Phasing using high intensity free-electron laser radiation

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

vorgelegt von

Lorenzo Galli

Hamburg, 2014

Folgende Gutachter empfehlen die Annahme:

der Dissertation:	Prof. Dr. Henry Chapman
	Prof. Dr. Andrea Cavalleri
der Disputation:	Prof. Dr. Christian Betzel
	Prof. Dr. Henry Chapman
	Prof. Dr. Gerhard Gruebel
	Prof. Dr. Henning Moritz
	Prof. Dr. Michael Ruebhausen
Datum der Disputation:	18 Dec. 2014

Author email: lorenzo.galli2312@gmail.com

To my parents

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

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

Hamburg, den

Acknowledgments

I am indebted to my supervisor, Henry Chapman, for accepting me in his group. I thank him for the trust he had in me, for giving me freedom to purse my own ideas and for his guidance. I will never forget the one-to-one meetings during which stream of ideas were flowing like a river, and experiments were being sketched on pieces of paper. His scientific vision is simply amazing.

I am grateful to all the members of the Coherent Imaging division: I spent wonderful years learning and collaborating with beautiful minds in an exquisitely international environment. In particular, I would like to thank Thomas White, for introducing me to the new research field when I joined the group, for being a constant source of answers and inspiration, and for his precious comments to this thesis. Thank to Miriam Barthelmeß for her essential help in the lab, to Mauro Prasciolu and Francesco Stellato for creating a beautiful working environment and for their time spent helping me in the lab or chatting together with a cup of coffee.

I will always be grateful to Irmtraud Kleine, for her prompt administrative support and assistance. She is the skeleton of the group.

I would like to thank my collaborators, in particular Ilme Schlichting and Thomas Barends from the MPI for Medical Research in Heidelberg, Carl Caleman from Uppsala University, Max Nanao from the EMBL in Grenoble, and Sang-Kil Son from the CFEL theory division, for the many successful collaborations, and the scientific support and guide.

Finally, I would like to thank all friends scattered all around Europe, for the long distance mutual backing and encouragement during the new experience abroad, my girlfriend Lisa and my parents for their constant love and presence.

Abstract

X-ray free-electron lasers (XFELs) provide extremely bright X-ray pulses of femtosecond duration, that promise to revolutionize structural biology, as they can be used to collect diffraction data from micrometer-sized crystals while outrunning radiation damage. The high fluence of the XFEL pulses induces severe electronic radiation damage to the sample, and especially the heavy atoms are strongly ionized by the X-ray radiation. The aim of this thesis is to test if it may be possible to use this specific radiation damage effect as a new approach to phasing.

By simulating serial femtosecond crystallography experiments at different X-ray fluence conditions, I describe that it is possible to use a Radiation damage-Induced Phasing scheme to retrieve the coordinates of the heavy atoms, and to correctly phase the model structure. Experimental data showed an effective reduction of the scattering power of a heavy atom inside a chemically modified protein, and of the sulfurs in a native protein. From the analysis of these experimental data, quantitative methods have been developed to retrieve information about the effective ionization of the damaged atomic species. The same analysis demonstrated that statistical methods can be used to sort the collected diffraction patterns, according to photon flux impinging on the sample. The knowledge of the real experimental conditions is critical for the success of high intensity phasing technique.

Abstract

Freie Elektronen Röntgenlaser (X-ray Free-electron Laser - XFELs) liefern extrem helle Röntgenpulse von Femtosekunden Dauer, die die Strukturbiologie zu revolutionieren versprechen, weil diese verwendet werden können, um Beugungsdaten von mikrometergroßen Kristalle zu sammeln bevor Strahlenschäden auftreten. Die hohe Photonen Fluenz der XFEL Impulse induziert schwere elektronischen Strahlenschäden an der Probe und vor allem die schweren Atome werden stark von der Röntgenstrahlung ionisiert. Das Ziel dieser Arbeit ist es, zu testen, ob es möglich sein kann, diese spezifische Strahlenschäden Effekte zur Phasierung zu verwenden.

Durch die Simulation von seriellen Femtosekunden Kristallographie Experimenten bei unterschiedlichen Röntgenphotonen Fluenz Bedingungen zeige ich, dass es möglich ist, ein Schema des Phasierens mit Strahlenschäden (Radiation damage-Induced Phasing - RIP) zu verwenden, um die Koordinaten der Schweratome zu bestimmen, und so im Folgenden die Modellstruktur korrekt zu phasieren. Experimentelle Daten zeigten eine effektive Verringerung des Streukraft eines schweren Atoms innerhalb eines chemisch modifizierten Proteins und von Schwefel in einem nativen Protein. Aus der Analyse dieser experimentellen Daten wurden quantitative Methoden entwickelt, um Informationen über die effektive Ionisierung der beschädigten Atomarten abzurufen. Die gleiche Analyse zeigte, dass statistische Verfahren verwendet werden können, um die gesammelten Beugungsmuster nach Photonenfluss der auf die Probe trifft zu sortieren. Die Kenntnis der realen Versuchsbedingungen ist notwendig für den Erfolg der Technik des Phasierens mit hohe Röntgenintensität (high intensity phasing - HIP).

List of papers

This thesis is based on my work as research assistant at the Center for Free-Electron Laser Science, Hamburg, within the Coherent Imaging Division and as member of the Graduate College "GRK 1355" at the University of Hamburg.

The results presented in this manuscript are mostly based on the following papers and on unpublished work.

My contributions spanned from simulation and data analysis, to sample characterization, sample delivery, and data collection.

1. **L. Galli**, T. R. M. Barends, S-K. Son, T. A. White, A. Barty, S. Botha, C. Caleman, R. B. Doak, K. Nass, M. Nanao, R. L. Shoeman, N. Timneanu, R. Santra, I. Schlichting and H. N. Chapman. *Phasing using high X-ray intensity*, to be submitted (2014).

2. **L. Galli**, S-K. Son, T. A. White, R. Santra, H. N. Chapman and M. Nanao. *Towards RIP with Free-Electron Laser radiation*, Journal of Synchrotron Radiation (accepted, 2014).

List of additional papers

The following list of publications is presented to show my additional scientific contributions during the graduate period. The topics that appear in these papers are not (or are only marginally) part of this dissertation. Papers in preparation are not included in this list.

1. T. A. White, A. Barty, M. Metz, D. Oberthur, C. Gati, **L. Galli**, O. Yefanov and H. N. Chapman. *Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers?*, under review (2014).

2. K. R. Beyerlein, C. Jooss, A. Barty, R. Bean, S. Boutet, S. S. Dhesi, R. B. Doak, M. Först, **L. Galli**, R. Kirian, J. Kozak, M. Lang, R. Mankowsky, M. Messerschmidt, J. C. H. Spence, D. Wang, U. Weierstall, T. A. White, G. J. Williams, O. Yefanov, N. A. Zatsepin, A. Cavalleri, and Henry N. Chapman. *Trace Phase Detection and Strain Characterization from Serial XFEL Crystallography of a Pr0.5Ca0.5MnO3 Powder*, Proceedings of the European Powder Diffraction Conference (EPDIC) (accepted, 2014).

3. A. D. Rath, N. Timneanu, F. R. N. C. Maia, J. Bielecki, H. Fleckenstein, B. Iwan, M. Svenda, D. Hasse, G. Carlsson, D. Westphal, K. Mühlig, M. Hantke, T. Ekeberg, M. M. Seibert, A. Zani, M. Liang, F. Stellato, R. Kirian, R. Bean, A. Barty, **L. Galli**, K. Nass, M. Barthelmess, A. V. Martin, A. Aquila, S. Toleikis, R. Treusch, S. Roling, M. Wöstmann, H. Zacharias, H. N. Chapman, S. Bajt, D. DePonte, J. Hajdu, and Jakob Andreasson. *Explosion dynamics of sucrose nanospheres monitored by time of flight spectrometry and coherent diffractive imaging at the split-and-delay beam line of the FLASH soft X-ray laser*, Optics Express (accepted, 2014).

4. C. Kupitz, S. Basu, I. Grotjohann, R. Fromme, N. Zatzepin, K. Rendek, M. S. Hunter, R. L. Shoeman, T. A. White, D. Wang, D. James, J. Yang, D. E. Cobb, B. Reeder, R. G. Sierra, H. Liu, A. Barty, A. L. Aquila, D. Deponte, R.

A. Kirian, S. Bari, J. J. Bergkamp, K. R. Beyerlein, M. J. Bogan, C. Caleman, T. Chao, C. E. Conrad, K. M. Davis, H. Fleckenstein, **L. Galli**, S. P. Hau-Riege, S. Kassemeyer, H. Laksmono, M. Liang, L. Lomb, S. Marchesini, A. V. Martin, M. Messerschmidt, D. Milathianaki, K. Nass, A. Ros, S. Roy-Chowdhury, K. Schmidt, M. Seibert, J. Steinbrener, F. Stellato, L. Yan, C. Yoon, T. A. Moore, A. L. Moore, Y. Pushkar, G. J. Williams, S. Boutet, R. B. Doak, U. Weierstall, M. Frank, H. N. Chapman, J. C. H. Spence and P. Fromme. *Serial time-resolved femtosecond crystallography of Photosystem II using a femtosecond X-ray laser*, Naure **513**, 261–265 (2014).

5. F. Stellato, D. Obertuer, M. Liang, R. Bean, C. Gati, O. Yefanov, A. Barty, A. Buckhardt, P. Fischer, **L. Galli**, R.A. Kirian, et al. *Room-temperature macromolecular serial crystallography using synchrotron radiation*, IUCrJ **1**, 204-212 (2014).

6. L. C. Johansson, D. Arnlund, G. Katona, T. A. White, A. Barty, D. P. DePonte, R. L. Shoeman, C. Wickstrand, A. Sharma, G. J. Williams A. Aquila, M. J. Bogan, C. Caleman, J. Davidsson, R. B. Doak, M. Frank, R. Fromme, **L. Galli**, I. Grotjohann, M. S. Hunter, S. Kassemeyer, R. A. Kirian, C. Kupitz, M. Liang, L. Lomb, E. Malmerberg, A. V. Martin, M. Messerschmidt, K. Nass, L. Redecke, M. M. Seibert, J. Sjöhamn, J. Steinbrener, F. Stellato, D. Wang, W. Y. Wahlgren, U. Weierstall, S. Westenhoff, N. A. Zatsepin, S. Boutet, J. C. H. Spence, I. Schlichting, H. N. Chapman, P. Fromme, and R. Neutze. *Structure of a photosynthetic reaction centre determined by serial femtosecond crystallography*, Nature Communications **4**: 2911 (2013).

7. L. Redecke, K. Nass, D. P. DePonte, T. A. White, D. Rehders, A. Barty, F. Stellato, M. Liang, T. R. M. Barends, S. Boutet, G. J. Williams, M. Messerschmidt, M. M. Seibert, A. Aquila, D. Arnlund , S. Bajt, T. Barth, M. J. Bogan, C. Caleman, T-C. Chao, R. B. Doak, H. Fleckenstein, M. Frank, R. Fromme, **L. Galli**, I. Grotjohann, M. S. Hunter, L. C. Johansson, S. Kassemeyer, G. Katona, R. A. Kirian, R. Koopmann, C. Kupitz, L. Lomb, A. V. Martin, S. Mogk, R. Neutze, R. L. Shoeman, J. Steinbrener, N. Timneanu, D. Wang, U. Weierstall, N. A. Zatsepin, J. C. H. Spence, P. Fromme, S. Schlichting, M. Duszenko, C. Betzel, and H. N. Chapman. *Natively inhibited trypanosoma brucei cathepsin B structure determined by using an X-ray laser*, Science **339**, 227–230 (2013).

Contents

Chapter 1

X-ray radiation

The descriptive nature of the physical world changed considerably after the discovery of X-ray radiation in the late $19th$ century. The term X-ray denotes a particular range of electromagnetic radiation, having energies between 100 eV and 100 keV. This range is loosely separated into *hard X-rays* (with photon energies above $2 - 5$ keV) and *soft X-rays*; hard X-rays have a higher penetration depth, while soft X-rays are easily absorbed in air and by any material. X-ray photons interact strongly with atoms, with a probability that can be roughly approximated to Z^3/E^3, where Z is the atomic number and E is the photon energy. This property makes them an ideal probe for medical imaging (radiography or tomography, for example). The X-ray spectrum presents sharp discontinuities at energies corresponding to electronic transitions of an atom called absorption edges. As X-ray photons carry enough energy to ionize atoms, they interact disruptively with matter. High X-ray doses are considered harmful for living tissues and, on the atomic level, they can induce damage and disrupt many molecular bonds (as explained in section 1.3). Due to their very short wavelength (10^3 times shorter or more than visible light), X-rays are the most widely used tools for acquiring high resolution images from structures which are invisible for optical microscopes; in particular, hard X-rays have wavelengths comparable to the length of atomic bonds, so they are also used to determine the positions of atoms in solids through the collection of diffraction images, with a technique called X-ray crystallography.

This first chapter of this thesis describes the principles and the consequences of X-ray diffraction, with particular attention to the problems of radiation damage and to the *de novo* methods of structural determination. In the first half of the chapter, the basic theory of X-ray scattering and the fundamental laws of crystallography are defined. The second half of the chapter deals with the phase problems in crystallog-

raphy (that is, hot to synthesize 3D images of the electron density from measured Fourier intensities) and with the effects of radiation damage.

1.1 Scattering of X-rays

X-ray diffraction is the result of the interaction of the electromagnetic radiation with the electrons of the atoms in the crystal. Since the dielectric polarizability is several orders of magnitude higher than the diamagnetic susceptibility, the electromagnetic interaction happens mainly through the oscillating electric field. The dielectric polarizability (α) is related to the refractive index n of a material of density N through the Clausius Mosotti equation:

$$N\alpha = 3\frac{n^2-1}{n^2+2}$$

and at very high temporal frequency it becomes very small, so the material becomes transparent to the radiation (see for example Chapter 32 in [1]). The direct consequence of this fact is that it is practically not possible to achieve atomic resolution using hard X-ray refractive lenses.

Dispersion (intended generally as the dependence of a physical property with frequency) and absorption also exist for X-rays: in particular X-ray absorption involves high-energy electronic transitions in the atomic core levels, and ionization. In a protein crystal this can lead to bond breaking and generation of free radicals, which degrades the sample quality and fixes boundaries to the data collection times.

If the interaction between the X-rays and the electrons is considered on a microscopic level, it can be seen as an induction of oscillatory motions of the electrons. As charges accelerate, they emit electromagnetic waves of the same frequency, while the phase difference between the scattered waves gives rise to diffraction phenomena.

1.1.1 Definition of Bravais and reciprocal lattice

1.1.1.1 The Bravais lattice

The lattice is a fundamental concept in the description of any crystalline solid. It is defined in $3D$ as an array of discrete points, which can be described using a position vector $\mathbf{R}$ of the form:

$$\mathbf{R} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}\,,$$

where n_1, n_2, n_3 are integers and $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are vectors not all in the same plane. These vectors are also called the basis vectors of the cell, and the volume of space that can fill the entire lattice (with no overlap or voids) when translated through some subset of the vector is called "unit cell". Since the choice of the basis vector has no particular restrictions, different types of unit cells can be chosen to define the same lattice. The cells containing only a single lattice point are called "primitive", while those containing multiple lattice points are referred to as "multiple" or "centered" cells. The directions specified by the $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ vectors are the crystallographic axes, while the angles between them are indicated by α, β, and γ, with α opposing $\mathbf{a}$, β opposing $\mathbf{b}$, and γ opposing $\mathbf{c}$. The convolution between the lattice and the content of the unit cell is commonly referred to as the crystal structure. Only particular geometrical figures can fill the entire space with no voids, so there exist only a finite number of possible unit cell symmetries. Crystal lattices can be classified according to the set of rigid operations (translations, rotations, reflections, and inversions) that transform the lattice onto itself. The set of nontraslational operations that leaves a point of the lattice fixed defines the crystal system, or family, of the lattice. In three dimensions, the possible lattices can be categorized in 7 crystal systems or families, reproduced in figure 1.1. To each crystal system a primitive cell can always be associated, but other types of lattices exist based on non-primitive lattices, which are hard to express as primitive cells. Despite the total possible combinations of symmetry operations and centering is 42, these can be reduced to only 14 independent space lattices, called Bravais lattices. The set of rotation and reflection operations that do not have translational component and which leave one point fixed (called in general point group operations), instead, defines 32 point groups. Finally, the set of symmetry operations that take a three-dimensional periodic object onto itself gives rise to 230 crystallographic space groups [2].

1.1.1.2 The reciprocal lattice

The reciprocal lattice is a Bravais lattice defined as the set of all wave vectors $\mathbf{K}$ (with $|\mathbf{K}| = 2\pi/\lambda$) that yield plane waves with the periodicity of a Bravais lattice. Analytically, $\mathbf{K}$ belong to the reciprocal lattice if

$$e^{i\mathbf{K}\cdot(\mathbf{r}+\mathbf{R})} = e^{i\mathbf{K}\cdot\mathbf{r}}$$

holds for any vector $\mathbf{r}$ and for any $\mathbf{R}$ defining the Bravais lattice [3]. The same relation can be rewritten as:

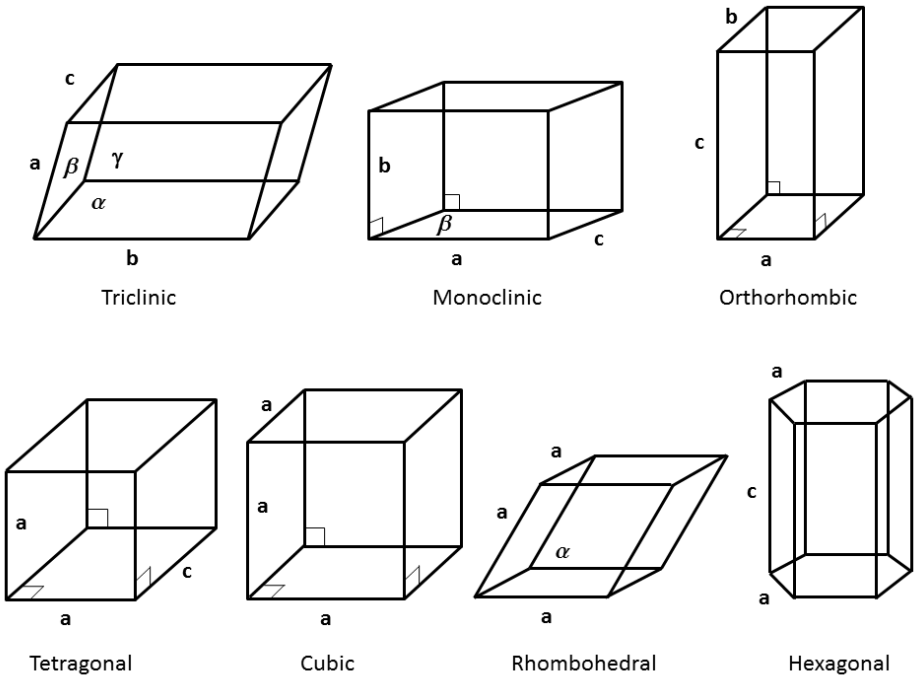

Figure 1.1: Sketch of the 7 lattice types.

$$e^{i\mathbf{K}\cdot\mathbf{R}} = 1\,. \tag{1.1}$$

1.1.1.3 Lattice planes and Miller indices

Given a lattice, a lattice plane is defined to be a plane generated by 3 non collinear lattice points. A set of parallel and equally spaced lattice planes will contain all the points of a Bravais lattice, and it can be described by a reciprocal lattice vector $\mathbf{K}$, normal to the planes and with length $n = 1/d$, where d is the distance between two consecutive planes (this follows immediately from the definition of reciprocal lattice). The coordinates of the shortest reciprocal lattice vector describing the plane are called the Miller indices of the plane, and are commonly expressed as three integer numbers (h, k, l) given a reciprocal lattice vector of the form:

$$\mathbf{K} = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3\,.$$

1.1.2 Determination of crystal structures by X-ray diffraction

1.1.2.1 Bragg and von Laue equations

The typical interatomic distances between atoms in solids are on the order of $1-2\,\text{Å}$. If one wants to investigate the atomic structure using an electromagnetic probe, must therefore utilize a wavelength at least that short, corresponding to an energy of $^{hc}/_{\lambda} \simeq 6 - 12\,\text{keV}$, which means in the X-ray region.

In crystalline materials, for certain sharply defined wavelengths and particular incident directions, intense scattered peaks can be observed. This fact was first observed in 1912 by W. Friedrich, P. Knipping and M. Laue [4], and explained later by W.H. and W.L. Bragg [5], by describing a crystal as made of sets of parallel planes of ions, spaced a distant d apart. The conditions for the appearance of an intense scattered peak are: the reflected wave has to be specular to the incident wave, and successive planes scatter in phase. For rays to interfere constructively, the path length difference between two consecutive planes must be a multiple of the incoming wavelength:

$$n\lambda = 2d\sin(\theta) \tag{1.2}$$

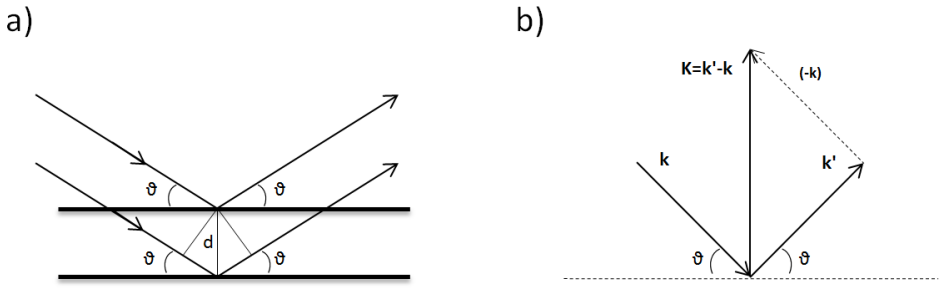

Figure 1.2: **a)** A Bragg reflection from a particular family of lattice planes, separated by a distance d. The incident and reflected rays are drawn for two consecutive planes. **b)** Vectors satisfying the Von Laue condition.

where θ is the incident angle, as drawn in figure 1.2a. Equation 1.2 is commonly referred to as Bragg's Law, and the scattered peak is named "Bragg peak" or "Bragg reflection".

The same equation can be derived without assuming specular reflections from idealized planes, but considering instead the crystal as form of identical objects occupying the Bravais lattice sites $\mathbf{R}$. Constructive interference between scattered waves from two of those objects can be observed, for an incoming wavelength λ, in a direction $\mathbf{n}'$ satisfying the equation:

$$\mathbf{d} \cdot (\mathbf{n} - \mathbf{n}') = m\lambda \,,$$

where $\mathbf{n}$ is the direction of the incident radiation and $\mathbf{d}$ the distance vector between the objects. Substituting the wave vector $\mathbf{k} = \mathbf{n}/\lambda$ and generalizing the equation for an array of scatterers occupying the Bravais lattice:

$$\mathbf{R} \cdot (\mathbf{k} - \mathbf{k}') = m$$

which can be rewritten as:

$$e^{(\mathbf{k}-\mathbf{k}')i} = 1 \,.$$

By recalling the equation defining the reciprocal lattice (1.1) we arrive at the condition that the change in wave vector is a vector of the reciprocal lattice ($\mathbf{K} = \mathbf{k} - \mathbf{k}'$).

