] C. Kittel. *Introduction to Solid State Physics (3rd edition)*. John Wiley & Sons, Inc., 1968.

[22] E. F. Garman and M. Weik. X-ray radiation damage to biological macromolecules: further insights. *Journal of Synchrotron Radiation*, 24:1–6, 2017.

[23] 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, R. B. Doak, F. R. N. C. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Hömke, C. Reich, D. Pietschner, L. Strüder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K.-U. Kühnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jönsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C.-D. Schröter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. M. Barends, R. Neutze,

S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, and J. C. H. Spence. Femtosecond x-ray protein nanocrystallography. *Nature*, 470:73–77, 2011.

[24] A. Barty, C. Caleman, A. Aquila, N. Timneanu, L. Lomb, T. A. White, J. Andreasson, D. Arnlund, S. Bajt, T. R. M. Barends, M. Barthelmess, M. J. Bogan, C. Bostedt, J. D. Bozek, R. Coffee, N. Coppola, J. Davidsson, D. P. DePonte, R. B. Doak, T. Ekeberg, V. Elser, S. W. Epp, B. Erk, H. Fleckenstein, L. Foucar, P. Fromme, H. Graafsma, L. Gumprecht, J. Hajdu, C. Y. Hampton, R. Hartmann, A. Hartmann, G. Hauser, H. Hirsemann, P. Holl, M. S. Hunter, L. Johansson, S. Kassemeyer, N. Kimmel, R. A. Kirian, M. Liang, F. R. N. C. Maia, E. Malmerberg, S. Marchesini, A. V. Martin, K. Nass, R. Neutze, C. Reich, D. Rolles, B. Rudek, A. Rudenko, H. Scott, I. Schlichting, J. Schulz, M. M. Seibert, R. L. Shoeman, R. G. Sierra, H. Soltau, J. C. H. Spence, F. Stellato, S. Stern, L. Strüder, J. Ullrich, X. Wang, G. Weidenspointner, U. Weierstall, C. B. Wunderer, and H. N. Chapman. Self-terminating diffraction gates femtosecond X-ray nanocrystallography measurements. *Nature Photonics*, 6:35–40, 2012.

[25] P. Fromme. XFELs open a new era in structural chemical biology. *Nature Chemical Biology*, 11:895–899, 2015.

[26] H. N. Chapman, S. P. Hau-Riege, M. J. Bogan, S. Bajt, A. Barty, S. Boutet, S. Marchesini, M. Frank, B. W. Woods, W. H. Benner, R. A. London, U. Rohner, A. Szöke, E. Spiller, T. Möller, C. Bostedt, D. A. Shapiro, M. Kuhlmann, R. Treusch, E. Plönjes, F. Burmeister, M. Bergh, C. Caleman, G. Huldt, M. M. Seibert, and J. Hajdu. Femtosecond time-delay X-ray holography. *Nature*, 448:676–679, 2007.

[27] P. Hart1, S. Boutet, G. Carini, A. Dragone, B. Duda, D. Freytag, G. Haller, R. Herbst, S. Herrmann, C. Kenney, J. Morse, M. Nordby, J. Pines, N. van Bakel, M. Weaver, and G. Williams. The Cornell-Pixel Array Detector at LCLS. Nuclear Science Symposium, Medical Imaging Conference, Anaheim, CA, 2012.

[28] S. Herrmann, S. Boutet, B. Duda, David Fritz, G. Haller, P. Hart, R. Herbst, C. Kenney, H. Lemkeb, M. Messerschmidt, J. Pines, A. Robert, M. Sikorski, and G. Williams. CSPAD-140k: A versatile detector for LCLS experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 718:550–553, 2013.

[29] S. Herrmann, P. Hart, A. Dragone, D. Freytag, R. Herbst, J. Pines, M. Weaver, G. A. Carini, J. B. Thayer, O. Shawn, C. J. Kenney, and G. Haller. CSPAD upgrades and CSPAD V1.5 at LCLS. Journal of Physics: Conference Series, 2014.

[30] G. Blaj, P. Caragiulo, G. Carini, S. Carron, A. Dragone, D. Freytag, G. Haller, P. Hart, J. Hasi, R. Herbst, S. Herrmann, C. Kenney, B. Markovic, K. Nishimura, S. Osier, J. Pines, B. Reese, J. Segal, A. Tomada, and Matt Weaver. X-ray detectors at the Linac Coherent Light Source. *Journal of Synchrotron Radiation*, 22:577–583, 2015.

[31] K. R. Ferguson, M. Bucher, J. D. Bozek, S. Carron, J.-C. Castagna, R. Coffee, G. I. Curiel, M. Holmes, J. Krzywinski, Marc M. Messerschmidt, M. Minitti,

A. Mitra, S. Moeller, P. Noonan, T. Osipov, S. Schorb, M. Swiggers, A. Wallace, J. Yin, and C. Bostedt. The atomic, molecular and optical science instrument at the Linac Coherent Light Source. *Journal of Synchrotron Radiation*, 22(3):492–497, 2015.