Since $\mathbf{k}$ and $\mathbf{k}'$ have the same magnitude, this also means that:

$$\mathbf{k} \cdot \hat{\mathbf{K}} = 1/2|\mathbf{K}| \,,$$

which is also called Laue condition. The $\mathbf{k}$ vector also defines a set of planes per-

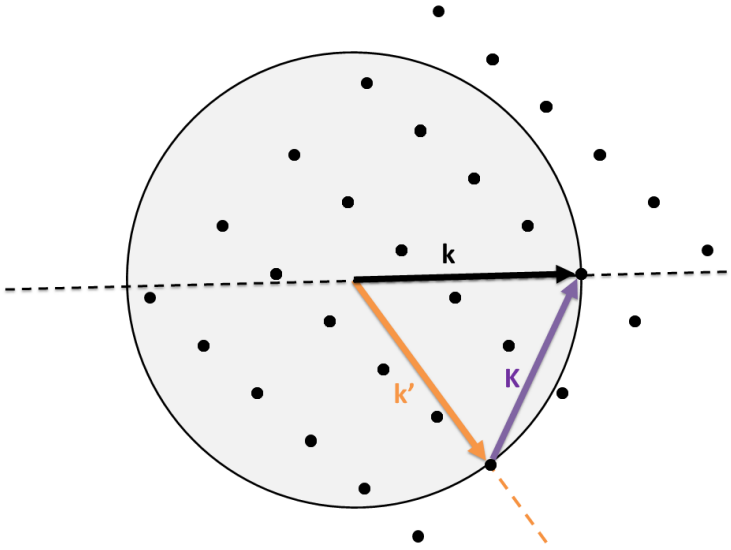

Figure 1.3: The Ewald construction. A sphere of radius $|\mathbf{k}|$ is drawn about the incident wave vector $\mathbf{k}$. Diffraction peaks will be observed in direction $\mathbf{k}'$, from the reciprocal lattice vector $\mathbf{K}$, if the vector lies on the surface of the sphere.

pendicular to the reciprocal lattice vector $\mathbf{K}$, which are called Bragg planes. From the relation between vectors of the reciprocal lattice and families of Bravais lattice planes (see subsection 1.1.1.3), and because the scattering is supposed elastic (so incident and scattered wave vectors have the same magnitude), $\mathbf{k}$ and $\mathbf{k}'$ make the same angle θ with the plane perpendicular to $\mathbf{K}$ as shown in figure 1.2b [3]. Therefore the Bragg and von Laue formulations are equivalent.

1.1.2.2 The Ewald construction

Since the set of Bragg planes is a discrete family, for a fixed incident direction and wave vector magnitude (i.e. X-ray energy) the probability of fulfilling the diffraction condition will be very low. In order to search experimentally for Bragg peaks, either the orientation of the crystal to the beam or the X-ray energy has to be modified. A simple geometric construction was conceived by Paul Peter Ewald to easily visualize those methods, here depicted in figure 1.3: given an incident wave vector $\mathbf{k}$, a sphere (also called Ewald sphere) of radius k is drawn about $\mathbf{k}$. The reciprocal lattice points are drawn as well. Diffraction peaks will be observed only if the surface of the sphere intersects a reciprocal lattice point.

To bring lattice planes to diffract, experimentally one has then to release the

constrains on **k**, either by changing the incident wavelength, or by rotating the crystal, which corresponds to rotating the reciprocal lattice. The first of these experimental approaches is called the Laue method, and consists of continuously changing the X-ray wavelength within a relatively broad range. The second method is instead the most widely used in crystallography, and it is known as the rotating-crystal method. Standard diffractometers use goniostats (or goniometers) to rotate the crystal, usually mounted on a cryoloop. Complete sets of diffraction data are collected by sampling the entire asymmetric unit, by rotating the crystal around one or more axes. Each diffraction image is recorded while the crystal is rotated by a small angle, generally 0.1° to 1.5°.

Another possible experimental method is the powder (or Debye-Scherrer) method. In this case the axis of rotation is varied over all possible orientations by using a sample in the form of crystalline powder: because the crystals are randomly oriented, the diffraction pattern will be the combination of all the diffractions from the single crystals. In this case the reciprocal lattice in the Ewald construction can be represented by a family of spheres of radius K, and the Ewald sphere will intersect the lattice in circles.

1.1.2.3 Finite crystals and imperfections

The infinite lattice is a useful idealization to describe mathematically the crystal diffraction, but does not correspond to the reality, where the physical crystal only fills up a finite portion of the space. Furthermore, the atoms or molecules when forming a crystal do not arrange themselves in a perfect 3-dimensional array, because of impurities and energy minimization effects (such as surface effects, as showed for example in [6]). These imperfections contribute to the formation of misaligned domains, so that a real single crystal is rather a mosaic crystal, composed of many domains aligned to within few tenths of a degree. The misalignment of the individual domains is described as the mosaicity of the crystal. Each domain will diffract at a slightly different orientation, so the single Bragg reflection will fall at a slightly different but possibly overlapping position on the detector, increasing and deforming the shape of the Bragg peak.

1.1.2.4 Reflection partiality

If a finite crystal having random orientation is exposed to an X-ray radiation, and a snapshot image is taken without the possibility of changing the experimental condi-

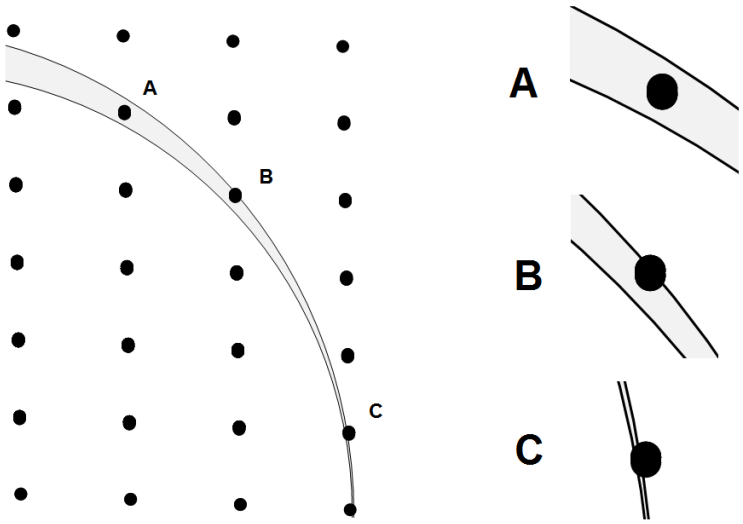

Figure 1.4: Section of the reciprocal space showing the Ewald sphere intersecting few reciprocal lattice points. B and C are only partially recorded.

tions, it can be expected that the reflections in the diffraction pattern would be only partially recorded. Furthermore, X-ray beams are generally neither monochromatic nor perfectly collimated (which means they usually have a small bandwidth and convergence angle), so the Ewald sphere assumes the form of an imperfect annulus. Figure 1.4 shows a possible experimental situation, where the reciprocal lattice points are drawn with a finite width due to possible crystal imperfections. Some of the lattice points in diffracting condition do not fully overlap with the Ewald sphere, so the diffracted intensity will contain partial Bragg reflection. Partially recorded reflections can be handled with different methods [7], knowing the experimental geometry and taking advantage of successive recorded patterns during a rotation series. In XFEL experiments, however, the jitter of the X-ray parameters and the unknown experimental geometry can bring reflection partiality to be one of the main source of errors. Nevertheless, White [8] described a method by which iterative post-refinements could be used to partially correct the merged data and improve the data quality.

1.1.2.5 The intensity of diffracted X-rays

The presence of a Bragg peak at a particular angle for a given crystal orientation provides information about the periodicity of the crystal, but does not give any

information about the real content of the Bravais lattice (i.e. the periodic molecular motif / the content of the unit cell). Such an information is contained partially in the intensity of each Bragg reflection.

Let us consider a monoatomic lattice containing n identical atoms in the unit cell, occupying the positions $\mathbf{d_1}, \mathbf{d_2}, .., \mathbf{d}_n$. The intensity of the radiation in a given Bragg peak will depend on the degree of interference of the scattered radiation from every atoms in the unit cell. If the Bragg peak is associated with a change in wave vector $\mathbf{K} = \mathbf{k}' - \mathbf{k}$, then the phase difference between X-rays scattered by two atoms in $\mathbf{d}_i$ and $\mathbf{d}_j$ will be $\mathbf{K} \cdot (\mathbf{d}_i - \mathbf{d}_j)$, so the amplitude of the two rays will differ by a factor $exp(i\mathbf{K} \cdot (\mathbf{d}_i - \mathbf{d}_j))$. The net scattering from the unit cell at the Bragg peak will then be proportional to the sum of all the atomic contributions:

$$\mathbf{F}_K = \sum_{j=1}^{n} e^{i\mathbf{K}\cdot(\mathbf{d}_i - \mathbf{d}_j)} \ . \tag{1.3}$$

This quantity is known as the geometrical structure factor, and the intensity of the Bragg peak is proportional to $|\mathbf{F}_K|^2$.

If the atoms in the basis are not identical, the structure factor in 1.3 assumes a more general form:

$$\mathbf{F}_K = \sum_{j=1}^{n} f_j(\mathbf{K}) e^{i\mathbf{K}\cdot(\mathbf{d}_i - \mathbf{d}_j)} \ , \tag{1.4}$$

where f_j is the atomic form factor, uniquely determined by the internal structure of the atoms occupying the position $\mathbf{d}_j$. The ideal atomic form factor is taken to be proportional to the Fourier transform of the electronic charge distribution of the corresponding atom, centered in $\mathbf{K}$:

$$f_j(\mathbf{K}) = -\frac{1}{e} \int d\mathbf{r} e^{i\mathbf{K}\cdot\mathbf{r}} \rho_j(\mathbf{r}) \ .$$

1.1.2.6 The Debye-Waller factor

The equations written in the previous subsection are valid under the assumption that all the species composing the crystal are fixed in absolutely rigid positions, which is only true in the ideal case where the atoms are at the absolute temperature of 0 K. In the real life, instead, the environment will donate thermal energy which makes the atoms vibrate about their equilibrium position, by a mean square atomic displacement $< u(0)^2 >$ which increases with the temperature. If the probability of

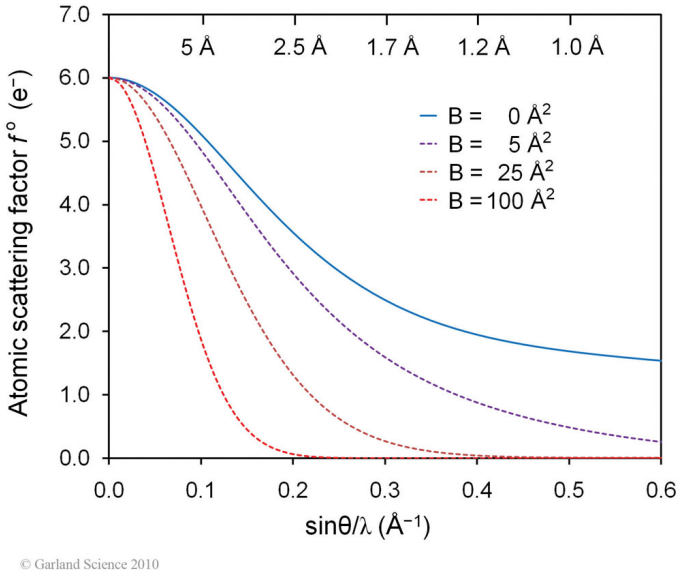

Figure 1.5: Scattering factor curves for carbon (z=6), calculated for different B-factors. A displacement of 1 Å corresponds to a B_{iso} of 79 Å^2. Reproduced with permissions from [10].

displacement by a quantity r' follows a simple Gaussian equation such as:

$$p(r') = \frac{1}{(2\pi U)^{1/2}} e^{\frac{-r'}{2U}} ,$$

with $U =< u(0)^2 >$, or if the displacement is small and with no preferred direction [9], then the resulting reduction of the atomic scattering factor, also called the Debye-Waller factor, is defined as:

$$T_s = e^{(-B_{iso}(\sin(\theta)/\lambda)^2)} . \tag{1.5}$$

The B_{iso} factor is called isotropic displacement parameter, or simply B-factor, and it is directly related to the mean square ionic displacement:

$$B_{iso} = 8\pi^2 < u(0)^2 > .$$

As a result, the atomic form factor will gain a Gaussian, wavelength- and angular-dependent term (see fig 1.5).

The atoms can also be displaced in the lattice because of disorder. Those two effects add phase differences in the scattering waves, which can be seen as a more

complicated attenuation factor. A similar effect, hard to distinguish from the B-factor, is the effect of partial occupancy. This happens when atoms or molecules (such as solvent or ligand molecules in a protein structure) are missing in some of the unit cells composing the macroscopic crystals, resulting in a general reduction of the scattering amplitude by an occupancy factor $n = [0 - 1]$.

In general, atomic bonds act as constraints, limiting the thermal movements along the bond direction. For this reason the B-factor is often defined as an anisotropic thermal factor (if the data quality and the quantity of information permit), represented by means of a 3-axis ellipsoid.

1.1.2.7 The Wilson plot

The atomic form factor, including the Debye-Waller factor of equation 1.5, can be then written as:

$$f_j^B = f_j e^{(-B_{iso}(\sin(\theta)/\lambda)^2)}$$

and the observed scattered intensity, in the presence of an isotropic thermal displacement, results:

$$I_{obs}^B \propto \sum_j^{atoms} (f_j^B)^2 = I_0 e^{(-B_{iso}(\sin(\theta)/\lambda)^2)} ,$$

where I_0 stands for the intensity on an absolute scale, in the case of a perfect crystal at 0 K temperature. Defining as k the scale factor between I_{obs}^B and I_0:

$$I_{obs}^B = k I_0 e^{(-2B_{iso}(\sin(\theta)/\lambda)^2)} ,$$

and taking the logarithm:

$$\ln \frac{I_{obs}^B}{I_0} = \ln k - 2B_{iso}(\sin(\theta)/\lambda)^2 . \tag{1.6}$$

Equation 1.6 has the general form of a straight line ($y = a + bx$), with the scale factor representing the intercept. This equation is often used as a check of the data quality, under the name of Wilson Plot. In protein crystals, however, the atomic positions are not distributed randomly at low resolution, and the Wilson plot looks generally as in figure 1.6.

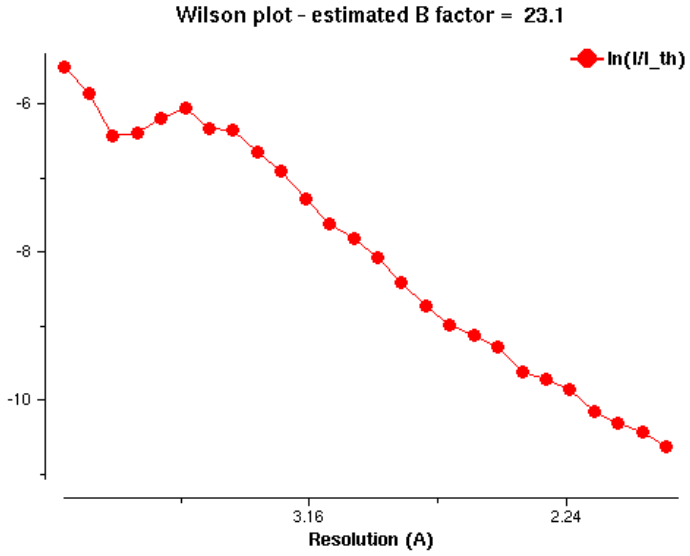

Figure 1.6: Example of Wilson plot used in macromolecular crystallography.

1.1.2.8 Friedel's Law

The Miller indices (hkl) and $(\bar{h}\bar{k}\bar{l})$ are defined with vectors having the same magnitude and direction, but opposite orientations. The families of planes described will then be the same, and so will be the structure factor. This statement is known as Friedel's Law, a property of the Fourier transform of a real-valued function, and has important consequences in crystallography. In particular, the squared amplitude $|\mathbf{F}|^2$ is centrosymmetric:

$$|\mathbf{F}(hkl)|^2 = |\mathbf{F}(\bar{h}\bar{k}\bar{l})|^2 \,,$$

and the phase ϕ of F is antisymmetric:

$$\phi(hkl) = -\phi(\bar{h}\bar{k}\bar{l}).$$

The pair of reflections hkl and $\bar{h}\bar{k}\bar{l}$ is called Friedel pair, while the two reflections are named Friedel mates.

1.1.2.9 Anomalous scattering factors

The classical description of elastic scattering was formulated by J.J. Thompson in 1906 and applies to free electrons, but it is also used with good approximation for the

bound electrons in atoms. In reality, the electrons occupying atomic orbitals must respond to the incident radiation according to their characteristic orbital frequency. In particular, the X-ray induced electron vibrations can resonate with the natural frequency of the bound electrons. This effect adds a perturbation to the free-electron-like "normal" factor f_0, which is usually described as a combination of two distinct "anomalous" terms f' and f'', so that:

$$f = f_0 + f' + if'' \,. \tag{1.7}$$

Here f is the true atomic scattering factor. Those anomalous terms are also called dispersive, since they - strongly - depend on the X-ray energy, while they are almost independent of the scattering angle because they derive from core electrons.

The anomalous scattering can be easily derived with a simple semi-classical model in which an atom is represented by a massive positively charged nucleus, surrounded by several electrons held at discrete binding energies, and an impinging electromagnetic wave described by an electric field $E_i \exp(-i\omega t)$. Treating a bound electron as a dampened oscillator with resonant frequency ω_s and dissipative frequency γ, the general dispersion term for the atomic scattering function is (see the appendix for a more complete treatment):

$$f = \frac{\omega^2}{(\omega^2 - \omega_s^2 + i\gamma\omega)} \,.$$

The most striking results are that the atomic structure factor displays a strong wavelength dependence, especially close to the resonance electron energy, where the imaginary dampening term prevent a discontinuity at ω_s, and that this latter term give an important out of phase contribution (the f'' term) to f.

A second repercussion of the imaginary term in the anomalous scattering is that the phase change of the scattered wave breaks the internal centrosymmetry within the collected dataset: under this condition the hkl and $\bar{h}\bar{k}\bar{l}$ reflection will have a phase shift and the intensity of the associated Bragg reflections will differ (see figure 1.7). Friedel's law, then, does not hold in the presence of anomalous scattering and this fact leads to important consequences in crystallography, as described in the next section.

1.1.2.10 The Patterson function

The Patterson function is based on the autocorrelation of the electron density map, and it is defined at any point **u** by a convolution integral over the unit cell volume,

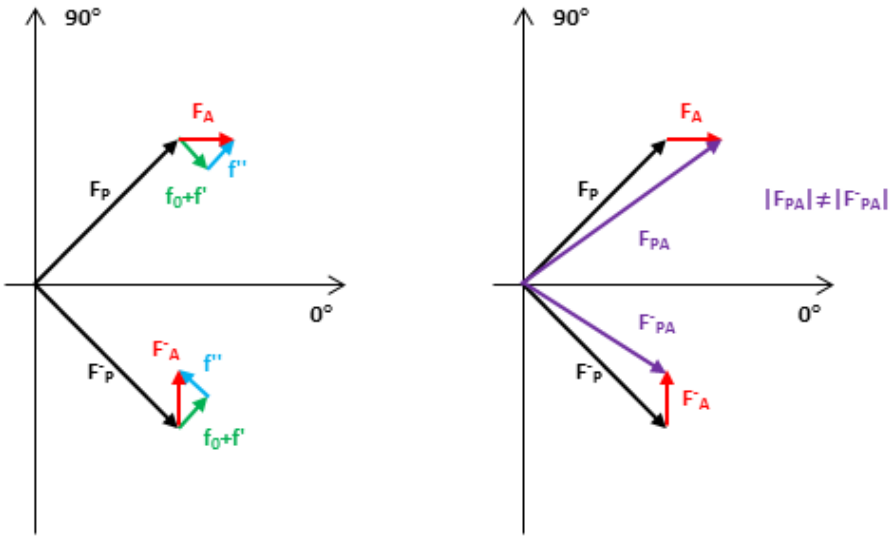

Figure 1.7: The breakdown of the Friedel's law due to anomalous scattering contribution. The F_P term represents the partial sum of normal contributions, while the other vectors are the contributions from anomalously scattering atoms. The Friedel pair is represented as a mirror copy of the *hkl* reflection, with a "-" superscript.

as:

$$P(\mathbf{u}) = \int_R \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})d\mathbf{r}\,, \tag{1.8}$$

where ρ is the electron density, and R represents the unit cell volume in the real space. The Patterson function has large values when both the electron density calculated at $\mathbf{r}$ and the translated density at $(\mathbf{r}+\mathbf{u})$ are high, that is when $\mathbf{u}$ is an interatomic distance vector. The map constructed with the Patterson function (called Patterson map) will then contain $N(N-1)$ peaks, if N is the number of atoms in the molecule, corresponding to the interatomic distances (not considering the "self-peaks" at $\mathbf{u}=0$). The construction of the map can be performed directly from the experimental intensities, without the knowledge of the phases: this follows from the Fourier convolution theorem, derived in the appendix.

The interpretation of the Patterson maps has a significant role in many experimental phasing techniques. In particular, they are often used for the determinations of marker atom positions (also called marker atom substructure) from isomorphous difference data (explained in the next section), in the determination of anomalously scattering atom positions, and during a molecular replacement experiment, to determine the orientation of the search model (see appendix 9.5).

1.2 Experimental phasing techniques in X-ray crystallography

1.2.1 Isomorphous replacement methods

The isomorphous replacement method is a general approach to *de novo* phasing, based on the determination of a marker atom substructure. Historically, isomorphous replacement was the phasing method adopted for the determination of the first three macromolecular structures: myoglobin [11, 12], hemoglobin [13] and the first enzyme, lysozyme [14].

This method relies on the possibility to have one or more isomorphous derivative crystals, the diffraction pattern of which can be subtracted from the experimental data on the native crystal, and the location of the source of the electronic difference (the marker substructure) can the obtained. Possible sources of difference may be introduced by adding heavy atoms into the native crystal, or by replacing one atom in the structure with one of another kind. Depending on how many derivatives are used, the method is called SIR (single isomorphous replacement) or MIR (multiple

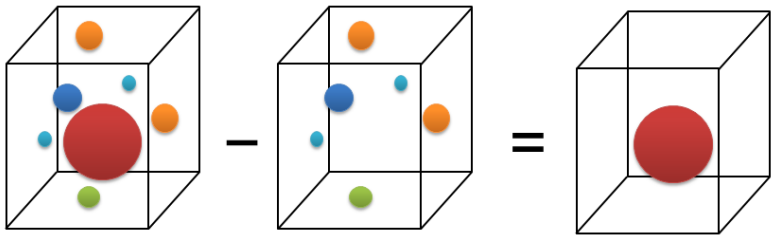

Figure 1.8: The determination of the marker atom substructure showed as a *gedankenexperiment* in real space. The first structure represents the derivative crystal, where the heavy atom is drawn as a big red sphere. The second crystal is instead the native. The light atoms cancel out and only the heavy marker atom is present in the difference crystal.

isomorphous replacement), and it can be combined with anomalous scattering (AS) methods, giving rise to SIRAS or MIRAS, respectively.

A stringent requirement for these methods to work is the isomorphism between derivatives and native crystals, i.e. the crystals should have the same internal structure and unit cell dimensions.

1.2.2 SAD and MAD phasing

The first experiment that proved the presence of anomalous X-ray scattering was performed in 1930 by Coster, Knol and Prins [15]: using a zinc blende (ZnS) sample and selecting the X-rays wavelength close to the absorption edge of Zn, they were able to demonstrate that Friedel's Law is not valid when the phase change is different for atoms in the same unit cell. It took however more than 20 years before Bijvoet and collaborators succeeded in using the deviations from Friedel symmetry to retrieve the absolute configuration of a small molecule [16]. Even after this breakthrough, anomalous methods were used mostly as aids to the more advanced techniques based on isomorphous differences, due to the limited choice of radiation sources. Only with the advent of synchrotron radiation sources it was possible to develop anomalous diffraction methods, which are now dominating among the *de novo* methods for the determination of crystal structures of biological molecules.

Not all the atoms composing a protein present a meaningful anomalous dispersion: for the typical X-ray energies used in crystallography, for example, light atoms such as H, C, N, and O have very low anomalous scattering, which can be usually neglected, while heavy atoms (i.e. species with a large number of electrons) can

display a moderate dispersive scattering. Phase information from measurements of anomalous diffraction can be derived by exploiting the interference between scattering from anomalous centers and that from the other atoms. The impact of each anomalous scatterer (R) on diffraction measurements can be evaluated calculating the contribution to the total diffraction as a sum of the components due to the total scattering factor, as expressed in equation 1.7:

$$\mathbf{F}_{A_R} = \mathbf{F}^0_{A_R} + \mathbf{F}'_{A_R} + i\mathbf{F}''_{A_R} = \left[1 + (f'/f^0) + i(f''/f^0)\right] \mathbf{F}^0_{A_R} \, .$$

In the presence of a single kind of anomalous scatterer, the total diffraction measurements associated to a particular Bragg reflection $\mathbf{h}$ at a given wavelength λ are given by:

$$\mathbf{F}(\mathbf{h}) = \mathbf{F}_T(\mathbf{h}) + \sum_R \left[(f'/f^0) + i(f''/f^0)\right] \mathbf{F}^0_{A_R}(\mathbf{h}) \, ,$$

where $\mathbf{F}^0_T$ is the total wavelength-invariant contributions from the f_0 components of the scattering factor. The observable quantity in a diffraction experiment is the intensity, proportional to $|\mathbf{F}^\lambda(\mathbf{h})|^2$, while the phases φ^0 are lost. Squaring the previous equation and separating the known factors from the unknown variables, it is possible to obtain the Karle-Hendrickson equation:

$$\begin{aligned} |\mathbf{F}^\lambda(\pm\mathbf{h})|^2 = \quad & |\mathbf{F}^0_T|^2 + \; a(\lambda)|\mathbf{F}^0_A|^2 + b(\lambda)|\mathbf{F}^0_T||\mathbf{F}^0_A| \cos(\varphi^0_T - \varphi^0_A) \pm \\ & c(\lambda)|\mathbf{F}^0_T||\mathbf{F}^0_A| \sin(\varphi^0_T - \varphi^0_A) \, . \end{aligned} \tag{1.9}$$

Here , $\mathbf{F}^0_A$ is the scattering part contributed solely by the normal scattering of the anomalous centers, and the a, b, c coefficients are defined by:

$$\begin{aligned} a(\lambda) &= (f'^2 + f''^2)/f_0^2 \\ b(\lambda) &= 2f'/f'_0 \\ c(\lambda) &= 2f''/f'_0 \, . \end{aligned}$$

The reflections $+\mathbf{h}$ and $-\mathbf{h}$ are Friedel mates, and the difference

$$\Delta F_{\pm\mathbf{h}} = |\mathbf{F}^\lambda(\mathbf{h})| - |\mathbf{F}^\lambda(-\mathbf{h})|$$

between the moduli of the Friedel mates (or of the rotational symmetry equivalents) is called Bijvoet difference. The difference between structure factor amplitudes at

two different wavelengths:

$$\Delta F_{\Delta\lambda} = |\mathbf{F}^{\lambda_1}| - |\mathbf{F}^{\lambda_2}|$$

with $|\mathbf{F}^{\lambda}| = \left(|\mathbf{F}^{\lambda}(\mathbf{h})| + |\mathbf{F}^{\lambda}(-\mathbf{h})|\right)/2$ is designated instead as the dispersive difference.

It can be seen from equation 1.9 that the Bijvoet difference depends on $\sin(\Delta\varphi = \varphi_T^0 - \varphi_A^0)$ and on $f''(\lambda)$, while the dispersive difference depends on $\cos(\Delta\varphi)$ and on $|f'(\lambda_1) - f'(\lambda_2)|$. So they provide orthogonal phase informations and they are complementary. Knowing $a(\lambda), b(\lambda), c(\lambda)$ from the evaluation of the anomalous coefficients, a set of equation of the form of 1.9 can be solved for the desired unknowns $|F_T^0|$, $|F_A^0|$, and $\Delta\varphi$. This experimental method is known under the name of MAD, and it assumes that multiple data are collected at (at least) two different wavelengths, chosen in order to maximize the dispersive differences.

A correct phase determination can also be achieved using the Bijvoet difference alone, i.e. with a single wavelength experiment (SAD). In this case the system of equations isn't complete, and in general the solution for the phase angle is ambiguous, as sketched in figure 1.9. This ambiguity can be nevertheless overcome with, for example, density modification techniques [17].

The expected scattering ratio, proportional to the Bijvoet difference, can be estimated, in the case of only one kind of anomalous scatterer and for zero scattering angle, using the equation proposed by Hendrickson and Teeter [18]:

$$\frac{< \Delta F >}{< F >} = \sqrt{2}\frac{\sqrt{N_A} f_A''}{\sqrt{N_P} Z_{eff}}\,, \tag{1.10}$$

where N_A and N_P are respectively the number of heavy atoms and the total number of non-hydrogen atoms in the protein, and Z_{eff} the effective atomic number (~ 6.7 for non-hydrogen protein atoms).

1.3 Radiation damage

The diffraction processes considered in the previous sections are only of type elastic. In the X-ray range used in crystallography, however, the scattering cross section is generally orders of magnitude smaller than the absorption cross section, so energy-loss processes are much more frequent than the elastic scattering. As can be seen from figure 1.10, for a pure-carbon sample exposed to 6 keV photons, the scattering cross section is $2.9{\cdot}10^{-9}\,\mu\mathrm{m}^2/\mathrm{g}$, while the absorption cross section is $1.05{\cdot}10^{-7}\,\mu\mathrm{m}^2/\mathrm{g}$

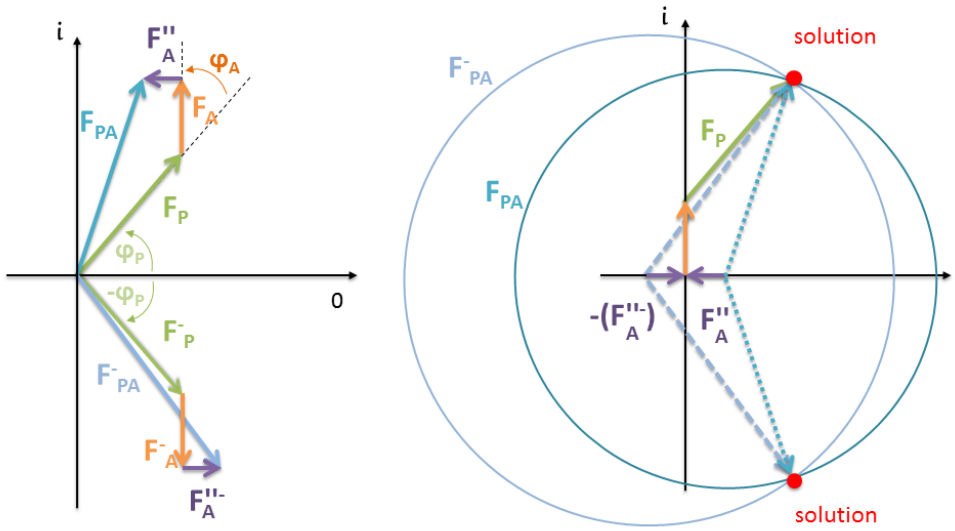

Figure 1.9: The left panel shows the the total structure factor ($\mathbf{F}_{PA}$) and its Bijvoet mate ($\mathbf{F}_{\overline{P}A}^{-}$), where the respective heavy atom contributions are divided in real and imaginary part. On the right side, the visual solution of the SAD phasing is sketched. The two possible solutions can be determined by drawing circles of radius $|\mathbf{F}_{PA}|$ and $|\mathbf{F}_{\overline{P}A}^{-}|$ from the corresponding origins.

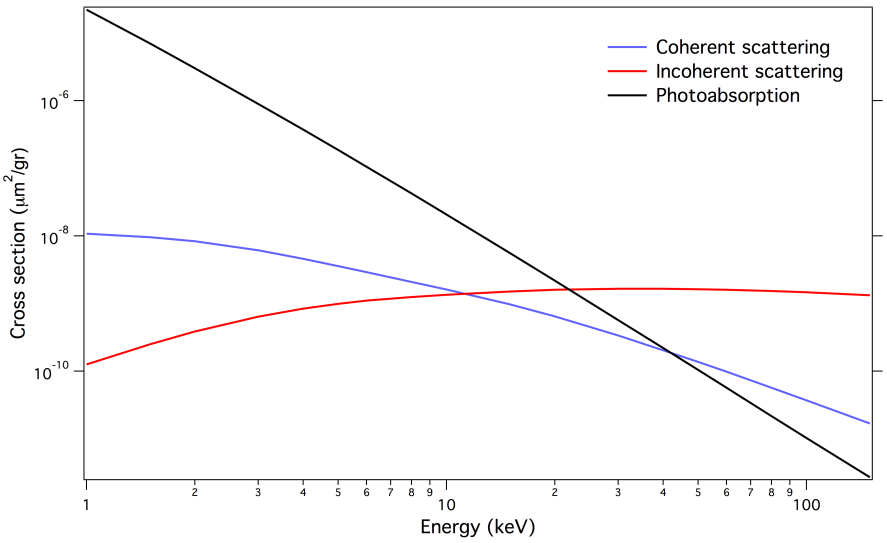

Figure 1.10: Atomic cross sections of carbon, for photoabsorption, elastic and inelastic scattering. Note that the X-ray energy are above the carbon K absorption edge

, meaning that for every scattered photon there are, on average, 36 photoionization events.

The inelastic scattering, responsible for the energy lost by an X-ray beam in a crystal through either photo-absorption processes or the inelastic (Compton) scattering, is the main source of radiation damage. At photon energies used for macromolecular crystallography (MX), the photoelectric effect has a much higher cross section and accounts for the majority of the energy deposited by the X-rays. Each of the created photoelectrons has enough energy to produce hundreds of other photoionization events (referred to as secondary damage), through either relaxation processes, such as the Auger decay, or electron-electron collisions, due to the short mean free path of the initially created photoelectrons. This cascade of ionization events can result in the formation of radical species in the crystal. In particular, biological crystals contain a percentage of solvent (20-80% in volume) which contributes to the creation of radicals. Some of the energy deposited in the sample is then converted into heat, resulting in a temperature rise in the sample.

Generally, the damage is manifest as an overall decrease in diffracted intensity and resolution. The measure of the energy loss is the "dose" received by the sample per unit of mass, and different effects of radiation damage on biological crystals have been found, since the early investigations in the '60s [19]. Here the description of

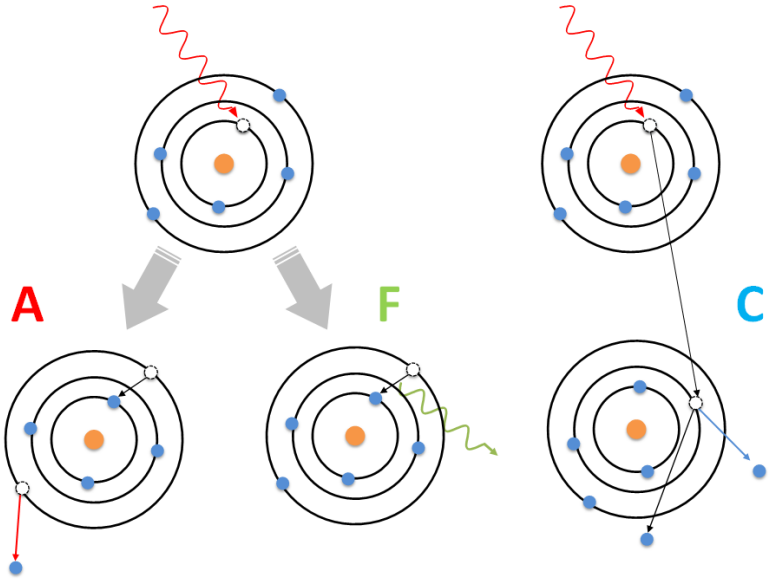

Figure 1.11: Cartoon of the main secondary damage processes: Auger decay (A), fluorescence (F), and electron-electron collision (C).

the dose as commonly used in MX and the effects of X-ray damage are explained.

1.3.1 The Dose

The dose is defined as the energy deposited in the sample per unit mass, and in the SI it is expressed in Gray (1 Gy = 1 J/kg). Since the dose quantifies the number of primary inelastic interactions per atom or molecule, it is directly related to the degree of radiation damage. The dose that a macromolecular crystal can tolerate before it loses half of its diffraction intensity was experimentally determined as 20 MGy [20]. A value commonly assumed as the experimental dose limit corresponds to a degradation of the average diffraction intensity by 70% of its initial value.

In the single atom case, the dose corresponds to the probability that an atom absorbs a photon, given by its cross section σ_a, multiplied by the X-ray fluence (energy per unit area):

$$D_{atom} = \frac{N_{ph} h\nu}{A} \sigma_A \,.$$

Expressing the equation for the energy deposited per unit mass, for a specimen of a

Element	Absorption cross section at 6 keV (Å^2)	Photon flux needed for single ionization (photons/μm^2)	Corresponding dose (GGy)
C	$2.2 \cdot 10^{-6}$	$4.5 \cdot 10^{13}$	103
N	$4.1 \cdot 10^{-6}$	$2.4 \cdot 10^{13}$	53.4
O	$7.2 \cdot 10^{-6}$	$1.3 \cdot 10^{13}$	30.5
S	$1.1 \cdot 10^{-4}$	$9.0 \cdot 10^{11}$	2.01
Fe	$7.6 \cdot 10^{-5}$	$1.3 \cdot 10^{12}$	2.86

Table 1.1: Absorption cross section of various elements at 6 keV, with the corresponding photon flux needed to induce a single ionization and the corresponding dose.

single atomic constituent with N_a atoms of mass m_a, and for a sample much thinner than the absorption length:

$$D = \frac{I_0 N_a}{m_a} \sigma_A \,,$$

with $I_0 = Nh\nu/A$. Under these assumptions, the dose is an atomic property, independent of the sample geometry or of the arrangement of atoms (with the exception of the atomic density).

At a photon energy of 6 keV, the atomic cross section varies between $10^{-14}\,\mu$m^2 for the light elements to $10^{-11}\,\mu$m^2 for heavier atoms [21], which results in a high penetration depth for X-rays into matter. Stated from another perspective, the photon flux needed for any atoms to absorb a single photon ranges between $4.5 \cdot 10^{13}$ photons/μm^2 for carbon to $1.3 \cdot 10^{12}$ photons/μm^2 for iron (see table 1.1). The corresponding doses are given by:

$$D_1 = h\nu \frac{N_A}{m_A} \,.$$

To stress the importance of the dose, in protein crystallography a dose of 30 MGy is often considered the highest tolerable for a cryocooled crystal, even though this dose is barely enough to ionize 0.06% of atoms of a pure-carbon sample exposed to 6 keV radiation. A widely used tool for computing the dose absorbed by a macromolecular crystal during an X-ray diffraction experiment, taking into account the sample geometry, the environment, and the absorption and attenuation, is RADDOSE [22].

1.3.2 Effects of radiation damage

The effect of the energy transfer from the X-rays into the sample, and consequently on the diffraction pattern, depends on the processes initiated by the photoionization. These processes depends on the exposure time and on the kinetic energy of the photoelectrons. As nicely illustrated by Chapman *et al.* [23], if we assume that no energy can flow out of the sample, and we consider time scales where the sample has reached thermal equilibrium, the temperature rise is given by the ratio between the dose and the heat capacity of the sample. Considering a sample with the heat capacity of water ($4800\,\mathrm{JkgK^{-1}}$) and a dose of $1\,\mathrm{MGy} = 10^6\,\mathrm{J/kg}$, the temperature rise will be of 208 K, while an X-ray dose of 1 GGy will heat the system up to about $200,000\,\mathrm{K}$. In synchrotron MX experiment, the exposure time is slow enough for this heat to be conducted away to the environment, by means of cryocooling systems or just the surrounding environment. An X-ray FEL pulse, however, can deliver doses of the order of 1 GGy in few tens of femtoseconds, creating a plasma that cools by expansion long after the pulse.

For a given instantaneous dose, the kinetic energy distribution of the produced photoelectrons has been found to be largely independent of the photon energy [24]. This approximation is best for samples consisting of light elements, such as C, N, and O. The photoelectrons generated from these light elements will have quite high energy, due to the low binding energy of their K shell electrons (294 eV for C). The generated core hole is predominantly filled by Auger decays, with decay times from 5 fs (for O) to 10 fs (for C). The emitted Auger electrons have a much lower kinetic energy, corresponding to an average velocity of about $100\,\mathrm{Åfs^{-1}}$. Photoelectrons and Auger electrons propagate through the sample and can cause an ionization cascade due to collisions with other atoms of the sample. It has been found that a single 5 keV photoelectron thermalizes in about 10 fs, producing around 10 core hole ionizations and a total of 240 ionizations within a range of about 100 nm [23]. Heavier atoms have higher inelastic cross sections, but also higher binding energy, so the energies of the photoelectrons emitted from these atoms are considerably lower than from light elements. The Auger decay, instead, competes with relaxation by fluorescence emission.

On a macroscopic scale, X-ray damage in MX is usually divided into two classes: global damage and specific damage. The former manifests as a loss of the measured reflection intensities, particularly at high resolution (few angstrom), as an expansion of the unit cell volume, as an increase of the thermal factor (B factor), and often as increase in mosaicity. Various metrics can be used to monitor this global damage, comparing the diffraction measurements at increased doses (see [25] for a review of

the metrics).

The specific structural damage is instead observed in particular covalent bonds, as a reproducible effect of the energy absorbed. Experiments have shown that disulphide bridges are particularly susceptible of X-ray damage [26, 27, 28], as well as C-S bonds in methionines [27], or bonds involving heavy atoms [29].

1.3.2.1 RIP phasing

About ten years ago [31] it was shown that specific X-ray damage could be used as a novel phasing method for native protein crystals. The method, called radiation induced phasing (RIP) utilizes the specific damage of X-ray-susceptible substructures, such as disulphide bridges of cystines, combined with a modified SIR workflow. This method is indeed similar to an isomorphous method, in which two (or more) datasets can be collected on the same crystal, and the first of this data is compared to the last one. In the presence of radiation damage, the two sets will show significant intensity differences, and the specificity of the damage to the susceptible chemical group can be thought as an isomorphic difference. In reality, a background of non-specific changes exists, such as a possible translation/rotation of the molecules in the crystal or an expansion of its unit cell, which introduce a non-isomorphism between the early collected data and the damaged one. Specific X-ray damage can be induced, for example, by breaking the S-S bonds in a molecule (see figure 1.12), with a short exposure to a highly ionizing radiation before the collection of the second dataset. In this way, the difference between the collected diffraction should be localised to the sulfurs, which can be localized with substructure determination programs (the experiment shown in this manuscript uses SHELXD [32]). The ionizing radiation can also be provided by an external source, such as an UV light. In particular, the UV energy can be chosen to match the absorption energy of the valence electrons involved in the cystine bonds, or to initiate indirect processes of radical formations.

Radiation damage always causes an overall decrease of the scattering power that is not taken into account during the scaling procedure, commonly adopted to bring two or more datasets on the same intensity scale (see for example the algorithms implemented in programs such as *Scaleit* [33] or *xscale* [34]). The RIP workflow compensates for the possible over-scaling of the damaged set by introducing a constant scale factor, k, and by performing parallel substructure determination processes with different values for k [35].

Like isomorphous replacement, RIP has the advantage of not being limited to wavelengths close to the absorption edge of the elements used as substructures;

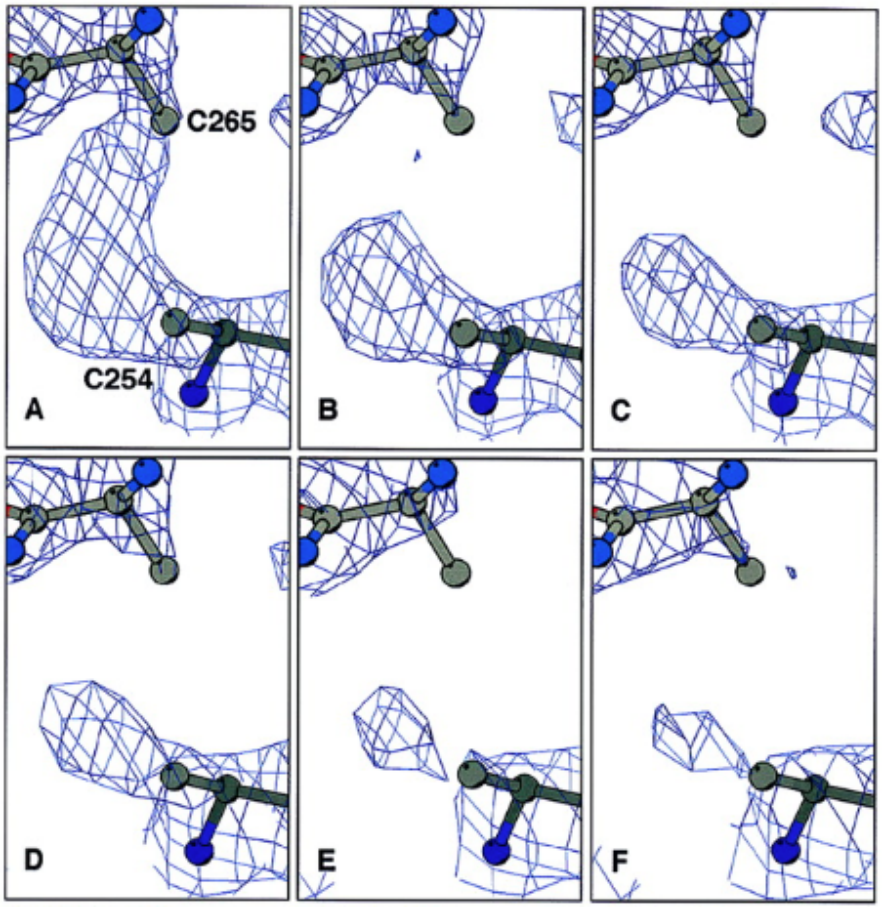

Figure 1.12: Sequential Fourier ($3Fo - 2Fc$) maps, showing the time course of cleavage of a disulfide bond in a protein crystal, exposed multiple times over an x-ray beam. Maps are contoured at $1.5\,\sigma$. Reproduced with permission from [30]. Copyright (2000) National Academy of Sciences, U.S.A.

furthermore, it does not require a derivative crystal, or in general more than one crystal.

In the next chapter, the effects of radiation damage with high intensity FEL radiation are described in details. In particular, it is shown that high X-ray fluences can significantly alter the scattering factors of the heavy atoms, similarly to a specific radiation damage effect. The conventional RIP scheme is adopted in chapter 5 to retrieve the coordinates of the photo-ionized species and to correctly phase the model structure.

Chapter 2

FEL radiation

Synchrotron radiation is generated when relativistic electrons (or charge particles in general) are accelerated in a magnetic field. There are three main types of magnetic structures commonly used to produce synchrotron radiation: bending magnets, wigglers and undulators. The former use a single magnet to curve the trajectory of an electron bunch, creating a fan of radiation around the bend. Wigglers and undulators use a periodic array of magnetic structures, so that the electrons experience a harmonic oscillation, which results in a narrow radiation cone emitted along the axial direction of the device.

This section provides a qualitative discussion about the origin of free electron laser radiation and the main properties of the FEL radiation are illustrated.

2.1 Bending magnet and undulator radiation

An electron experiencing radial acceleration as it travels around a circle emits radiation through a broad angular pattern. When the electron velocity is highly relativistic, however, the angular pattern is much compressed when seen in the laboratory frame of reference. This can be shown from the Lorentz transformation of the angles (see appendix 9.1):

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

where θ' is the angle observed in the frame of reference moving with the electron, and θ is in the laboratory. $\beta \equiv v/c$ is the relative velocity between frames and for relativistic electrons $\beta \simeq 1$, so for arbitrarily large emission angles θ' the radiation

is folded in the forward direction of half angle $\theta \simeq 1/2\gamma$.

For electrons traveling in a ring, one can estimate that the photon energies radiated depends on the time width of the observed radiation from a given point, through the Heisenberg's uncertainty principle ($\Delta E \Delta t \geq \hbar/2$):

$$\Delta E \geq \frac{2e\hbar B\gamma^2}{m} ,$$

so it is proportional to the magnetic field intensity B and to the electron velocity.

If the electron velocity is perturbed by a periodic magnetic structure, a small amplitude oscillation will start to occur, and the electron will radiate. If the angular excursions are small compared to the natural radiation width ($\theta \simeq 1/2\gamma$), the device is called undulator. The wavelength of the emitted radiation will depend on the magnetic period λ_u, but Lorentz contraction and relativistic Doppler shift will lead to a reduction in the radiated wavelength by a factor of $2\gamma^2$. Indeed, since the electron moves with relativistic velocity towards the periodic magnetic array, it will see a contracted period of $\lambda' = \lambda/\gamma$, and will emit dipole radiation with frequency $f' = c/\lambda'$. In the laboratory reference frame the radiation wavelength is further reduced by relativistic Doppler shift and becomes, for small angles θ relative to the undulator axis:

$$\lambda_n = \frac{\lambda_u}{2\gamma^2 n}\left(1 + \frac{K^2}{2} + \gamma^2\theta^2\right) ,$$

where n is the number of magnet periods and

$$K \equiv \frac{eB\lambda_u}{2\pi mc}$$

is called the magnetic deflection parameter. Thus, a periodic magnetic structure of a few centimeters can lead to observed X-ray wavelengths in angstrom.