[32] L. Strüder, S. Epp, D. Rolles, R. Hartmann, P. Holl, G. Lutz, H. Soltau, R. Eckart, C. Reich, K. Heinzinger, C. Thamm, A. Rudenko, F. Krasniqi, K.-U. Kühnel, C. Bauer, C.-D. Schröter, R. Moshammer, S. Techert, D. Miessner, M. Porro, O. Hälker, N. Meidinger, N. Kimmel, R. Andritschke, F. Schopper, G. Weidenspointner, A. Ziegler, D. Pietschner S. Herrmann, U. Pietsch, A. Walenta, W. Leitenberger, C. Bostedt, T. Möller, D. Rupp, M. Adolph, H. Graafsma, H. Hirsemann, K. Gärtner, R. Richter, L. Foucar, R. L. Shoeman, I. Schlichting, and J. Ullrich. Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 614(3):483–496, 2010.

[33] M. Liang, G. J. Williams, M. Messerschmidt, M. M. Seibert, P. A. Montanez, M. Hayes, D. Milathianaki, A. Aquila, M. S. Hunter, J. E. Koglin, D. W. Schafer, S. Guillet, A. Busse, R. Bergan, W. Olson, K. Fox, N. Stewart, R. Curtis, A. A. Miahnahri, and S. Boutet. The Coherent X-ray Imaging instrument at the Linac Coherent Light Source. *Journal of Synchrotron Radiation*, 22:514–519, 2015.

[34] D. P. DePonte, U. Weierstall, K. Schmidt, J. Warner, D. Starodub, J. C. H. Spence, and R. B. Doak. Gas dynamic virtual nozzle for generation of microscopic droplet streams. *Journal of Physics D: Applied Physics*, 41(19), 2008.

[35] R. A. Kirian, S. Awel, N. Eckerskorn, H. Fleckenstein, M. Wiedorn, L. Adriano, S. Bajt, M. Barthelmess, R. Bean, K. R. Beyerlein, L. M. G. Chavas, M. Domaracky, M. Heymann, D. A. Horke, J. Knoska, M. Metz, A. Morgan, D. Oberthuer, N. Roth, T. Sato, P. L. Xavier, O. Yefanov, A. V. Rode, J. Küpper, and H. N. Chapman. Simple convergent-nozzle aerosol injector for single-particle diffractive imaging with X-ray free-electron lasers. *Structural Dynamics*, 2(041717), 2015.

[36] N. Roth, S. Awel, D. A. Horke, and Jochen Küpper. Optimizing aerodynamic lenses for single-particle imaging. *Journal of Aerosol Science*, 124:17–29, 2018.

[37] I. Gaponenko and C. O'Grady. SLAC confluence website. Available at `https://confluence.slac.stanford.edu/display/PSDM/psana+python+Setup`.

[38] M. Dubrovin. SLAC confluence website. Available at `https://confluence.slac.stanford.edu/display/PSDM/Common+mode+correction+algorithms#Commonmodecorrectionalgorithms-#1-commonmodepeakfindingalgorithm`.

[39] A. Munke, J. Andreasson, A. Aquila, S. Awel, K. Ayyer, A. Barty, R. J. Bean, P. Berntsen, J. Bielecki, S. Boutet, M. Bucher, H. N. Chapman, B. J. Daurer, H. DeMirci, V. Elser, P. Fromme, J. Hajdu, M. F. Hantke, A. Higashiura, B. G. Hogue, A. Hosseinizadeh, Y. Kim, R. A. Kirian, H. K.N. Reddy, T.-Y. Lan, D. S.D. Larsson, H. Liu, N. Duane Loh, F. R.N.C. Maia, A. P. Mancuso,

K. Mühlig, A. Nakagawa, D. Nam, G. Nelson, C. Nettelblad, K. Okamoto, A. Ourmazd, M. Rose, G. van der Schot, P. Schwander, M. M. Seibert, J. A. Sellberg, R. G. Sierra, C. Song, M. Svenda, N. Timneanu, I. A. Vartanyants, D. Westphal, M. O. Wiedorn, G. J. Williams, P. L. Xavier, C. H. Yoon, and J. Zook. Coherent diffraction of single Rice Dwarf virus particles using hard X-rays at the Linac Coherent Light Source. *Scientific Data*, 3, 2016.

[40] T. Ekeberg, M. Svenda, C. Aberge, F. R. N. C. Maia, V. Seltzer, J.-M. Claverie, M. Hantke, O. Jönsson, C. Nettelblad, G. van der Schot, M. Liang, D. P. DePonte, A. Barty, M. M. Seibert, B. Iwan, I. Andersson, N. Duane Loh, A. V. Martin, H. Chapman, C. Bostedt, J. D. Bozek, K. R. Ferguson, J. Krzywinski, S. W. Epp, D. Rolles, A. Rudenko, R. Hartmann, N. Kimmel, and J. Hajdu. Three-Dimensional Reconstruction of the Giant Mimivirus Particle with an X-ray Free-Electron Laser. *Physical Review Letters*, 114(098102), 2015.

[41] J. M. Kirkpatrick and B. M. Young. Poisson Statistical Methods for the Analysis of Low-Count Gamma Spectra. *IEEE Transactions on Nuclear Science*, 56(3):1278–1282, 2009.