Furthermore, the relative spectral bandwidth of an undulator radiation is much narrower than that of a bending magnet emitting at the same wavelength, and it's proportional to the number N of oscillation periods.

It can be shown (see for example pag.153 of [36]) that the average power radiated by electrons generating a current I into the central radiation cone of half angle $\theta = 1/\gamma\sqrt{n}$ is:

$$\bar{P}_{cen} = \frac{\pi e\gamma^2 I}{\epsilon_0 \lambda_u}\frac{K^2}{(1 + K^2/2)^2} , \tag{2.1}$$

with ϵ_0 the electric constant and λ_u the undulator period. The average power is therefore proportional to the number N_e of electrons in the bunch (since $N_e \propto I$).

2.2 Free-electron laser principles

Spontaneous undulator radiation is the workhorse of third-generation synchrotron facilities. The radiated power, as given by equation 2.1, assumes that the motion of electrons composing the bunch is uncorrelated, because of the random arrangement of them in the bunch. Thus the power is proportional to the electron current, since there is no correlation between the phases of the electrons, so only the intensity adds. Under favorable conditions, the electromagnetic wave generated inside an undulator copropagates with the electron beam in the forward direction and exchanges energy with the electrons. The copropagating radiation, indeed, overtakes the electrons in one undulator period λ_u by the resonant wavelength λ', and it can exchange energy with the electrons over many undulator periods. Depending on the relative phase of the electrons to the plane wave, some of them can gain energy from the radiation, while others will lose energy to the radiation. As the faster electrons catch up with the slower electrons, a periodic density modulation of the electron bunch begins to develop about the radiation wavelength in the undulator. This modulation is commonly referred to as "microbunching". For a sufficiently long undulator and a bright electron beam, the radiated intensity grows exponentially along the undulator distance as shown in figure 2.1. This growth will eventually stop as the electron beam microbunching reaches a maximum saturation level, when the longitudinal space-charge field between electrons matches in strength the bunching process. This process, called self-amplified stimulated emission (SASE), is the working principle of SASE free electron lasers (FELs).

2.2.1 SASE FEL properties

The amplification process of the SASE FEL due to the microbunching has a strong effect on the coherence properties of the produced radiation. Since only the wavelengths close to the resonance are exponentially amplified, the SASE FEL can reach almost full transversal coherence [37, 38]. Because of the stochastic generation of electrons in the electron gun, the temporal property of a SASE FEL is that of a chaotic polarized light. A simplified model of chaotic light can be represented, in the time or frequency domain, as a superposition of Gaussian pulses; the resultant wave is a relatively regular oscillation (see figure 2.2) interrupted only a few times. In the time domain, the number m_c of regular regions, given by the ratio of the bunch length to the average length of the regular region (the coherence length), is commonly referred to as the number of coherence modes. The evolution of each of these modes is nearly independent from the others, and their intensity fluctuation

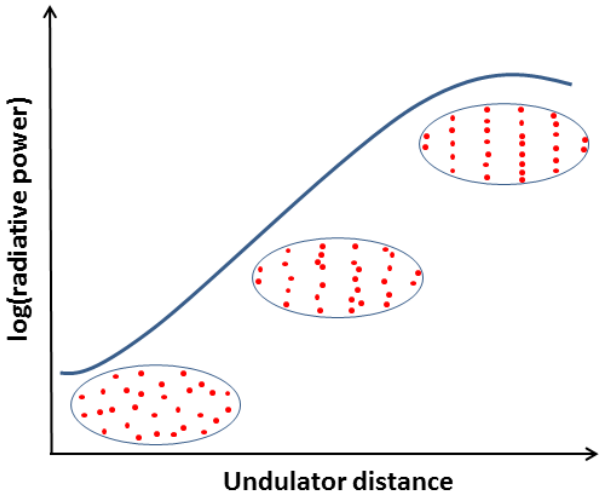

Figure 2.1: Growth of the radiation power and the electron beam microbunching as a function of the undulator distance.

can be described using the central limit theorem, as normally distributed. If the whole pulse is integrated, then the single fluctuations will be smoothed out, and the variance will be reduced to $\sqrt{m_c}$.

2.2.2 The LCLS and the CXI endstation

The Linac Coherent Light Source (LCLS) is a 2 km long FEL located at the SLAC National Accelerator Laboratory, at Stanford, USA. The machine utilizes 1 km of the previously existing SLAC linear accelerator, while the SASE process is initiated in a 132 m long series of undulators. The LCLS provides X-ray pulses at 120 Hz, between 270 eV and 10 keV. The typical pulse energy is about 2 mJ, and the pulse length can be adjusted between 40 and 300 fs, in FWHM (shorter pulses can be achieved by reducing the pulse energy). Due to the SASE process, the produced X-ray pulses are almost fully spatially coherent, while the expected bandwidth at saturation is, in the hard X-ray range, around 0.2%, with a similar photon wavelength jitter. The shot jitter of the X-ray arrival time is about 50 fs per minute.

All the FEL experiments described in this thesis have been performed at the Coherent X-ray Imaging (CXI) endstation at LCLS. This consists of a flexible instrumentation suite for hard X-ray diffraction experiments in a vacuum environment, well suited for serial crystallography techniques. The endstation is located

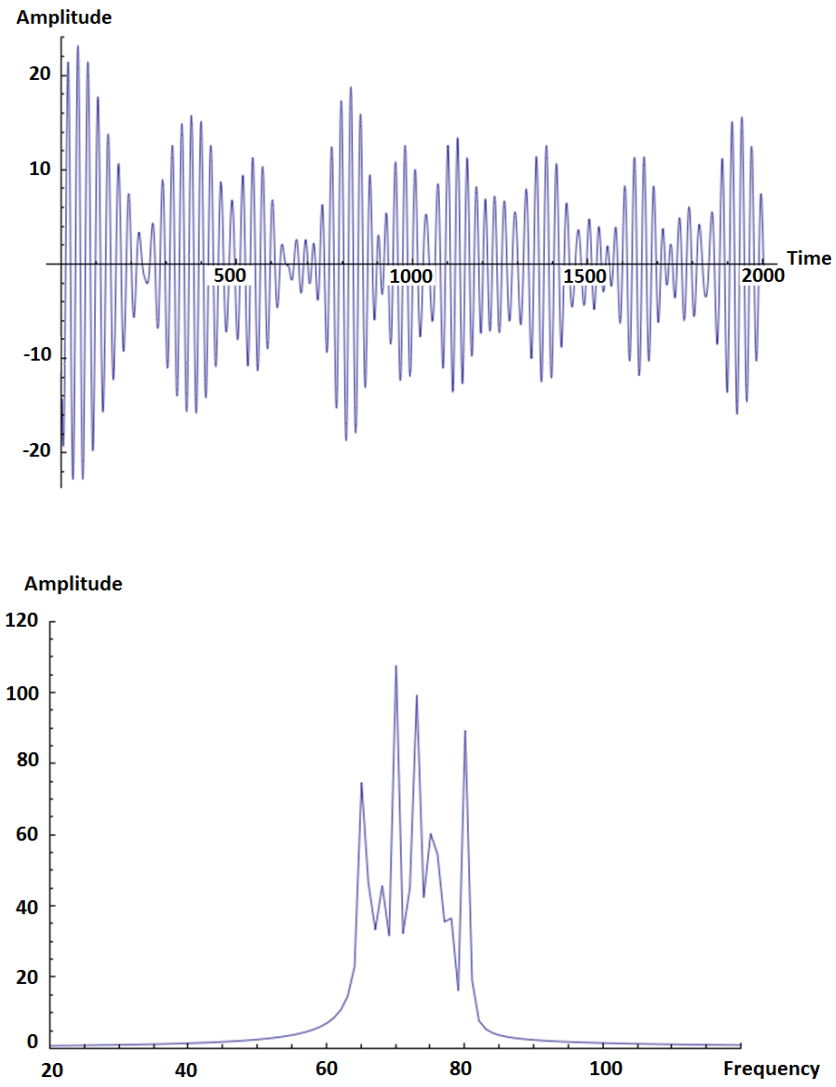

Figure 2.2: Top: random superposition of 100 Gaussian wave packets with random phase and 25% frequency spread. Bottom: the intensity spectrum corresponding to that wave packet.

383 m away from the exit of the undulators and it uses the FEL beam delivered by a grazing incidence mirror system. The geometry of the mirror acts as a high-energy filter, with a cutoff of about 25 keV (which can be produced by the high-order harmonics), while the rapid beam divergence sets the lowest X-ray energy available at the endstation to 2 keV. The FEL radiation can be focussed at the CXI instrument by two different Kirkpatrick-Baez (KB) mirror systems: a 1 μm focus system and a 0.1 μm system. These KB optics are connected to two different sample chambers, each equipped with a sample injector system and a CSPAD detector (described in section 3.2). The other important instruments are a pulse intensity monitor, which can provide a measurement of the photon intensity entering the endstation, on a shot-by-shot basis, and a set of motorized silicon attenuators, used to reduce the intensity of the beam before the interaction region. The endstation is also equipped with an external femtosecond optical laser that can be synchronized with the FEL pulse arrival time and delivered to the interaction region as a pump system for time-resolved experiments, with sub-*ps* time resolution.

2.2.3 Seeded FELs

The energy modulation of the electron beam can be induced by an external radiation, to optimize the amplification process. This method, called "seeding", allows to generate FEL radiation with a spectral bandwidth much narrower than the one of a SASE FEL (10^{-4} FWHM), with a much smaller photon energy fluctuation. The FELs adopting the seeding technique are commonly referred to as "seeded". Depending on the desired X-ray range of operation, as well as the seeding process itself, the seeding source may vary. In the soft X-ray range, an external pulsed laser can be adopted as master seed, and the FEL wavelength can be achieved through a high harmonic generation process involving multiple undulators [39, 40]. This method is used successfully at the FERMI@Elettra FEL in Trieste, Italy, allowing wavelengths down to 20 nm [41]. At hard X-ray wavelengths there are no existing external sources capable of driving the seeding process, so the modulating radiation must come directly from the FEL itself. A "self-seeding" scheme has been tested at LCLS, and involves a SASE lasing from a first undulator, which successively recombines with the same electronic bunch, by means of a delay line to retard and monochromatize the emitted radiation.

2.3 Diffraction before destruction

2.3.1 Ionization at high X-ray fluence

High intensity FEL-diffraction experiments lead to a rapid destruction of the sample, due to the Coulomb explosion. However, electronic damage dynamics may influence the scattering mechanism during the first femtoseconds, when the atomic motion is still negligible. As described in section 1.3, a single photoionization is followed by other ionization processes, such as relaxation effects or collisions events, which create an electron cascade in a very short time scale (tens of femtoseconds). At the dose rates delivered by an X-ray FEL, the ionization events are so numerous that they turn the sample into a plasma, which thermalises by expansion only after the pulse.

Since, for light elements, the X-ray photo-absorption probability is higher for inner-shell electrons than for valence electrons, it is possible to create atoms with empty inner shells if the dose rate exceeds the Auger decay rate. This effect is called "X-ray transparency", or "frustrated absorption", because the photo-absorption cross section of these "hollow" atoms is strongly reduced, and consequently electronic damage is suppressed [42]. The scattering cross section is however proportional to the number of remaining atoms, making X-ray transparency beneficial for diffraction experiments. The dose required to remove all valence electrons depends on the pulse duration, since it will depend on the collisional processes. Assuming a pulse longer than 10 fs and an average energy of 25 eV required to ionize a valence electrons, the dose rate needed to saturate collisional ionization is around 1.3 GGy, while for short pulses only multiple photoabsorption and relaxations events can result in a complete removal of electrons from an element, requiring a much higher dose of 20 GGy [23].

2.3.2 Ionic displacement and Bragg termination effects

The motion of the ions created by the FEL radiation can be modeled by molecular dynamics [43] or plasma physics codes [44, 45]. The former shows at high dose rates (above 1 MGy/fs) a random and isotropic displacement of the ions formed by the X-ray pulse. In a plasma, the ionic motion can be described by a diffusion equation, which is determined by the ion velocities and the collision frequencies. The evolution of the root-mean square (RMS) ion displacement depends on the diffusion constant $D(t)$, which increases with the dose, through the diffusion equation

$$\sigma(t) = \sqrt{2N_d D(t)t}\,,$$

with N_d the dimensionality of the system. Barty *et al.* have shown [46] that the RMS displacement increases approximately as $t^{3/2}$ for pulses longer than 10 fs (see figure 2.3). The same figure shows that the RMS displacement roughly increases with the square root of the intensity (or dose rate). Anisotropic displacements can however happen in presence of heavy atoms, due to the higher cross sections and the rapid ionization of them.

The total diffraction pattern of a sample exposed to FEL radiation is the pulse-integrated sum of waves scattered from the atoms during the plasma formation and expansion. The accumulated scattered signal can be written as:

$$I(q;T) \propto I_o T |F(q)|^2 g(q;T) \,, \tag{2.2}$$

with $F(q)$ the structure factor of the sample and $g(q;T)$ the dynamic disorder factor:

$$g(q;T) = \frac{1}{T} \int_0^T e^{-4\pi^2 q^2 \sigma^2(t)} dt \,, \tag{2.3}$$

similar to the Debey-Waller factor defined in section 1.5. From equations 2.2 and 2.3 it can be calculated that the Bragg diffraction terminates when the RMS displacement exceeds $1/(2\pi q) = d/(2\pi)$, with d the interplanar spacing of the reflection. Higher-resolution reflections thus "turn off" sooner, leading to lower counts on the detector, but eventually also the low angle diffraction terminates, even before the end of the X-ray pulse, thus producing an apparent pulse length shorter than the real one.

2.3.3 Atomic scattering factors at high X-ray intensity

At low resolution, the atomic scattering factor of an ionized atom can be assumed to be proportional to the number of bound electrons. Generally, an inner shell ionization event will shift the absorption edge of the atom to higher energy, since there is less screening effect on the remaining core electrons. The shape of the absorption edge depends on the charge state of the ion, i.e. on the atomic configuration on the remaining electrons, as can be seen in figure 2.4. The immediate consequence of these is that not only the scattering strength of an atom, but also its dispersion correction will be modified by intense radiation, and that the scattered intensity will depend on the population of the ions created during the pulse as well as on their dynamics. The heavy atoms in particular, due to the higher cross sections and the electronic cascade that can follow the first inner shell photoionization, are affected by electronic damage and their standard dispersion coefficients (defined in equation

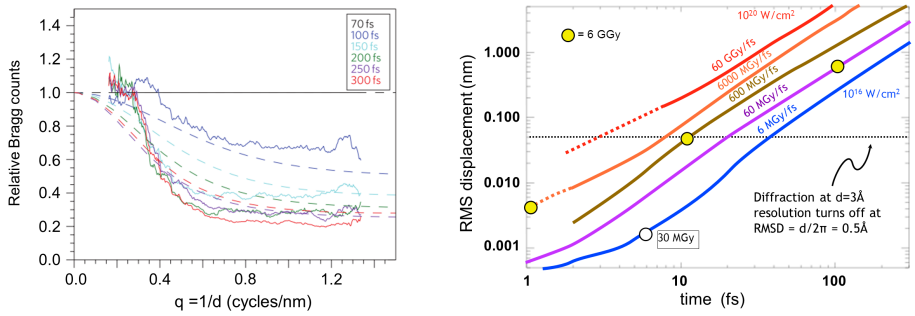

Figure 2.3: Left: Experimental evidences of the Bragg termination effect on the powder patterns of a photosynthetic protein collected at LCLS, reproduced with permissions from [46]. Right: the average RMS displacement for pulses at various dose rates as a function of the pulse duration, calculated by plasma dynamic code. Courtesy of Henry Chapman.

1.7) will require significant corrections depending on the dose rates.

To simulate the electronic damage dynamics at high intensity, a toolkit named XATOM has been developed by Son *et al.* [47], based on nonrelativistic quantum electrodynamics and perturbation theory. The atomic processes implemented in XATOM requires lots of calculations to provide numerical results, and the simulation time grows exponentially with the complexity of the system (proportional to the number of electronic levels to consider).

2.3.3.1 Evaluation of the scattered intensity with a single heavy atom species

Close to the absorption edge of a heavy element, the scattering cross section of the light atoms in a protein is generally much lower (for example $\sigma_{Fe}/\sigma_C \simeq 300$ at 8 keV), so one can assume that only the heavy atom species are effected by electronic damage. If only a single heavy atom species H is considered, the scattered intensity per solid angle, from a protein P, generated by a spatially uniform X-ray beam, can be evaluated at the reciprocal point $\mathbf{Q}$, as suggested by Son *et al.* [48], with:

$$\frac{dI(\mathbf{Q},\omega)}{d\Omega} = \mathcal{F}C(\Omega)\int_{-\infty}^{\infty} dt g(t) \sum_I P_I(t)|F_P^0(\mathbf{Q}) + \sum_{j=1}^{N_H} f_{I_j}(\mathbf{Q},\omega)e^{i\mathbf{Q}\cdot R_j}|^2 \,, \qquad (2.4)$$

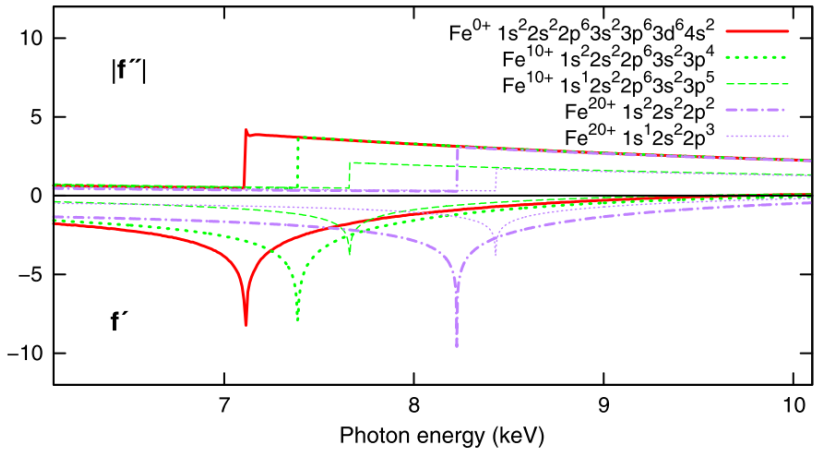

Figure 2.4: Dispersion coefficients for several charge states of Fe with different electronic configurations. Reprinted with permission from [48]. Copyright (2011) by the American Physical Society.

where I_j represents the electronic configuration of the jth out of N_H heavy atoms, ω is the wavelength of the incoming radiation, and $P_I(t)$ is the probability among all the possible electronic configurations $\{I\} = (I_1, I_2, ..., I_N)$ at the time t. The X-ray flux is represented by: $\mathcal{F}$ (the pulse fluence) , $g(t)$ (the normalized pulse envelope), and by the polarization factor $C(\Omega)$. The atomic form factor $f_{I_j}(\mathbf{Q}, \omega)$ includes the dispersion corrections at high intensity, while $F_P^0(\mathbf{Q})$ is the standard molecular form factor for the protein, as in equation 1.4.

Assuming that all the heavy atoms are ionized independently, the previous equation can be simplified to:

$$\begin{aligned}\frac{dI(\mathbf{Q},\omega)}{d\Omega} = \mathcal{F}C(\omega)[&|F_P^0(\mathbf{Q})|^2 + |F_H^0(\mathbf{Q})|^2\tilde{a}(\mathbf{Q},\omega)+ \\ &|F_P^0(\mathbf{Q})||F_H^0(\mathbf{Q})|b(\mathbf{Q},\omega)\cos(\varphi_P^0(\mathbf{Q}) - \varphi_H^0(\mathbf{Q}))+ \\ &|F_P^0(\mathbf{Q})||F_H^0(\mathbf{Q})|b(\mathbf{Q},\omega)\sin(\varphi_P^0(\mathbf{Q}) - \varphi_H^0(\mathbf{Q}))+ \\ &N_H|f_H^0(\mathbf{Q})|^2\{a(\mathbf{Q},\omega) \;-\; \tilde{a}(\mathbf{Q},\omega)\}] \quad ,\end{aligned}$$

with the introduction of new "MAD" coefficients:

$$
\begin{aligned}
a(\mathbf{Q},\omega) &= \frac{1}{\{f_H^0(\mathbf{Q})\}^2}\sum_{I_H}\bar{P}_{I_H}\left|f_{I_H}(\mathbf{Q},\omega)\right|^2 \\
b(\mathbf{Q},\omega) &= \frac{2}{f_H^0(\mathbf{Q})}\sum_{I_H}\bar{P}_{I_H}\left\{f_{I_H}^0(\mathbf{Q})+f'_{I_H}(\omega)\right\} \\
c(\mathbf{Q},\omega) &= \frac{2}{f_H^0(\mathbf{Q})}\sum_{I_H}\bar{P}_{I_H}f''(\omega) \\
\tilde{a}(\mathbf{Q},\omega) &= \frac{1}{\{f_H^0(\mathbf{Q})\}^2}\int_{-\infty}^{\infty}dtg(t)\left|\tilde{f}_H(\mathbf{Q},\omega,t)\right|^2 .
\end{aligned}
\tag{2.5}
$$

Here the electronic configurations of the independent heavy atom species are indicated with I_H, while the atomic form factors are still divided as in equation 1.7, with the exception of the "dynamical" form factor $\tilde{f}_H(\mathbf{Q},\omega,t) = \sum_{I_H} P_{I_H}(t) f_{I_H}(\mathbf{Q},\omega)$ which is a coherent average of the configuration-specific form factors $f_{I_H}(\mathbf{Q},\omega)$ at a given time. $\bar{P}_{I_H} = \int_{-\infty}^{\infty} dtg(t) P_{I_H}(t)$ is the pulse-weighted average population for the particular configuration I_H. The coefficients in 2.5 are atom specific and can be calculated theoretically using XATOM, if the atom is not too heavy. An example of the dependence of those coefficients on the photon energy, for a given X-ray fluence, can be found in [48] and it is here proposed in figure 2.5. It is worth noting that $\tilde{a}$ represents the effective scattering strength of the heavy atom.

2.3.3.2 The full MAD equation

As can be seen from figure 2.6, if the power of the X-ray pulse is high enough, also the effective scattering strength of the lighter atoms will be reduced, and the assumption made for retrieving the previous equations will not be valid any more. In the more generalized case where all the atoms in the samples scatter anomalously, equation 2.4 becomes:

$$
\frac{dI(\mathbf{Q},\omega)}{d\Omega} = \mathcal{F}C(\Omega)\int_{-\infty}^{\infty}dtg(t)\sum_I P_I(t)\left|\sum_X\sum_{j=1}^{N_X} f_{I_j^X}(\mathbf{Q},\omega)e^{i\mathbf{Q}\cdot R_j^X}\right|^2
$$

with X representing an atomic species in the molecule. In this scenario, the number of global configurations increases with a power law, making this equation too complicated to be handled. A possible solution is to separate the configuration index I into the single atom species, so that:

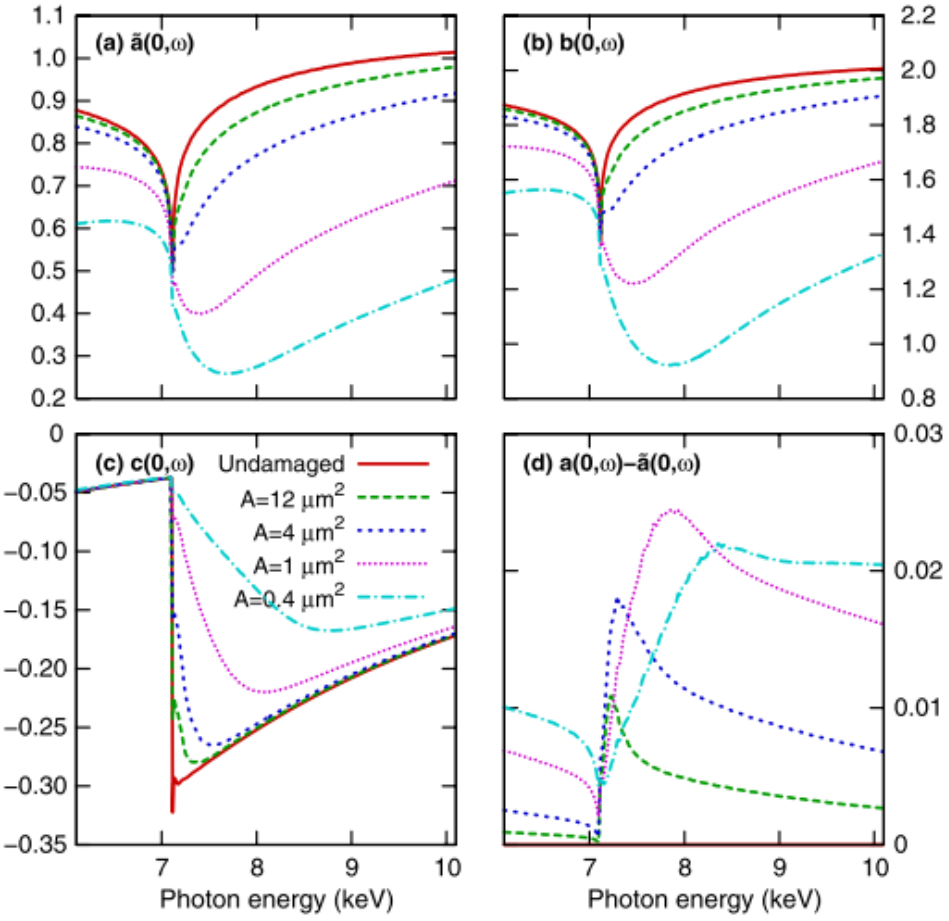

Figure 2.5: MAD coefficients computed with XATOM, assuming a Fe atom interacting with a 10 fs FWHM pulse having a fluence of $2 \cdot 10^{12}$ photons/A, where A is the X-ray spot area written in the graph. Reproduced with permission from [48]. Copyright (2011) by the American Physical Society.

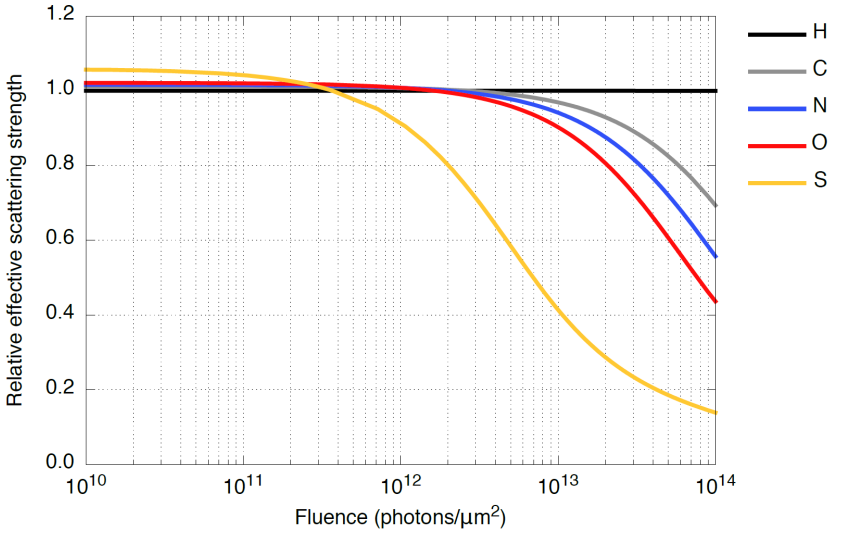

Figure 2.6: Relative effective scattering strength (coefficient $\tilde{a}$) as a function of the X-ray fluence, for different atomic species, at 6keV.

$$\frac{dI(\mathbf{Q},\omega)}{d\Omega} = \mathcal{F}C(\Omega)\sum_X[|F_X^0|^2\tilde{a}_X(\mathbf{Q},\omega) + N_X|f_X^0|\{a_X(\mathbf{Q},\omega) - \tilde{a}_X(\mathbf{Q},\omega)\}+ \sum_{Y>X}|F_X^0||F_Y^0|\{B_{XY}(\mathbf{Q},\omega)\cos\Delta\phi_{XY}^0 + C_{XY}(\mathbf{Q},\omega)\sin\Delta\phi_{XY}^0\}], \quad (2.6)$$

where

$$\Delta\phi_{XY}^0 = \phi_X^0(\mathbf{Q}) - \phi_Y^0(\mathbf{Q})$$
$$B_{XY}(\mathbf{Q},\omega) = \frac{2}{f_X^0 f_Y^0}\int_{-\infty}^{\infty} dtg(t)\left[\Re\{\tilde{f}_X(t)\}\Re\{\tilde{f}_Y(t)\} + \Im\{\tilde{f}_X(t)\}\Im\{\tilde{f}_Y(t)\}\right]$$
$$C_{XY}(\mathbf{Q},\omega) = \frac{2}{f_X^0 f_Y^0}\int_{-\infty}^{\infty} dtg(t)\left[\Re\{\tilde{f}_X(t)\}\Re\{\tilde{f}_Y(t)\} - \Im\{\tilde{f}_X(t)\}\Im\{\tilde{f}_Y(t)\}\right] \quad (2.7)$$

The symbols used are the same as in the subsection 2.3.3.1, with the omission of the $\mathbf{Q}$-dependence for the normal scattering and form factors. Compared to equation 2.4, the $b(\mathbf{Q},\omega)$ and $c(\mathbf{Q},\omega)$ coefficients are replaced with the biatomic $B_{XY}(\mathbf{Q},\omega)$ and $C_{XY}(\mathbf{Q},\omega)$. In this formulation, the assumptions of independent atom ionization and the synchronized ionization of the same atomic species (expressed by the $\tilde{f}$ term) still hold.

2.4 Determination of the anomalous coefficients at high X-ray intensity

2.4.1 Transmission experiment

The imaginary part of the scattering factor can be directly measured with a transmission experiment. At high X-ray intensity, the expression of the transmission coefficient (T) can be generalized as in [49], imposing the same assumptions used in the subsection 2.3.3. In particular, the ratio between the number of transmitted photons (N_{ph}) through a sample of thickness x and the number of incident photons, $N_{ph}(0)$, can provide a direct measurement of the anomalous coefficient $\tilde{c}$, via:

$$\frac{N_{ph}(x)}{N_{ph}(0)} \approx 1 + \frac{4\pi\alpha}{\omega} n_H \tilde{c}(\mathcal{F}, \omega) x \ ,$$

where α is the fine-structure constant, ω is the X-ray wavelength, and n_H is the number density of the heavy atom species. $\tilde{c}$ is defined as:

$$\tilde{c}(\mathcal{F}, \omega) = \sum_{I_H} \overline{P}_{I_H}(\mathcal{F}, \omega) f''_{I_H}(\omega) \ ,$$

where I_H indicates the possible electronic configuration of the heavy atom and $\overline{P}_{I_H}$ its time-averaged population. $\tilde{c}$ is related to the fluence-dependent anomalous coefficient c defined in equation 2.5, through:

$$c(\mathcal{F}, \omega, \mathbf{Q}) = \frac{2}{f^0_H(\mathbf{Q})} \tilde{c}(\mathcal{F}, \omega) \ .$$

Consider the specific case of a Fe solid target with a thickness of 200 nm, exposed to pulse of a $5 \cdot 10^{12}$ photons/μm^2, the expected variation of the transmission is around 6%. Experimentally, the transmission in a single shot could be measured with up to 2% accuracy (after accurate calibration of a pair of beam intensity monitors), and one could achieve much better than 0.1% error by averaging thousands of shots and by binning shots by incident pulse energy.

2.4.2 Fluorescence measurements

At low-intensity X-ray regime, fluorescence measurements can be used to determine the anomalous scattering coefficient f", through the optical theorem [50]:

$$f''(\omega) = -\frac{\omega}{4\pi\alpha} \sigma_P \ . \tag{2.8}$$

In this case, the fluorescence signal can be used to retrieve the photoabsorption cross section σ_P, proportional to it. At high-intensity, instead, the fluorescence signal is no longer linearly proportional to the photoabsorption cross section, due to the saturation of the one-photon absorption. In the case of neutral Fe, the fluence required to saturate the one-photon absorption is about $3 \cdot 10^{11}$ photons/μm^2 at 7.6 keV (slightly above the Fe K-edge), much below the expected fluence available at the LCLS facility. Nonetheless, assuming that the fluorescence yield is linearly proportional to the photoabsorption cross section, it is possible to convert the fluorescence yield into the fluence-dependent anomalous coefficient $\tilde{c}$ using 2.8, as:

$$\tilde{c}(\mathcal{F}, \omega) = \gamma\omega \frac{N_{fluo}}{N_{ph}} \ ,$$

where N_{ph} is the number of incident photons, N_{fluo} the measured fluorescence and γ is a scaling factor. Additional information on the oxidation states of the heavy atoms could be provided from a high resolution fluorescence spectra.

2.4.3 Scattering measurements

Once c is determined from both the transmission and fluorescence measurements, the other high intensity coefficients can be extracted from the scattered intensity using equation 2.5. In the case of a simple crystalline compound containing only the heavy atom species, the equation can be further simplified, as:

$$\frac{dI(\mathbf{Q}, \omega)}{d\Omega} = \mathcal{F}C(\omega)[|F_H^0(\mathbf{Q})|^2 \tilde{a}(\mathbf{Q}, \omega) + N_H |f_H^0|^2 (a(\mathbf{Q}, \omega) - \tilde{a}(\mathbf{Q}, \omega))] \ .$$

Similarly, by using a known protein system, one can use the phase information (accessible for example through a simple molecular replacement) to determine the missing coefficients. If the scattering measurements are performed in the proximity of an absorption edge, the anomalous terms from the Friedel pairs will also provide an additional estimation of the c coefficient.

Chapter 3

Serial femtosecond crystallography

Serial crystallography is a novel method for structure determination that was first proposed to overcome the main bottleneck afflicting conventional X-ray crystallography: the radiation damage effects. When used in combination with a bright and ultrashort FEL radiation, this technique can mitigate the problem of radiation damage by utilizing pulses that are briefer than the timescale of most damage processes [43] at the expense of the sample, which reduces to a plasma when exposed to the intense radiation. The serial femtosecond crystallography (SFX) technique addresses this problem by continuously exchanging the crystal exposed by using a liquid jet carrying crystals at high concentration and by sequentially collecting still diffraction images from many thousands of crystals hit by the X-ray beam. SFX was first demonstrated in 2009 by Chapman *et al.* [51] at the LCLS on a large membrane protein. The available experimental conditions at that time limited the resolution of the retrieved structure to 8.5 Å, but successive experiments have demonstrated the possibility of solving a known protein structure to high resolution [52], then with the determination of unknown biological information of an enzyme [53]. Recently, the possibility of retrieving phase informations with anomalous techniques has been reported by Barends *et at.* [54].

The typical SFX experiment can be schematized as in figure 3.1. The sample is introduced in the experimental chamber in form of a liquid jet. Pulsed X-rays are focused onto the liquid column by means of focusing elements, such as berillium lenses or Kirkpatrick-Baez (KB) mirror systems. Diffraction patterns are recorded at the rate of the FEL pulses with a fast detector. In the following paragraphs, the

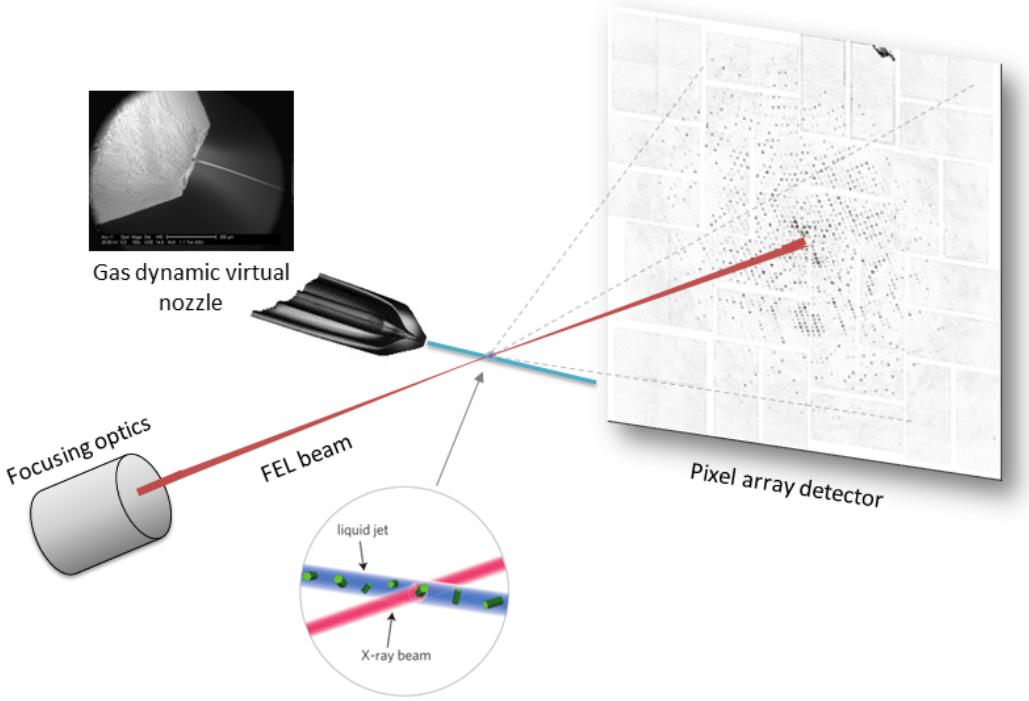

Figure 3.1: Scheme of a typical SFX setup.

main elements of the setup used for experiment at the CXI beamline at LCLS are described individually in detail.

3.1 Sample injection

Single crystal diffraction patterns can be obtained at a hit rate (defined as the number of patterns containing crystal diffraction, or "hits" divided by the number of frames collected) proportional to the dimension of the interaction volume and the sample concentration. The quality of the recorded pattern depends on the diffraction quality and the size of the crystal, but also on the level of background, which is mainly due to the diffraction signal of the liquid solution. The solvent scattering appears as a moderate background, and for an aqueous solution shows up as a broad ring around $3-4$ Å resolution (the so called "water ring"). To reduce the background, the size of the liquid should ideally match the crystalline dimensions, which are generally on the order of a few micrometers. The most used injector

system, capable of producing a stable μm-sized liquid jet, is the gas dynamic virtual nozzle (GVDN) [55]. This device generates a liquid stream focused by a coaxially flowing gas, typically He, through gas dynamic forces. A typical GVDN used for FEL experiments is shown in figure 3.2: an inner fused silica fiber, between 10 and $50\,\mu$m of inner diameter, carries the liquid solution. The fiber is inserted into an outer borosilicate or ceramic capillary, so that the gas can flow in the $\sim 150\,\mu$m wide annulus between the two. The head of the fiber is ground to a sharp tip and the outer capillary is shaped to match the fiber, creating a very tight aperture. This geometry allows to focus the liquid from a jet of tens of microns to a micrometer size jet, by means of a converging gas stream. The outer part of the borosilicate can be shaped to reduce the material around the aperture, which can attenuate the high-angle scattered radiation in case of an interaction region too close to the injector. Two HPLC pumps or gas regulators are used to push independently the liquid and the gas through the injector. The standard working parameters for a micrometer-sized steady jet are $10 - 30\,\mu$l/min and 500 psi of pushing pressure, in the case of an aqueous solution.

Other types of sample injection systems have been proposed and tested, depending on the viscosity of the medium carrying the crystals. For very high viscosity solutions, the most successful system is the lipidic cubic phase injector [56]. This consists of a hydraulic stage, together with a sample reservoir and a nozzle. The hydraulic stage is used to amplify the gas pressure from a HPLC system up to $10,000$ psi, needed to extrude gel-like solution contained in the reservoir. The nozzle part utilizes the co-flowing gas focusing technique (mainly to keep the solution straight), similar to the GVDN.

3.2 The CSPAD detector

Photon counting detectors (i.e. photodetector capable to detect single photons) are often used at synchrotron facilities due to their high count rates, large areas, and low noise levels. The integration time and the dynamic range of those detectors do not satisfy, however, the high frame rate and the instantaneous count rates produced by a FEL source. Indeed, the few-fs long X-ray pulses can generate pixel counts greater than 10^{17} photons/s [57] (as sketched in figure 3.3) with a continuous frame rate up to 120 Hz, for the case of LCLS.

The detector available at the CXI endstation is a $2D$ X-ray pixel array detector named CSPAD (Cornell-SLAC Pixel Array Detector). This detector comprises 64 ASICs (Application Specific Integrated Circuits) bonded on 32 silicon sensors,

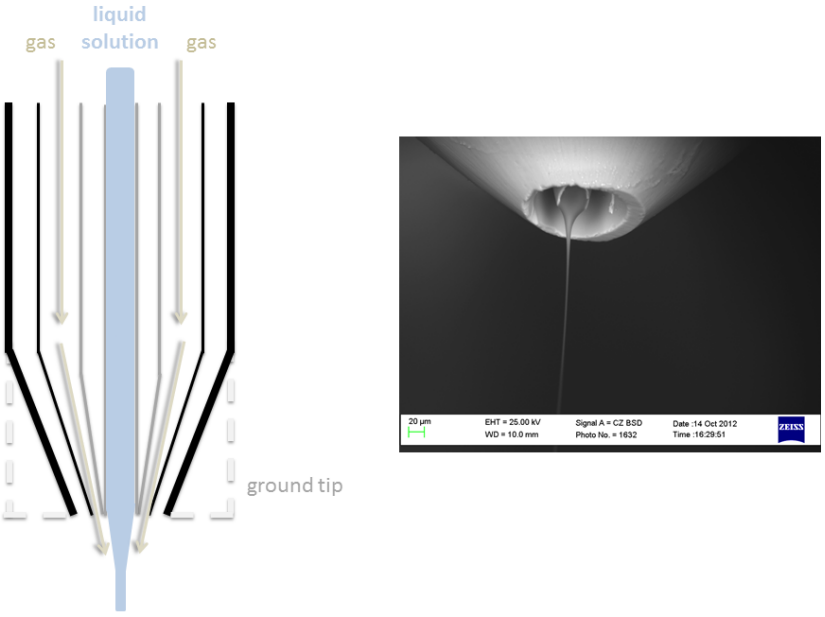

Figure 3.2: Left: a schematic of the GVDN. Right: SEM image of a running GVDN, courtesy of Rick Kirian.

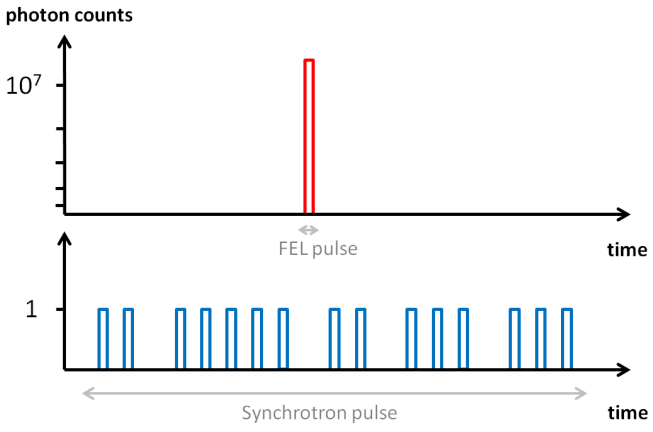

Figure 3.3: Comparison between the almost instantaneous pixel counts generated from a single FEL pulse and the single-photon counts typical of detectors used at synchrotron sources, during a single synchrotron pulse. The few-fs long X-ray pulses from a FEL can generate pixel counts greater than 10^{17} photons/s, which makes impossible to use single-photon counting detectors.

arranged in 4 independent quadrants with a central square aperture as beam hole. Each quadrant is mounted on a translator plate, so that the beam hole can be varied from 1 to 9.5 mm. A single ASIC is composed of 192×185 110 μm pixels, forming a 1516×1516 pixel array, arranged into an approximately 17 cm square detector with small gaps between the tiles (as can be seen from figure 3.1, ca 86% of the detector area is active). This detector has been specifically developed for imaging scattered X-rays from single molecules and small crystals at 120 Hz repetition rate, with a large dynamic range and a good signal-to-noise ratio. The detector features a per-pixel programmable gain control, that can be used to create an arbitrary two-level gain pattern across the detector, which allows simultaneously single photon sensitivity and thousands of counts per pixel.

Currently, two CSPADs are available at the CXI endstation, mounted in vacuum on independent detector stages. For high resolution data collection, the sample to detector distance can be varied continuously from 50 mm to 550 mm without breaking the vacuum. Smaller versions of the CSPAD, composed of only $140k$ pixels (i.e. 4 tiles), have also been used as beam monitoring or fluorescence detectors.

The large number of components of the CSPAD and the mobility of portions of it makes the knowledge of its metrology a key ingredient for any experiment recorded with this detector. As explained later in the dissertation, the metrology of each of the tiles has to be updated for every experiment, using visual references or iterative algorithms.

3.3 SFX data analysis methods

At the present time, the pre-processing of SFX data and the indexing stages differ mostly from standard methods, requiring new software specifically developed for handling huge amounts of diffraction snapshots. Indeed, since the detector readout frequency is often set to 120 Hz, the yield per hour is $432,000$ patterns, meaning that a large volume of data (intended both as memory size and as absolute number of images) is collected within a very short time. After the data processing, a final list of intensities and the associated errors is output, which can be then further processed with standard crystallographic programs. As will be shown in the next chapters, this current workflow has limitations, and the need of new tools and algorithms, specifically designed for FEL data, is emerging.

In the following sections, the programs used for SFX data analysis are described, following the natural steps from the raw data processing to the reconstruction of the final electron density map and the refined model of the structure, with a particular

emphasis on the main differences between SFX and the other canonical methods.

3.3.1 Pre-processing

One of the main differences between data acquired with SFX and data acquired from conventional crystallography methods is that not every recorded patterns contains crystal diffraction. This is because the positions of crystals inside the liquid suspension cannot be controlled, and the jet position itself can move away from the beam focus from time to time. Furthermore, even with a supercomputer, the amount of time spent to analyze the single frame can be high, while only $\sim 10\%$ of the detector readouts, on average, contain useful information. The first step of the analysis is then to extract and separate the images containing crystal diffraction from the blank shots. For the data described in the next chapters, the pre-processing is done with an open-source program called "Cheetah" [58]. To identify the diffraction patterns, the program performs a search through the single image and identifies clusters of pixels above a defined count threshold, labeling them as peaks. The number and size of the found peaks can be used to discriminate crystal diffraction from blank shots or other unwanted images. Cheetah can also perform other pre-processing tasks, like background subtraction, detector corrections, and creation of condensed data such as image stacks or profiles. The most common pre-processing parameters are explained in details in the appendix 9.6.

3.3.2 Indexing

The HDF5 images labelled by Cheetah as "hits" are later indexed using a specific component of the CrystFEL software suite, *indexamajig* [59]. Indexing is performed with conventional algorithms, such as the distributed and parallel subarray (DPS) FFT-based algorithm (implemented in the program called MOSFLM [7, 60, 61]), the DirAx algorithm [62], or XDS [34]; the indexing tool applies those programs using the list of peaks locations of the single image. If the unit cell parameters are known, the indexing is considered successful either if the lattice parameters found by the indexer match the ones of the unit cell, within a user-defined tolerance, or if they can be made to match by an affine transformation. In the case of unknown lattice parameters, the indexing tool can also give a first estimation of the unit cell parameters. In both cases, the found crystal orientation is then used to predict the location of diffracted spots on the image, and a minimum percentage of found peaks has to lay close to the predicted locations in order to consider an image indexed.