[42] J. Baglivo and D. Olivier. Methods for Exact Goodness-of-Fit Tests. *Journal of the American Statistical Association*, 87:464–469, 1992.

[43] M. L. Hazelton. *Methods of Moments Estimation.* Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg, 2011.

[44] B. J. Daurer, M. F. Hantke, C. Nettelblad, and F. R. N. C. Maia. Hummingbird: monitoring and analyzing flash X-ray imaging experiments in real time. *Journal of Applied Crystallography*, 49(3):1042–1047, 2016.

[45] S. Hahn. Structure and mechanism of the RNA Polymerase II transcription machinery. *Nature Structural and Molecular Biology*, 11(5):394–403, 2004.

[46] B. Nagler, A. Aquila, S. Boutet, E. C. Galtier, A. Hashim, M. S. Hunter, M. Liang, A. E. Sakdinawat, C. G. Schroer, A. Schropp, M. H. Seaberg, F. Seiboth, T. van Driel, Z. Xing, Y. Liu, and H. J. Lee. Focal Spot and Wavefront Sensing of an X-Ray Free Electron laser using Ronchi shearing interferometry. *Scientific Reports*, 7(13698), 2017.

[47] J. Andreasson, A. V. Martin, M. Liang, N. Timneanu, A. Aquila, F. Wang, B. Iwan, M. Svenda, T. Ekeberg, M. Hantke, J. Bielecki, D. Rolles, A. Rudenko, L. Foucar, R. Hartmann, B. Erk, B. Rudek, H. N. Chapman, J. Hajdu, and A. Barty. Automated identification and classification of single particle serial femtosecond X-ray diffraction data. *Optics Express*, 22(3):2497–2510, 2014.

[48] H. K.N. Reddy, C. H. Yoon, A. Aquila, S. Awel, K. Ayyer, A. Barty, P. Berntsen, J. Bielecki, S. Bobkov, M. Bucher, G. A. Carini, S. Carron, H. Chapman, B. Daurer, H. DeMirci, T. Ekeberg, P. Fromme, J. Hajdu, M. F. Hanke, P. Hart, B. G. Hogue, A. Hosseinizadeh, Y. Kim, R. A. Kirian, R. P. Kurta, D. S.D. Larsson, N. Duane Loh, F. R.N.C. Maia, A. P. Mancuso, K. Mühlig, A. Munke, D. Nam, C. Nettelblad, A. Ourmazd, M. Rose, P. Schwander, M. Seibert, J. A. Sellberg, C. Song, J. C.H. Spence, M. Svenda, G. Van der Schot, I. A. Vartanyants, G. J. Williams, and P. L. Xavier. Coherent soft X-ray diffraction imaging of coliphage PR772 at the Linac coherent light source. *Scientific Data*, 4(170079), 2017.

[49] M. F. Hantke, T. Ekeberg, and F. R. N. C. Maia. Condor: a simulation tool for flash X-ray imaging. *Jouurnal of Applied Crystallography*, 49:1356–1362,

2016.
[50] M. Jette and M. Grondona. Slurm: Simple Linux Utility for Resource Management. Proceedings of ClusterWorld Conference and Expo, San Jose, California, 2003.
[51] A. Barty, R. A. Kirian, F. R. N. C. Maia, M. Hantke, C. H. Yoon, T. A. White, and H. Chapman. Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *Jouurnal of Applied Crystallography*, 47:1118–1131, 2014.
[52] F. R. N. C. Maia. The Coherent X-ray Imaging Data Bank. *Nature Methods*, 9:854–855, 2012.
[53] N. Duan Loh, S. Eisebitt, S. Flewett, and V. Elser. Recovering magnetization distributions from their noisy diffraction data. *Physical Review E*, 82(061128), 2010.
[54] J. A Rodriguez, R. Xu, C.-C. Chen, Y. Zou, and J. Miao. Oversampling smoothness: an effective algorithm for phase retrieval of noisy diffraction intensities. *Journal Applied Crystallography*, 46:312–318, 2013.
[55] S. Marchesini, H. N. Chapman, A. Barty, C. Cui, M. R. Howells, J. C. H. Spence, U. Weierstall, and A. M. Minor. Phase aberrations in diffraction microscopy. page 380–382, 2006.
[56] A. Allahgholi, J. Becker, L. Bianco, R. Bradfordf, A. Delfs, R. Dinapoli, P. Goettlicher, M. Gronewald, H. Graafsma, D. Greiffenbergb, B. H. Henrich, H. Hirsemann, S. Jack, R. Klanner, A. Klyuev, H. Krueger, S. Lange, A. Marras, D. Mezza, A. Mozzanica, I. Perova, Q. Xia, B. Schmitt, J. Schwandt, I. Sheviakov, X. Shi, U. Trunk, and J. Zhang. The adaptive gain integrating pixel detector. *Journal of Instrumentation*, 11(C02066), 2016.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1764

Editor: The Dean of the Faculty of Science and Technology

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

Distribution: publications.uu.se
urn:nbn:se:uu:diva-372987

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019