The success of the indexing procedure is quantified by a number, the "indexing yield", defined as the percentage of indexed patterns over the number of hits. The indexing yield is therefore dependent on the quality of the pre-processing, as well as on the amount of diffraction patterns containing hits from more than a single crystal (in the case of a highly concentrated crystal solution). A more realistic success rate quantifier is the indexing yield as a function of the number of peaks found in the pattern, as shown in figure 3.4 for the case of a highly concentrated solution of lysozyme crystals. The plot shows that the best indexing rate is found for patterns containing few tens of Bragg peaks (this number depends on the experimental conditions and on the type of protein crystal), while at lower and higher numbers of found peaks the indexing algorithm fails due to, respectively, false positives (i.e. diffraction peaks not coming from the protein crystal, represented in the figure as a negative exponential decay) or multiple hits, represented by a double Gaussian centered at twice and three times the mean value of the Gaussian curve used to fit the indexable patterns. At the same time it is possible to check *a posteriori* the quality of the peak-finding algorithm used and the parameters set for the indexing.

CrystFEL is being constantly upgraded and improved; new indexing algorithms and novel ideas appears at every version: the quality of the processed data, therefore, strongly depends on the features used. The data showed in this thesis have been processed with the version 0.5.2, using specific parameters optimized for the particular dataset, which are described in the relevant sections.

3.3.3 Merging of intensities

Indexamajig writes a "stream" of information about each processed frame, such as the location and integrated intensity of the found peaks, and the unit cell parameters resulting from the indexing. Also in the stream is recorded the reflection list of the predicted peaks location of the individual diffraction pattern. For a sufficiently large dataset, each reflection will be sampled multiple times and it will be measured in a range of intensity values. These individual intensity measurements are merged using a simple procedure consisting of a Monte Carlo integration over the *3D* reflection profiles [63], explained in details in the Appendix 9.7. An example of histogram of measurements from a single Bragg reflection is given in figure 3.5. Despite the large number of weak or negative intensity observations, the final Monte Carlo integrated intensity is relatively high (about 1100 counts, as marked in the figure with a red dashed line). The Monte Carlo integration could also compensate for factors such as the crystal size distribution, the quality of the crystals, and the stochastic nature

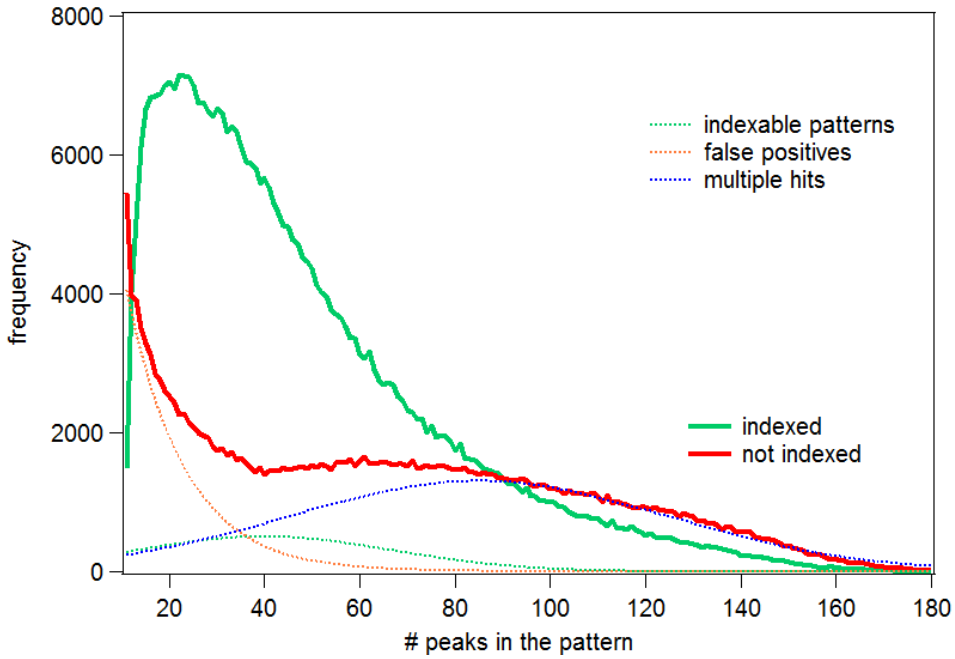

Figure 3.4: Histograms of the indexed and not-indexed patterns as a function of the number of Bragg peaks found in the image. The indexing rate (or yield) is the ratio between the red line and the sum of the red and the green lines. The dashed curves represent the contributions to the not-indexed curve coming from false positives, multiple hits and single hits that could not be indexed. The fit was performed using a single exponential decay to represent the false positives, and three Gaussian curves, with mean values of respectively x (for the indexable patterns), $2x$, and $3x$.

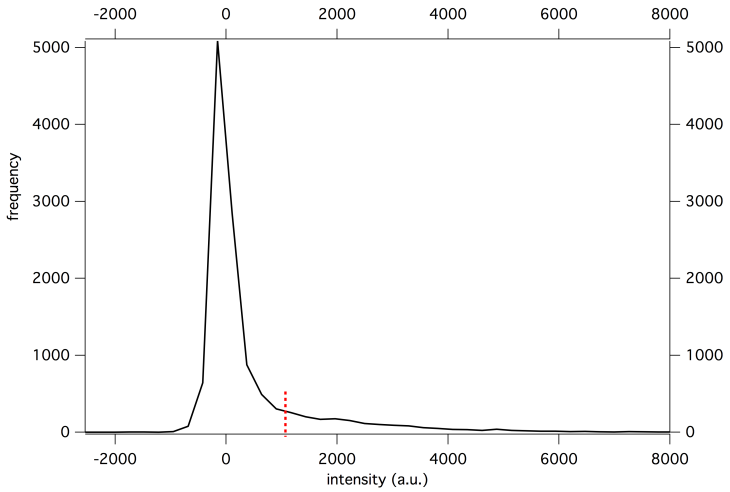

Figure 3.5: Example of histogram of measurements of a Bragg reflection, from the dataset described in chapter 4.

of the lasing process: it is hoped that these quantities can be averaged out, resulting in a constant scale factor equally affecting the intensities.

The error associated to the intensity of each reflection are estimated from the variance of the intensity distribution of each Bragg reflection.

The program *process_hkl* (CrystFEL) performs the merging of the individual intensities using this method, taking into account the symmetry of the structure. The quality of the final reflection list can be also improved performing simple scaling of the intensities, using a previously unscaled list from the same set of data in a two-pass process.

3.3.4 Evaluation of the data quality

Traditional data quality metrics, such as R_{merge}, defined as

$$R_{merge} = \frac{\sum_{\mathbf{h}} \sum_i |I_i(\mathbf{h}) - \overline{I(\mathbf{h})}|}{\sum_{\mathbf{h}} \sum_i I_i(\mathbf{h})} ,$$

do not give a meaningful measure of the data quality of a SFX set, due to the high multiplicity values (i) of each Bragg reflection $\mathbf{h} = (hkl)$. The SFX community adopts newly defined figures of merit, obtained by equally splitting the data into two sets (usually, even-numbered and odd-numbered pattern), which are merged independently. The agreement between the two resulting intensity lists is then examined,

for example, defining a metric R_{split} as:

$$R_{split} = \frac{1}{\sqrt{2}} \sum_{\mathbf{h}} \frac{|I_{even}(\mathbf{h}) - I_{odd}(\mathbf{h})|}{\frac{1}{2}(I_{even}(\mathbf{h}) + I_{odd}(\mathbf{h}))} ,$$

where I_{even} represents the intensity of a reflection from the odd-numbered patterns, and I_{odd} is the equivalent reflection from the even-numbered patterns. Since the comparison is done from intensities merged using half of the entire dataset, the convergence of the whole set is expected to be $\sqrt{2}$ better, which is why the factor appears in the equation.

Other metrics used in this thesis include:

$$R_{ano} = \sum_{\mathbf{h}} \frac{|I^{+}(\mathbf{h}) - I^{-}(\mathbf{h})|}{\frac{1}{2}(I^{+}(\mathbf{h}) + I^{-}(\mathbf{h}))} ,$$

with $I^{\pm}$ the Friedel pairs, the Pearson correlation coefficient (CC) and the R_{split}/R_{ano}. Statistics such as the mean $I/\sigma(I)$, the redundancy, and the completeness as a function of the resolution shells are also useful quality metrics. Examples of those metrics and statistics are given in the next chapters.

3.4 Time-resolved protein crystallography

The future of X-ray protein crystallography is most probably connected to the possibility of observing induced dynamics of biological molecules. The SFX technique and the possibility of synchronizing the FEL beam to an optical laser with a time resolution down to sub-picoseconds [64, 65] in a pump-probe setup are one of the few and most interesting tools that allow to extend crystallography to the time domain. In particular, A.Aquila *et al.* [66] showed that the optical-pump-SFX-probe technique is able to distinguish induced changed in a crystal structure due to the optical laser damage, while C. Kupitz *et al.* [67] have shown that the structure of a light-driven metastable state of the Photosystem II can be retrieved at moderate resolution.

Chapter 4

High-intensity SFX

The high X-ray doses produced by single XFEL pulses in very short time scales allow diffraction patterns to be recorded from much smaller crystals than have been examined at synchrotron radiation facilities, overcoming the exposure limitation set by radiation damage, without the need for cryogenic cooling of the sample. To date, the smallest protein crystals yielding near atomic resolution (2.0 Å) structure using an XFEL is the polyhedrin contained in granulosis virus particles. These particles are less than $0.02\,\mu\text{m}^3$ in volume, i.e. hundreds of times smaller than the smallest crystals collected at synchrotron radiation facilities [68, 69, 70].

4.1 The granulovirus

Cydia pomonella granulosis virus (CpGV) is part of the baculoviridae, a family of viruses that, together with cypoviruses, are natively embedded in extremely stable protein crystals called polyhedra. this stability is mainly due to the very low solvent content (23%) [68], which produces a sealed crystalline shell to protect the virus from environmental damage [71].

CpGV particles have a narrow size distribution, with an average size of about $210 \times 210 \times 400\,\text{nm}^3$, as shown in figure 4.1. Only about 30% of this volume consists of crystalline protein, for a total of only $8,000$ unit cells.

CpGV particles were produced by infecting larvae and purified as described in [72].

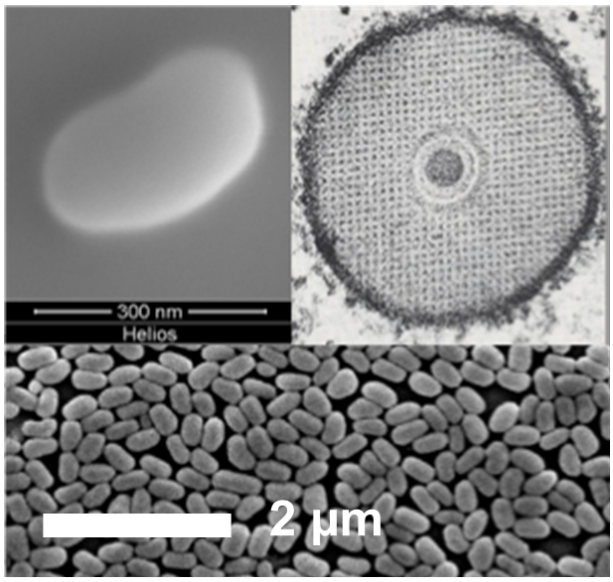

Figure 4.1: Electron microscope image of CpGV particles. Top: a section of CpGV recorded with a transmission electron microscope, showing the crystalline shell protecting the viral body. Courtesy of Peter Metcalf.

4.2 The LCLS experiment

X-ray diffraction patterns were collected from CpGV particles at the $1\,\mu$m focus sample chamber of the LCLS CXI endstation. The particles were suspended in water and flowed across the X-ray beam using a helium-focused GDVN of $50\,\mu$m inner diameter capillary producing a liquid jet of $3-4\,\mu$m of diameter. X-ray pulses with a photon energy of 7.9 keV and duration of 50 fs were focused to about $1\,\mu\text{m}^2$. Assuming an average of 2.7 mJ X-ray pulse energy and a beamline transmission efficiency of 60%, the maximum dose per pulse can be calculated as 1.3 GGy. About 3.5 hours of data collection yielded about $487,000$ diffraction patterns containing more than 20 Bragg peaks, as identified with *Cheetah*. A series of indexing trials and geometry refinement runs (see section 6.3.1 for a description of geometry refinement procedures) resulted in a total of $82,603$ indexable crystals diffraction patterns.

Figure 4.2 shows the distribution of the average Bragg peak intensities as a function of the incoming X-ray energy, as recorded from a beam intensity monitor located upstream of the beamline. Since the distribution of crystal volumes is quite narrow, the spread visible in the figure can be mainly attributed (in the absence of major data processing errors) to the profile of the focused beam as sampled by the randomly-positioned crystals.

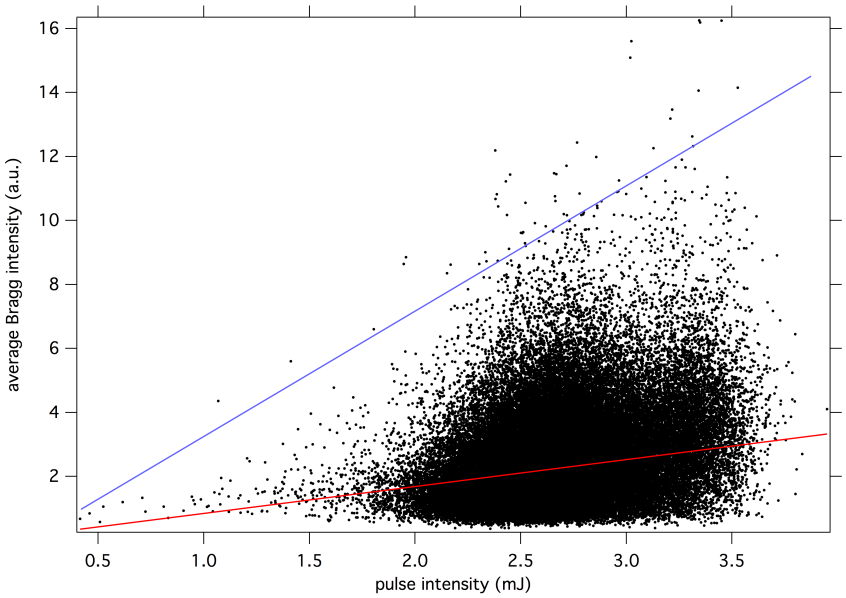

Figure 4.2: Scatter plot of the average Bragg intensity as a function of the pulse energy. Each point represents a single indexed diffraction pattern. The solid lines are the results of linear fittings.

4.2.1 Data analysis

Polyhedrin crystalises in a cubic point group, with a lateral dimension of about 103 Å. Its space group ($I23$) presents an indexing ambiguity, since there is a choice of two orientations that give rise to a particular set of locations of Bragg peaks on the diffraction pattern. This ambiguity is not resolved by the adopted indexing algorithms, but it might be overcome *a posteriori*, by examining the relative intensities of the Bragg peaks. In order to estimate the correct orientation for each indexed pattern, we adopted the algorithm devised by Brehm & Diederichs [73] and implemented in *ambigator* (*CrystFEL* suite). This script computes the average Pearson correlation between a single diffraction pattern and the rest of the dataset, and suggests the orientation which produces the best correlation. In the case of a SFX dataset, however, the partiality of the recorded Bragg reflections may decrease the calculated correlation between two patterns, so that a small fraction of the indexed patterns will still be twinned.

Fully-integrated counts were obtained using *process_hkl* and resulted in a dataset with a highest resolution of 2.0 Å (the statistics of the data quality are reported in table 4.1). Data processing was followed by structure solution by molecular replacement, performed using *Phaser MR* [74], using the structure of wildtype bac-

Wavelength (Å)	1.56 (7.9 keV)
Pulse fluence (photons/μm^2)	$1 \cdot 10^{12}$
Corresponding dose	1.3 GGy
Space group	$I23$
Cell dimensions (Å)	$a = b = c = 103.3$
Number of "hits"	$487,085$
Number of indexed patterns	$82,603$ (17%)
Highest resolution (Å)	2.06
Completeness	100% (99.95%*)
I/σ(I)	9.59 (0.92*)
R_{split} (%)	7.89 (116.2*)
CC (%)	0.99 (0.60*)
Redundancy	6008 (1258*)

Table 4.1: Overall SFX statistics. The values with * refer to the highest resolution shell.

ulovirus polyhedra (55% sequence identity) as a starting model. The initial solution was subjected to automated model building and refinement using the phenix package [75] alternated with manual model building with COOT [76]. The final model was validated with PDB_REDO [77] and presented a final *Rwork* of 14.5% ($Rfree = 18.9(5)\%$) . Figure 4.3 shows the resulting electron density map around a portion of the refined structure.

4.2.2 Discussion

In figure 4.2, the average peak intensity of the majority of the patterns does not follow the expected trend as a function of the pulse intensity, i.e. they are not clusered around a line. The lack of correlation between the average Bragg intensity

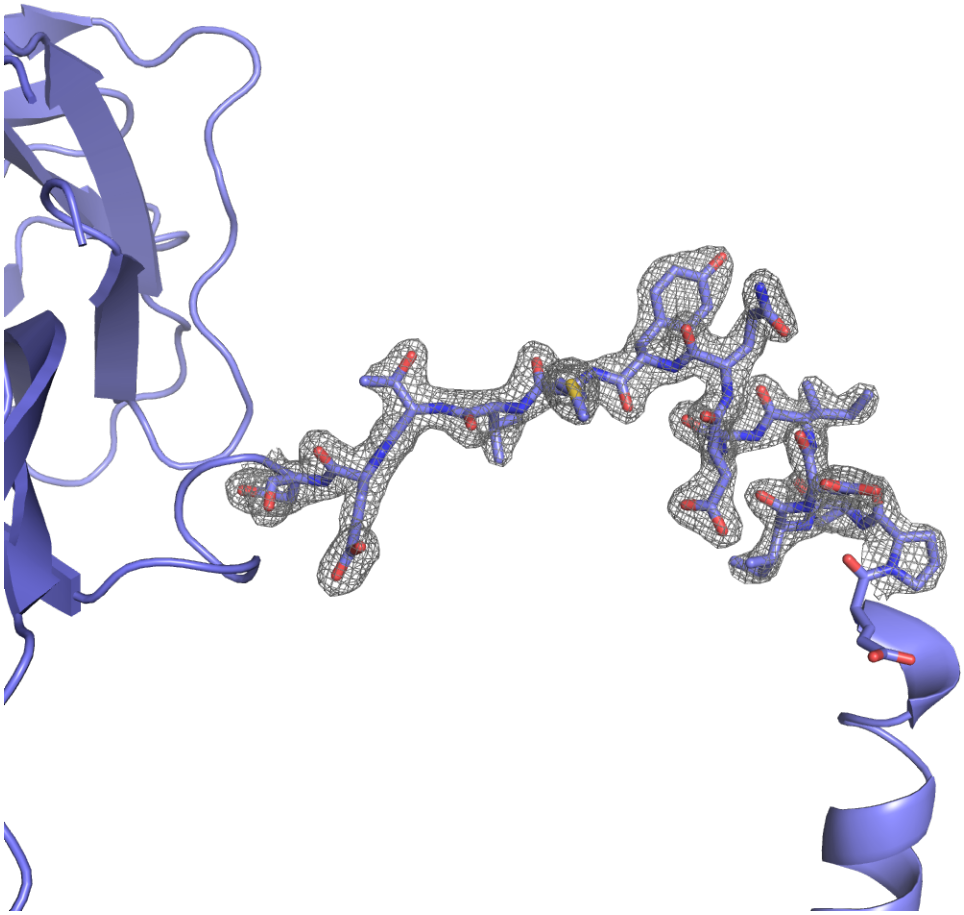

Figure 4.3: Section of the retrieved electron density map superimposed to the refined model. The map is countored at 1.5 sigma.

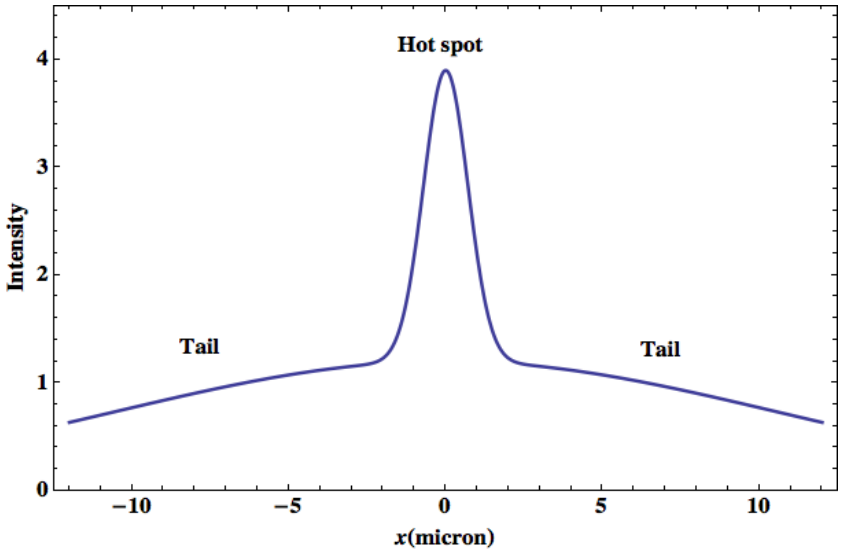

Figure 4.4: Suggested beam intensity profile.

and the pulse energy indicates the presence of a large, low intensity region which extends much further than a few microns. The collected dataset, then, does not fully utilize the highest available pulse fluence, but a portion of it, which can be roughly estimated from the average Bragg intensity as 1/4 of the peak fluence. The solid lines in the figure are the results of two linear fits. In particular, the red line represents the average trend of the Bragg intensity as a function of the impinging pulse energy, where the single data points used were weighted by the number of Bragg spots found in the image. The blue line is instead derived from a small percentage of selected, very bright diffraction patterns, showing the expected trend in the case of a perfect beam spot. A possible X-ray beam intensity profile is shown, as a visual example, in figure 4.4.

The great majority of the diffraction patterns collected are limited in resolution by the largest scattering angle recordable on the detector, indicating that diffraction from even smaller crystals can be recorded. Moreover, the SFX method allows to increase the signal-to-noise ratio (SNR, defined as the ratio between the average intensity of the reflection and the associated sigma) of the single Bragg peaks by averaging hundreds of observations, so that even very small signal can be distinguished from a relative large background noise. This is nicely proved in the top graph of figure 4.5, where the signal to noise ratio of selected high resolution Bragg peaks is plotted as a function of the times the reflection has been observed (i.e. the multiplicity of the reflection). The graph also shows that the improvement of the

data quality is faster at low resolution, because of the stronger scattering strength at low scattering angles. Given a number of indexed patterns, reflections at different resolution are predicted with different frequencies, due to the combination of crystal symmetry, dead areas on the detector, and the parameters used to model the X-ray beam, such as the beam divergence and the X-ray bandwidth.

The distribution of the final integrated intensities is presented in the bottom graph of the figure, together with the estimated background (defined as the sigma of the intensity distribution). The black line is the average signal-to-noise ratio calculated for the reflections in corresponding resolution shell. A longer data collection can reduce the estimated sigma level, increasing the SNR. For sufficiently large datasets, the SFX method should allow to distinguish signal with less than a single photon, since the final reflection intensity is the arithmetic average of the single observations. Thus, the dynamic range of the measurements can be increased by several orders of magnitude: ideally, the lowest recordable intensity corresponds to $1/n$ photons, with n the number of collected patterns, in the case where one single photon is recorded on one image, and the remaining $n - 1$ patterns do not contain signal. The highest intensity, however, is limited by the number of counts the detector can tolerate. High fluence measurements on larger crystal, then, may cause severe damage to the detector, since bright scattered reflections can easily saturate it. This can be avoided by utilizing solid attenuators, placed after the interaction region to protect the detector from the intense Bragg reflections, as described in chapters 6 and 7.

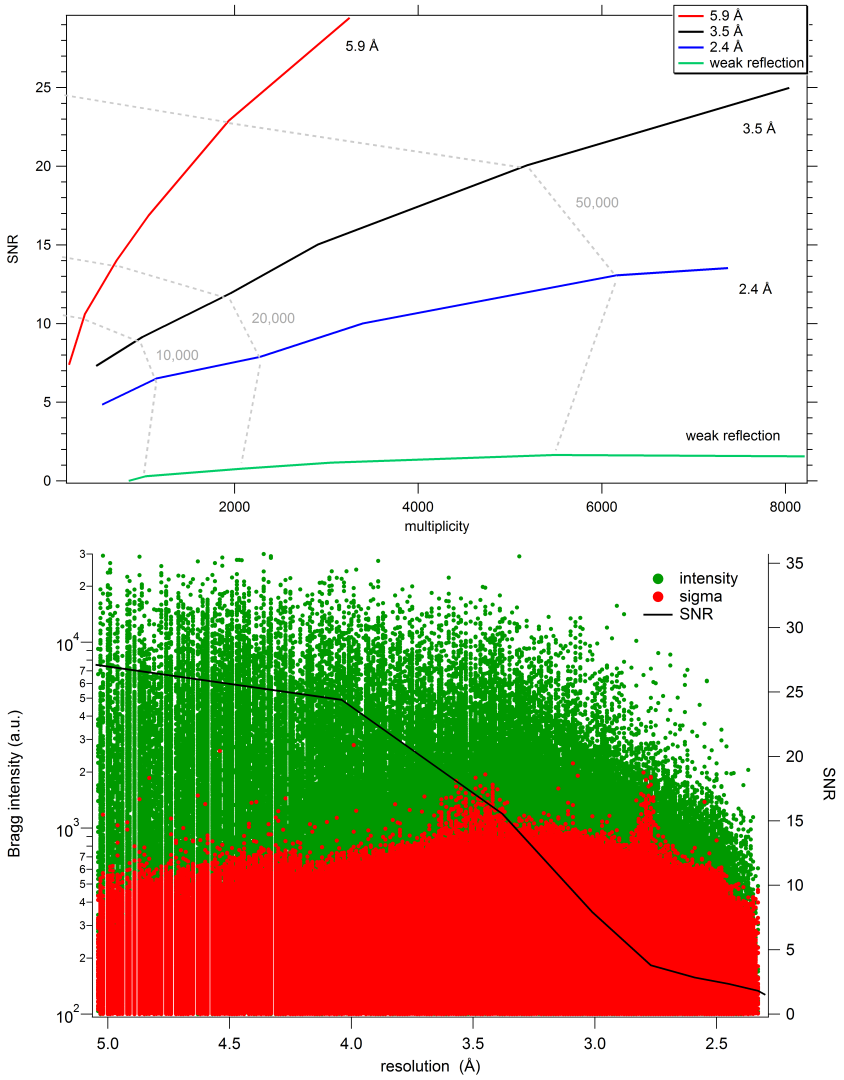

Figure 4.5: Top: the trend of the signal to noise ratio for selected Bragg reflections at different resolution, as a function of the multiplicity. The grey dotted lines corresponds to the number of indexed patterns required. Bottom: scatter plot of the integrated reflections and the associated sigma, as a function of the resolution. The average signal to noise ratio is over-plotted as a black line.

Chapter 5

HI-RIP simulations

So far, the reformulation of anomalous scattering at high X-ray intensity has been described theoretically, and the generalized equation 2.6, for the case of a monochromatic, flat-top X-ray beam has been introduced. This theory predicts that the X-ray fluences already achievable with existing FELs are enough to induce ionization that could significantly alter the scattering factors of the heavy atoms, hindering the direct application of anomalous phasing methods. This is analogous to the situation in synchrotron macromolecular crystallography, where radiation damage, if unaccounted for, can dramatically hinder anomalous phasing [78]. In the SFX case, ionization introduces the exciting possibility of determining phases *de novo* by varying the fluence and hence the scattering factors of the heavy atoms. This new approach could represent a powerful method of experimental phasing which would not require the modification of the native protein crystals if the sulfur atoms already present in the structure could be used. Since ionization occurs at all sulfur atoms and not just those that are found in disulfide bonds, this kind of electronic modification could be of broader applicability than synchrotron RIP, which requires disulfide bonds and/or metals. To avoid confusion with damage phasing by synchrotron radiation damage at cryo temperatures [31, 79], and to highlight the central role of the X-ray intensity, we call this new technique "High-Intensity Radiation Induced Phasing" (HI-RIP).

The CXI endstation at LCLS is, to date, the most intense hard X-ray FEL beamline currently available, which can provide up to 0.5 mJ per pulse in the photon energy range between $6 - 10\,\mathrm{keV}$, inside a focal region of $0.2\,\mu\mathrm{m}$ of diameter. At 6 keV the cross section of sulfur is the highest and the maximum fluence achievable reaches $10^{13}\,\mathrm{photons}/\mu\mathrm{m}^2$. As can be seen from figure 2.6, at this photon density the scattering strength of sulfurs is reduced to 40% of the normal value, while the lighter atoms are only minimally affected by the ionization processes, suggesting

that a HI-RIP experiment is feasible.

In this chapter, it is demonstrated that HI-RIP can in theory be used to determine substructures of radiation damage as well as determine high quality phases, by simulating two serial femtosecond crystallography experiments, at high and low X-ray fluence, under experimental conditions that mimic those available at the LCLS. Furthermore, the effect of different experimental parameters such as variable fluences and the number of patterns on the quality of the substructure solution and phasing is explored. The last section is about interesting results found assuming that the crystal aligned themselves with respect to the liquid flow direction.

5.1 Simulation of an SFX experiment

I used the *Trypanosoma brucei Cathepsin B* (CatB) structure recently solved at LCLS (RCSB code 4HWY) [53] to test the method. This protein consists of 340 residues, and includes 19 S atoms (in 5 Methionines and 14 Cysteines). The complex-valued structure factors of equation 2.6 (and hence their phases) were computed separately for each atomic species present in the protein structure using the *sfall* program [80] (Collaborative Computational Project No. 4 [81]); Friedel mates were averaged with an ad hoc script. The coefficients $B_{XY}(\mathbf{Q},\omega)$, $C_{XY}(\mathbf{Q},\omega)$, $a(\mathbf{Q},\omega)$, and $\tilde{a}(\mathbf{Q},\omega)$ were calculated with the XATOM toolkit using equations 2.5 and 2.7, for an X-ray energy of 6 keV ($\omega = 2.066$ Å), considering a 10 fs-long top-hat X-ray pulse. The final scattered intensity for each Bragg reflection was calculated using equation 2.6. The scattering contribution of the bulk solvent region (i.e. the unit cell volume occupied by amorphous solvent) was generated starting from the PDB using the *ano_sfall.com* script [82, 83] and it was summed to the scattering factors of the protein. The SFX experiment was simulated with *partial_sim*, which is part of the CrystFEL suite. The program takes a list of fully integrated reflection intensities and generates partial reflection intensities of randomly oriented crystals, adding noise to simulate other measurement errors. The detector geometry was chosen to reproduce the 64 tiles of the CSPAD installed at the CXI endstation. The sample-to-detector distance was set to 11 cm, giving a resolution limit of about 2.7 Å at the corners of the detector for 6 keV energy. The simulated noise and the beam parameters (in particular the beam profile radius and the beam divergence) were selected in order to produce *Rsplit* and $< I/\sigma >$ values as close to a real experiment as possible (see figure 5.1). I generated two sets of $300,000$ patterns, one with 10^{13} photons/μm^2 (as "high fluence") and another using 10^{11} photons/μm^2 (as "low fluence"). The simulated noisy partial intensities were then merged as described in

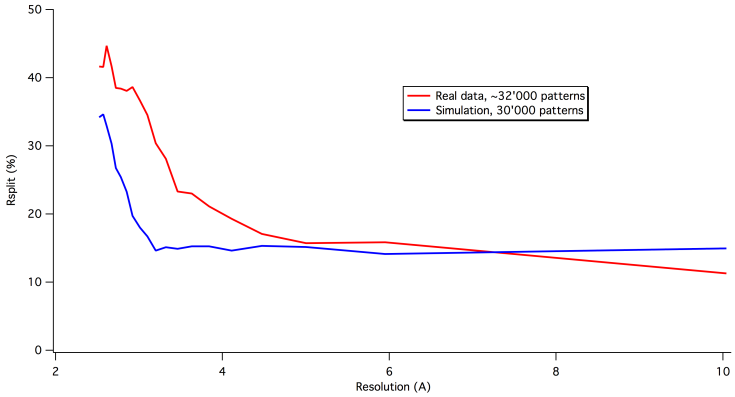

Figure 5.1: Comparison of the R_{split} metric as a function of the resolution for simulated (blue) and real data (red). The real data belongs to the experiment described in the next chapter.

section 3.3.3. To test the convergence of the Monte Carlo procedure, I processed smaller subsets of the complete data set ranging from $1,000$ to $300,000$ patterns. The RIP workflow [35] was then utilized to determine heavy atom substructures and determine phases.

5.2 Phasing

The first step in experimental phasing by HI-RIP is the determination of the "damage substructure". RIP substructures are generally qualitatively different from those of anomalous dispersion and isomorphous replacement, and require specialized methods [35, 84]. Because RIP substructures consist of many weak sites, iterative improvement of substructures (see appendix 9.4) and the down-weighting of the damaged (high fluence in HI-RIP) dataset are necessary. The down-weighting of the damaged dataset is done in a simple manner, in which an initial scale is determined using conventional programs, and then down weighted by the scale factor k (as defined in section1.3.2.1). In some cases, both techniques are required for the successful determination of the damage substructure, whereas in others only one of these methods is sufficient. Furthermore, it has been observed that a peak in the correlation coefficient (CC) of SHELXD [32] substructures as a function of scale factor is an excellent predictor of RIP signal, which complements the conventional indicators of signal strength such as $R_{\text{isomorphous}}$ [35]. In comparing the high and low fluence datasets, such a relationship has been observed, as displayed

in figure 5.2. Because the damage substructure and a reference phases set from a refined model are known, it is also possible to show that the phased difference $F_{\text{lowfluence}} - F_{\text{highfluence}}$ map calculated with model phases by the program ANODE [85] showed peaks, and that these peak heights (called "RIP peak heights") were quite high (up to 25.6 σ) over sulfur positions in Methionines and Cysteines, as expected from figure 2.6. The top ten strongest peaks were over the sulfur positions of residues $C107$, $C219$, $M131$, $C154$, $C122$, $C136$, $M138$, $C119$, $C158$, and $C192$ with peak heights of 25.6, 25.3, 24.9, 24.1, 28.9, 23.7, 23.2, 22.9, 22.4, and 22.2 (compared with σ). Model phases are not available when trying to determine a new structure; therefore, a more useful measure of whether there is adequate signal are the substructures quality metrics that are produced by SHELXD. The most frequently used metric in substructure solution is the ratio of $CC(all)$ to $CC(weak)$ (the definition of these metrics can be found in the appendix 9.4). In a plot of $CC(all)$ against $CC(weak)$, a contrast between trials with high values of $CC(all)/CC(weak)$ typically indicates that substructures are at least partially correct. Plots of $CC(all)/CC(weak)$ for the simulated data do indeed reveal a contrast in solutions (as shown in figure 5.2), with the best occurring at a k of slightly less than 1.0, which is consistent with what is observed in synchrotron RIP [35]. Because in this test the "correct" substructure is known from a peak search of the RIP peak map from ANODE, the resultant substructures could also be compared with this reference structure with another program called *phenix.emma* [75], which revealed that substructures were largely correct at a variety of values of k, with a maximum correctness at $k = 0.96$ (see figure 5.3). Taken together, these results indicate that there is adequate signal in a HI-RIP experiment to determine correct radiation damage substructures, in the same manner as RIP. Next, I determined whether, together with the experimentally determined substructures, the differences between high and low fluence datasets were enough to determine phases. The best substructures from each k were input into SHELXE [32] for phase calculation, phase improvement, and model building. Initial figure of merit weighted phase errors (after one round of phase improvement by SHELXE) were very good at 38°, and this improved and converged to 28° after an additional round of substructure and phase improvement. Unlike in synchrotron X-ray and UV RIP, where the main benefit from substructure improvement is the identification of sites with "negative" electron density in RIP difference maps caused by, for example, side chain rearrangements, these rearrangements are not anticipated on the time scale of an SFX experiment. Therefore in this case, substructure improvement serves the purpose of identifying weaker sites in the $F_{\text{lowfluence}} - F_{\text{highfluence}}$ map that were not identifiable initially by SHELXD. Having established that, in our initial testing conditions, standard RIP analysis could be

HF (photons/μm²) \ LF (photons/μm²)	10^{11}	$5 \cdot 10^{11}$	10^{12}	$5 \cdot 10^{12}$
$5 \cdot 10^{11}$	XXX			
10^{12}	XXX	XXX		
$5 \cdot 10^{12}$	20,000	30,000	40,000	
10^{13}	10,000	20,000	20,000	80,000

Table 5.1: Number of patterns needed per dataset to achieve a wMPE better than 40 degrees. HF= "high fluence", LF="low fluence".

used to both determine radiation damage substructures and to produce interpretable electron density maps, I studied the effects of various changes to the experimental conditions. Since the number of patterns that can be collected during an X-ray FEL experiment is often limited due to practical reasons such as limited beam time and/or sample, and the fact that 10^{13} photons/μm^2 is a current upper limit for the photon density in a single FEL pulse, I simulated the effect of both parameters on phasing. I found that substructure solution and correct phases could be achieved with high fluences of $5 \cdot 10^{12}$ photons/μm^2 even down to $20,000$ patterns (see for example the purple curve in the wMPE of figure 5.4), a value that is comparable to the average number of patterns required for solving SFX structures with standard methods [52, 54]. With a very large number of patterns ($n > 300,000$), slightly lower fluences of $1 \cdot 10^{12}$ photons/μm^2 might be used as well. The critical parameter is not the ratio of fluences, but rather the difference between fluences. This is because the effective scattering strength of sulfur does not decrease linearly with increasing fluence, but its relative change is the highest between $5 \cdot 10^{12}$ photons/μm^2 and 10^{13} photons/μm^2(see figure 2.6). In general, a larger number of patterns improves the quality of the HI-RIP solution because of improved averaging of errors in the Monte Carlo integration of intensities. Moreover, the RIP peak height is a good predictor of the phasing success: for these simulations, a RIP peak height of at least $16 - 17\ \sigma$ led to a good phasing result.

Table 5.1 summarizes the results of the simulations, reporting the number of pattern needed to achieve a RIP solution with mean phase error smaller than 40 degrees, as a function of the low and high fluences.

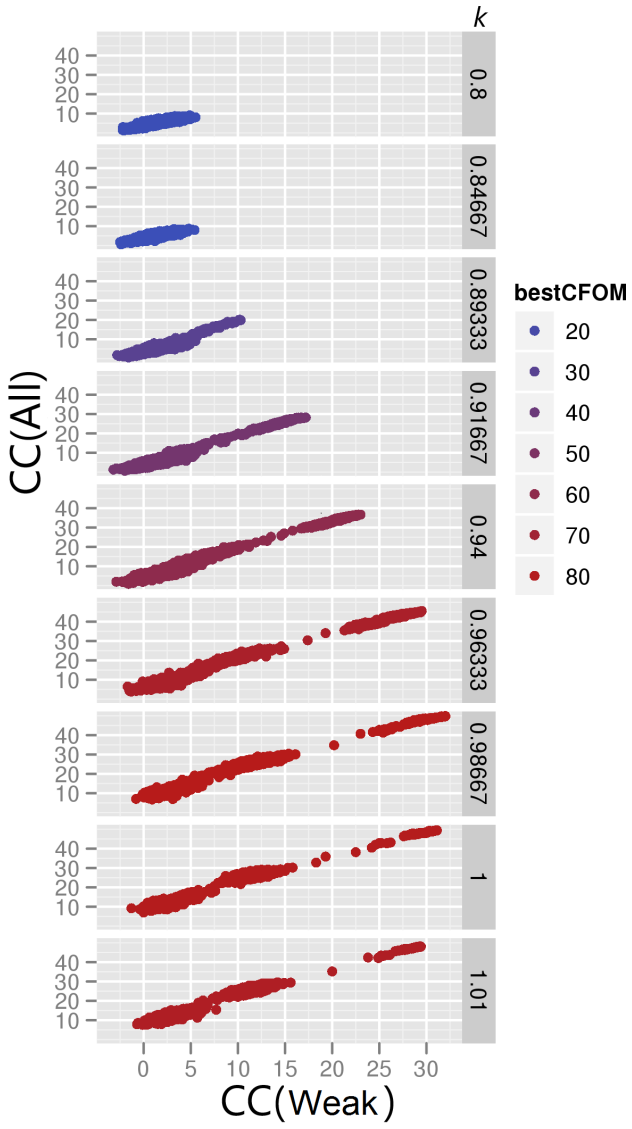

Figure 5.2: Example of the improvement of the SHELXD correlation coefficient ratios ($CC(all)/CC(weak)$) as a function of the scaling coefficient k (shown on the right side of each plot). The intensity for the high fluence data set was $1 \cdot 10^{13}$ photons/µm^2, and $1 \cdot 10^{11}$ photons/µm^2 for the low fluence. Each set consisted of $100,000$ patterns.

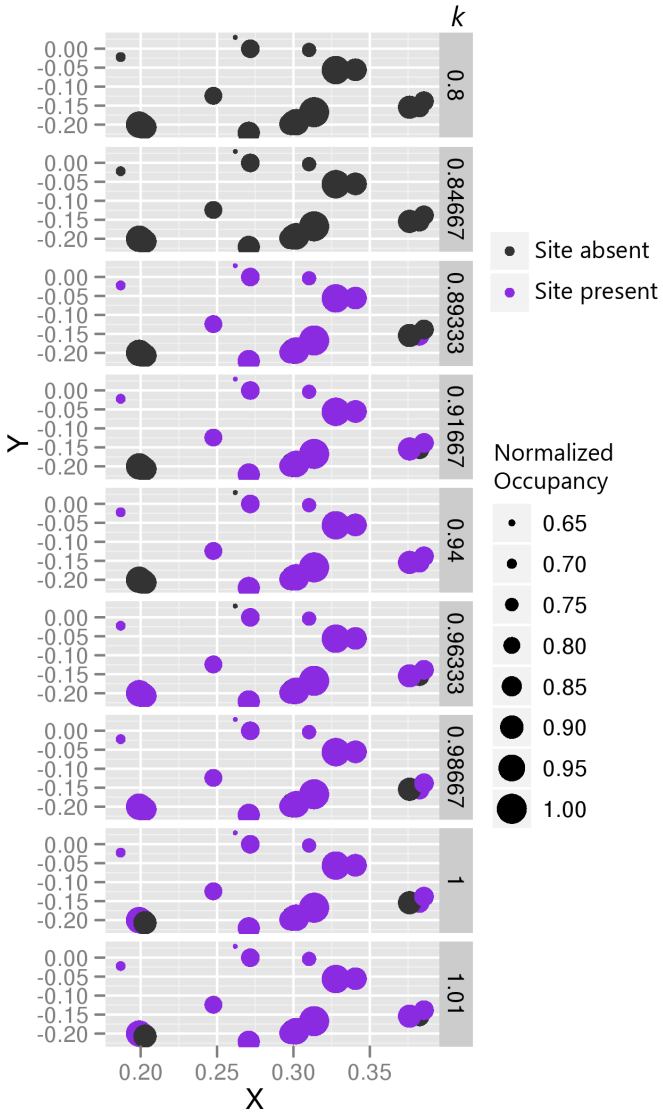

Figure 5.3: Correctness of SHELXD substructures. The substructure with the highest $CC(all)/CC(weak)$ was compared to a reference substructure from ANODE using *phenix.emma*. Purple circles indicate a correctly identified atom. Black circles indicate that a site was not identified. "X" and "Y" are the fractional unit cell coordinates of the sites, as a fraction of the X and Y axis. The diameter of each site represents the RIP peak hight, and is normalized to the most intense peak height in the difference map.

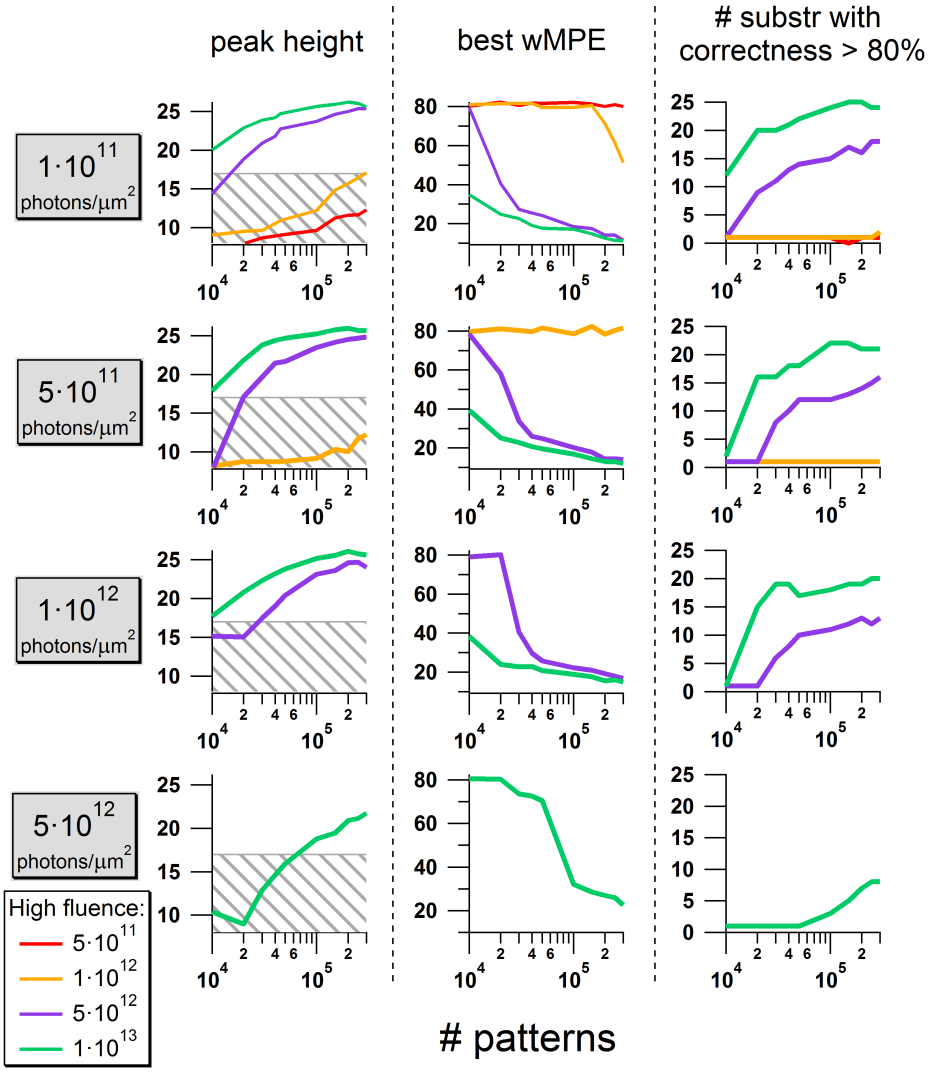

Figure 5.4: RIP peak height, best weighted mean phase error (wMPE), and the number of substructures having a correctness of greater than 80%, as a function of the number of patterns used. The corresponding low fluence is written in the grey boxes on the left-hand side, while the high fluence is shown with different colors (see legend on the bottom-left corner).

5.3 Simulation of particular experimental conditions

5.3.1 Simulations of flow-aligned crystals

Experimental evidence showed that crystals of needle-like or cylindrical shape are prone to flow align along the jet direction. Since the crystallographic axes are often related to the ones of a cylinder (i.e. one of the crystal axis will coincide with the crystal growing direction), the orientation of the collected diffraction pattern will not be purely random, but some Bragg peaks will intersect the Ewald sphere more often than others. CatB crystals fall in this class (as can be seen in section 6.1). This also means that the multiplicity and at the same time the convergence of the Monte Carlo integration will differ from the complete random case considered in this chapter. Further simulations were then performed to mimic the flow aligned case, by assuming a fixed liquid stream direction (set along the $\hat{\mathbf{z}}$ direction) and by allowing only certain crystal orientations. In particular, since the crystallographic c axis of the CatB coincides with the long axis of the rod-shaped crystals, the magnitude of the $\mathbf{c}^*$ component along the $\hat{\mathbf{z}}$ direction was taken to be greater than 80% of $|\mathbf{c}^*|$, while the $\mathbf{a}^*$ and $\mathbf{b}^*$ were completely free, allowing rotations. At these conditions, the RIP peak height was found to be up to 10 σ higher than for the random orientation case, allowing to better locate the sulfur substructures with fewer patterns. This is probably due to the higher SNR of some Bragg reflections, which are recorded more often than if the crystals had random orientations. Also the RIP phasing solution converged more rapidly as a function of the number of patterns considered, indicating that flow aligned crystals could be well suited for a HI-RIP phasing experiment.

5.3.2 Simulation of crystals with identical orientations

The most extreme HI-RIP experiment consists of two data sets collected on the same set of crystals, at different fluences. In this condition, the orientation of the hit crystals will be the same. This case has no immediate experimental approach, since it is almost impossible to reproduce an experiment where a crystal can survive the intense FEL beam without being vaporized, or where thousands of crystals can be set to a known orientation before the interaction with the X-rays.

Simulation were carried out using a single list of seeds to generate the crystal orientations, instead of a random one. Similarly to the flow aligned case, the results present a very fast convergence of the phasing solutions, faster than the previous

cases considered. This is indeed expected, since the partiality and the convergence of the integration is identical between the two sets of data, so these errors are mutually excluding.

5.4 Discussion

The success of the HI-RIP technique relies on the relative strength of the scattering factors of the heavy atoms at the two fluences used, and on the accuracy of the structure factor measurement. The former is mainly related to the specifications of the experimental facility, such as the available photon flux, its shot-to-shot variation, and the spatial photon distribution inside the beam profile. The latter is influenced by errors introduced by the SFX technique, such as reflection partiality, inhomogeneous crystal size or quality, and non-isomorphism. The Monte Carlo integration of intensities can help to average out these error sources, but a large number of observations are required. Despite noise has been added to the simulations, which yields data with merging statistics similar to those found with experimental data, the several assumptions made in the simulations might be difficult to achieve using current FEL technology. First, it is possible that the shot-to-shot variation in intensities might contaminate the high-fluence diffraction data with data recorded at lower fluences, reducing the ionization contrast. This could be avoided by using measurements of the pulse intensity from a beam diagnostic monitor after the interaction region, sorting the diffraction snapshots as a function of the pulse fluence (a similar attempt is made for the experimental data presented in chapter 7 using a beam intensity monitor located at the entrance of the endstation). Second, the actual beam profile is not a top hat function but has wings of lower intensity surrounding the focus. Crystals which pass through these regions would be exposed to lower intensities than those passing through the focal spot. Currently there do not exist published beam profiles for either of the optical layouts at CXI. We have, however used our workflow to make crude estimates of the effect of a more Gaussian beam profile, by considering high fluence datasets created by averaging intensities from snapshots at three different photon flux densities: $10^{13}\,\text{photons}/\mu\text{m}^2$, $5 \cdot 10^{12}\,\text{photons}/\mu\text{m}^2$, and $10^{12}\,\text{photons}/\mu\text{m}^2$. Since the low fluence data are only affected by a negligible ionization effect, we used a single set at $10^{11}\ photons/\mu m^2$. We saw that, using $50,000$ patterns, phasing is still possible as long as less than 40% of the reflection intensity measurements are generated from the tail of the beam. In particular, a dataset composed of $20,000$ patterns at the highest fluence, $20,000$ at medium, and $10,000$ at low fluence showed a RIP peak height of about 21 and a best wMPE

of about 20 degrees (roughly twice as large as the pure beam case). Doubling the amount of low fluence patterns at the expenses of higher fluence data reduces the RIP contrast much below 16, so that no phasing is possible. This indicates that the HI-RIP method can tolerate a more gaussian beam profile. The simulation, however, does not fully model the case of a gaussian beam of the same (or smaller) size as the crystalline sample, where the single diffraction pattern will contain both high fluence and low fluence scattering. This condition has to be modeled with a new set of equations, by introducing different electronic responses to various X-ray fluences over the sample. Experimentally, these problems could potentially be avoided by using an X-ray beam size much larger than the crystal size. Finally, to treat the ionization dynamics in the present work, we have used an independent atomic model, ignoring charge rearrangement with neighboring atoms. The molecular environment will affect all anomalous coefficients calculated within the independent atomic model. In our model, we also neglect resonant absorption processes, shakeup or shakeoff processes [86], and collisional ionization [87], which induce further ionization. These assumptions might cause a discrepancy between simulated data and experimental data and will be incorporated into future simulations.

Chapter 6

HI-HIP experiment using a native protein

The sulfur single-wavelength anomalous diffraction (S-SAD) phasing method allows the determination of native protein structures without requiring chemical modification or an homologous structure to be known. This *de novo* phasing technique presents, however, different problems connected to the long wavelength at which the sulfur K-edge lies (2.47 keV). To resolve near-atomic-resolution structural features, a diffraction experiment must be carried out at a photon energy higher than 6 keV, since the data are usually recorded in the forward scattering direction (so the 2θ angle does not exceed 90°). At this wavelength, the Bijvoet differences upon which the S-SAD phasing relies are very weak: for an average protein containing a single sulfur atom every 30 residues, the difference is about 2% at 6 keV (calculated using equation 1.10). Low anomalous signal requires very accurate data collection, which often means longer acquisition times and consequently a higher risk of radiation damage effects (that are more severe at longer wavelength). The difficulty of the technique is also proven by the very low number of deposited structures solved with S-SAD, compared to other X-ray methods (133 Vs. $89,367$[1]). All these difficulties might be overcome by the use of high-intensity X-ray FEL radiation, exploiting the preferential bleaching of the sulfur atoms at energies higher than the sulfur K-edge, and using the serial crystallography method, as described in the previous chapter. In particular, the anomalous signal (intended here as the f' component) can be in-

[1]Results from queries dated 12/06/14 using the Protein Data Bank website (www.pdb.org). For S-SAD, the advanced search "Structure Determination Method" was used, in combination with "Text search". Due to the absence of a well defined S-SAD search criteria, this number could be inaccurate.

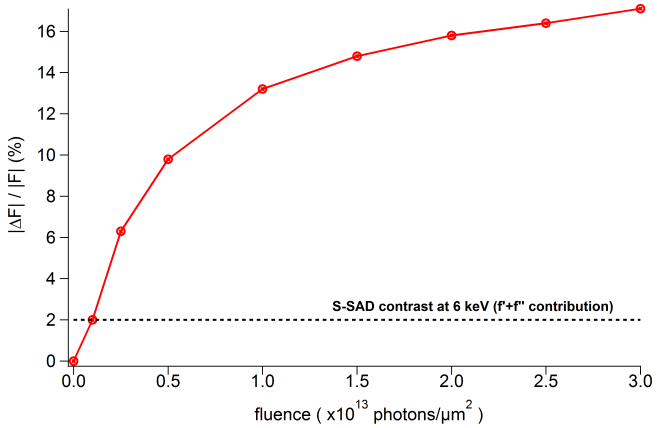

Figure 6.1: Plot of $< \Delta F > / < F >$ as a function of the pulse fluence, calculated for lysozyme at 6 keV, compared to the SAD contrast at low fluence, for the same wavelength.

creased at high intensities, overtaking the Bijvoet differences of the S-SAD as can be seen in figure 6.1. The acquisition of data under various X-ray pulse conditions could also allow the creation of a set of linearly-independent equations of the form 2.6 that are overdetermined as in the case of MAD. Moreover, the technique is not limited by the chosen wavelength, so one can utilize longer wavelengths without increasing the radiation damage (due to the femtosecond pulses of the FEL) or at shorter wavelengths to collect higher-resolution diffraction (provided that the pulse fluence is still sufficient to induce moderate ionization effects).

This phasing methodology, that I will refer to as "HIP", for "high intensity phasing", has the power to become the method of choice for native protein structure determination at X-ray FELs. Differently from the previously defined HI-RIP method (see the previous chapter), HIP requires the previous knowledge of the high intensity anomalous coefficients of the heavy atoms, and makes use of the Karle-Hendrickson equation to directly solve the phase problem.

In the previous chapter, SFX simulations showed that the available FEL sources can provide enough photon flux to ionize the sulfur atoms and that, by reducing the highest accessible flux by about two orders of magnitude, it is possible to utilize the HI-RIP technique with the conventional RIP (or SIR) phasing workflow. In this chapter, a first attempt of a HIP/HI-RIP experiment on native protein is described.

6.1 The in-vivo grown Cathepsin B crystals

The native protein sample employed for this experiment is the Cathepsin B: an enzyme belonging to the class of cysteine proteases, which degrade polypeptides. The form of the enzyme used is specific to the *Trypanosoma brucei* (TB) parasite, and it has been identified as a potential drug target for the treatment of the sleeping sickness disease, which affects about 60 million people in central Africa [88]. The glycosylated form of TB-CatB has been recently solved with the SFX method [53]. For this fortunate experiment, rod-shaped microcrystals were used, which were grown with an *in-vivo* method inside insect cells. In the *in-vivo* crystallization technique, a certain type of host cell is infected by a genetically-modified virus carrying the protein expression gene inside its viral genome. The infected cell can then assemble the protein, producing a crystal which is limited in dimension by the cell size and by the amount of protein that is produced. The CatB crystals can be grown from *Spodoptera frugiperda* insect cells infected by a recombinant *Baculovirus*. These crystals have typical dimensions of $5 - 15\,\mu$m in length and about $0.9\,\mu$m in width (see figure 6.2), and they survive in an aqueous suspension after purification from the mother cells. This makes CatB crystals a good sample for SFX, due to the aptitude of producing a stable liquid jet from watery solution and due to the homogeneity of the crystal size. Needle-like shaped crystals, indeed, when pushed in a liquid solution through the aperture of a GDVN, tend to align themselves along the liquid flow direction, so that the long axis of the crystal lies - within a few degrees - along that direction. This also means that the X-ray beam will intersect the crystal along a direction perpendicular to the liquid flow, penetrating most of the times a similar amount of unit cells. Since for the CatB the long axis of the crystal corresponds to the c-axis of the (tetragonal) unit cell, it is possible to have a visual proof of the flow alignment by plotting the orientation vectors: in the case of a flow-aligned crystalline sample, the vectors will form a cluster around the line defined by the liquid stream direction. An experimental proof if shown in figure 6.3, where the calculated orientation vectors are plotted wogether with the liquid stream direction, retrieved from a beamline microscope image. A direct consequence of flow alignment is that some Bragg peaks will be in diffraction conditions more often than others, so the multiplicity will not be homogeneous over the observable volume of the reciprocal space.

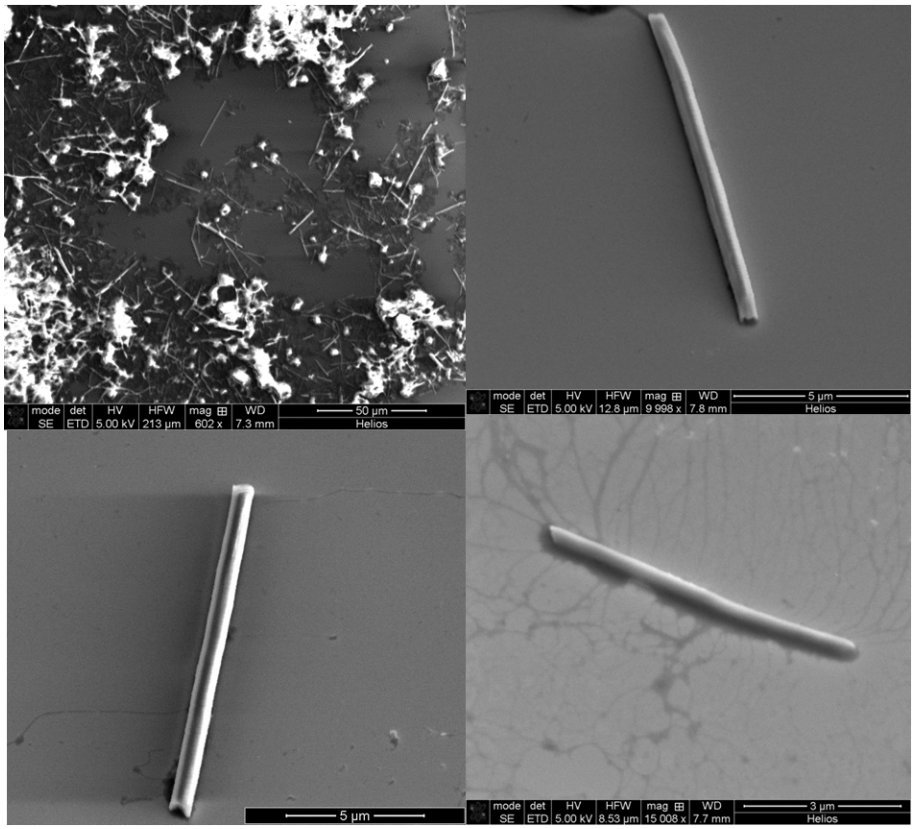

Figure 6.2: CatB crystals imaged with a Scanning Electron Microscope. The top-left image shows the typical crystal concentration, from a dried drop of solution. Courtesy of Francesco Stellato.

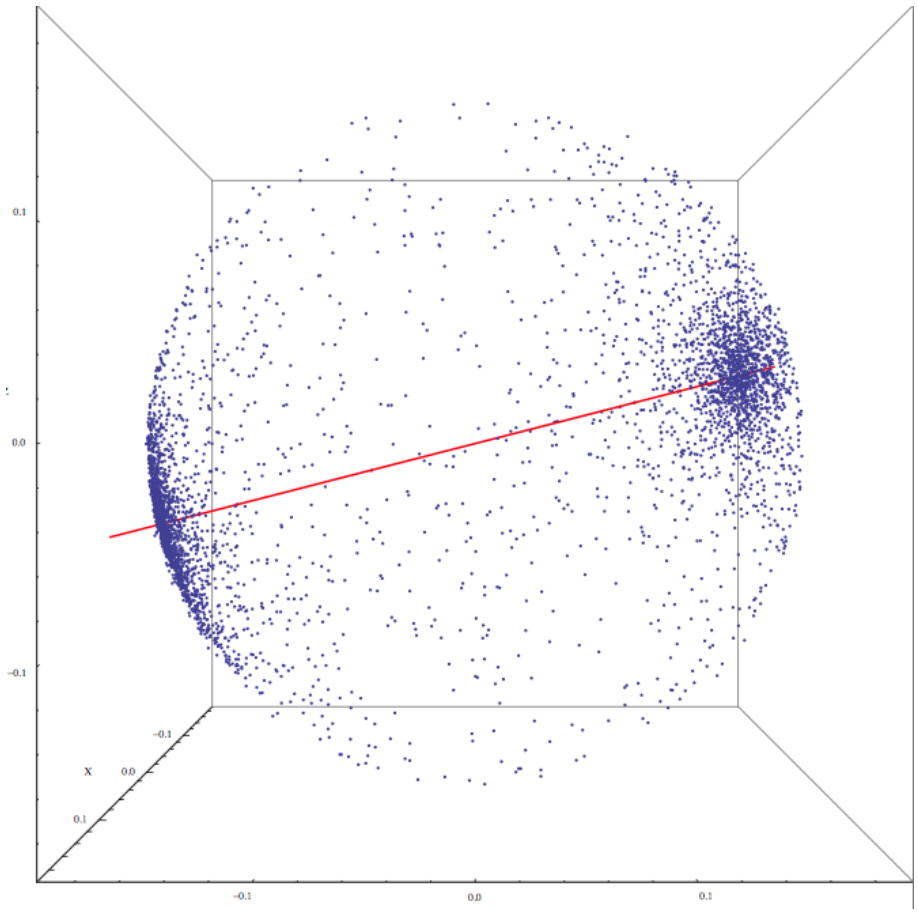

Figure 6.3: The sphere of orientation vectors of CatB crystals from a SFX experiments. Each blue dot represents the reciprocal space vector $c*$. The clusters of dots indicate flow alignment along the direction of the liquid jet, retrieved from an image of the jet, recorded with a beamline microscope, and represented as a red line.

6.2 The experiment

The first attempt of a HIP experiment was carried on during the L669 LCLS beamtime (June 2013), using the 0.1 μm sample chamber of the CXI instrument. Two SFX datasets were collected on CatB crystals at different X-ray fluences, using 6 keV photons: a first set was collected with an attenuated beam, using silicon filters of different tickness to reduce the flux of the FEL beam, with resulting transmissions between 1% and 27%. A second dataset was instead collected at full photon flux; in this case the detector was protected by a tailor-made attenuator, placed few centimeters after the interaction region, to reduce the strong scattered signal that could have damaged the CSPAD. The transmission of this attenuator was 25%, constant over the 2θ scattering angle.

The beamline efficiency was estimated to be around 20% at 6 keV, meaning that the average pulse intensity during the experiment was about 0.5 mJ at that wavelength. The loss of beam energy is mainly due to the divergence of the LCLS beam (produced about 384 m away from the endstation), which causes the X-rays to fall off the beamline mirror aperture. Other important sources of energy loss are: the transmission efficiency of the KB mirrors and the absorption of a 150 μm thick diamond window on the front of the sample tank. Assuming a focal spot size of 0.2 μm in diameter, the pulse fluence was about $4 \cdot 10^{12}$ photons/μm^2, in the absence of further attenuation.

Crystals were flown across the X-ray beam in a pre-filtered water solution containing approximately 10^9 crystals/ml, using a liquid jet of about 2 μm of diameter and running with a flow rate between 10 and 25 μl/min. 60 fs long FEL pulses were focused onto the liquid stream, about 50 μm away from the nozzle tip. At these experimental conditions, about 3% of the recorded frames contained diffraction from randomly-oriented CatB crystals.

6.3 Data analysis

A total of 101,080 images were identified by *Cheetah* as crystal hits and further processed by *indexamajig*. Indexing trials were initially performed to optimize the unit cell parameters, starting from the values of a previously-deposited structure (PDB code "4HWY") [53], and to refine the detector geometry, as explained in the next subsection. Random diffraction patterns from each recorded run, of the duration of about 5 minutes, were visually inspected to find possible regions of the detector to exclude; these excluded regions contained, for example: bright scattered

streaks from the water jet close to the central beam hole, shadows from the nozzle tip at high resolution, or features from the post-sample attenuator. 69,925 diffraction patterns were successfully indexed by *indexamajig* (using the "mosflm" and "dirax" algorithms) after these careful inspections. Two separated reflection lists were created, dividing the data in high fluence (HF) and low fluence (LF) according to the expected radiation intensity at the interaction region. Monte Carlo integrated intensities were computed with *process_hkl*, keeping the Friedel mates in the asymmetric unit as separate reflections and requiring that only reflections observed more than 10 times were integrated and recorded. Table 6.1 shows the detailed statistics of the two datasets collected as well as the values of some quality metrics determined from the final reflection list, after the Monte Carlo integration. It can be seen from Table 6.2 that the diffracted intensity at high resolution is weak and it is affected by a high level of uncertainty; the I/σ level, in particular, is lower than 2.0 at 3.26 Å and the R_{split} starts to increse rapidly at about that resolution.

6.3.1 Geometry refinement

One of the most important steps of the analysis workflow is the geometry refinement. This procedure aims at retrieving the correct locations of each of the detector tiles with respect to the X-ray beam direction, and the sample-to-detector distance. There are two main methods for performing the refinement, each of which can be started only after the collection of a SFX dataset (usually performed during the first hours of the experiment using a highly-concentrated solution of well diffracting protein crystals, such as lysozyme). The first of these refinement procedures consists in a "manual" adjustment of the coordinates of the single ASICs: *Cheetah* can sum the found hits into a single image, reproducing the "virtual powder pattern" that, for a sufficiently large dataset, will show diffraction rings (as in figure 6.4) similarly as in a conventional crystalline powder diffraction experiment. These rings will span several tiles and they can be used as a visual reference. The detector geometry can be adjusted to make the rings accurately circular. The *hdfsee* program in CrystFEL has specific tools for this purpose.

At very low resolution the rings can overlap, while at high resolution they usually are barely visible, so the metrology derived from a powder pattern will have moderate errors. A program currently under development in the *Coherent Imaging Division at CFEL*, can tackle the task of the geometry refinement automatically, using the information about the indexed patterns contained in the stream file produced by *indexamajig*. This program calculates the distance between the center of

	High fluence (HF)	**Low fluence (LF)**
Wavelength (Å)	2.066 (6 keV)	
Pulse fluence (photons/μm^2)	$4 \cdot 10^{12}$	$4 \cdot 10^{10} - 1 \cdot 10^{12}$
Corresponding dose	37 GGy	$0.37 - 1$ GGy
Space group	$P4_22_12$	
Cell dimensions a, b, c (Å)	124.4 124.4 53.9	
Number of "hits"	53,733	47,347
Number of indexed patterns	37,389 (69.6%)	32,536 (68.7%)
Highest resolution (Å)	3.26	3.26
Completeness	100% (100%*)	100% (100%*)
I/σ(**I**)	5.19 (1.86*)	5.86 (1.76*)
$\mathbf{R_{split}}$ (%)	18.1 (55.3*)	14.8 (59.1*)
CC (%)	0.96 (0.54*)	0.97 (0.54*)
Redundancy	541 (513*)	615 (566*)

Table 6.1: SFX data statistics. The metrics were calculated with CrystFEL, considering the Friedel pairs as distinct reflections. The values with * refer to the highest resolution shell.

1/d (nm^{-1})	**d (Å)**	**Compl. (%)**	**Redund.**	**I/σ(I)**	**R_{split}** (%)	**CC** (%)
0.835	11.98	100	659	13.77	7.8	0.98
1.430	6.99	100	509	8.26	13.0	0.94
1.705	5.87	100	575	7.23	14.3	0.93
1.909	5.24	100	619	6.75	15.1	0.93
2.076	4.82	100	627	6.44	15.4	0.92
2.22	4.51	100	574	5.85	17.1	0.90
2.347	4.26	100	516	5.13	20.1	0.86
2.461	4.06	100	541	4.42	22.4	0.85
2.566	3.90	100	456	3.85	27.9	0.77
2.662	3.76	100	457	3.42	30.2	0.73
2.753	3.63	100	508	3.12	32.5	0.73
2.837	3.52	100	518	2.78	36.8	0.68
2.917	3.43	100	513	2.54	38.6	0.73
2.993	3.34	100	531	2.32	46.1	0.57
3.065	3.26	100	513	1.86	55.3	0.54
3.145	3.18	100	466	1.68	60.7	0.49
3.215	3.11	100	454	1.20	79,6	0.40

Table 6.2: SFX data statistics for various resolution shells, for the LF dataset. In red is indicated the highest resolution at which the data were truncated.

each found diffraction peak and the location of the closest predicted Bragg peak. This information is used to compute the average displacement of each pair of ASICs (since each ASIC in the pair is physically connected to the other), and the geometry that minimize this displacement is output. In figure 6.4, an example of the residual disagreement after the manual geometry refinement is shown for this particular experiment; it can be seen that for the central tiles (containing the highest statistics) the predicted peak locations are on average $2-3$ pixels away from the corresponding Bragg peak. For big enough datasets, the geometry refinement program can also attempt to rotate the tiles in order to find the best agreement between the predicted and the found peaks location. This automatic procedure is usually iterated to further refine the metrology, since it can generally improve the number of indexable patterns.

6.4 Substructure determination and phasing attemts

In order to measure relative differences due to the photoionization processes, the scattered intensities of the two datasets have to be set on a common scale. The scaling was achieved with CCP4 *Scaleit* [33], by treating the low fluence data as a derivative set and the high fluence as native, and by applying the scaling function to the former. The best scaling was found using a final Wilson scaling after the step of least-squares determination of isotropic temperature factors.

A first proof of difference signal can be obtained from the phased difference map, i.e. a map where the difference of the structure factors of the two datasets is plot using the phases obtained from a molecular replacement run, performed using *Phaser* [74]. The known search model is the previously-deposited 4HWY.pdb, and its refined version can be superimposed on the map as a visual reference, as in figure 6.5. This figure shows that some of the most intense regions of the map corresponds to sulfur positions, in particular to CYS 158, CYS 215, and MET 138. The highest of these peaks goes to about $7\,\sigma$.

6.4.1 Estimation of ionization from occupancy

The estimation of the occupancy of the sulfur sites from an occupancy refinement process could be used as a possible indicator of the change in scattering strength, localized on these atoms. The photoionization, indeed, results in a loss of scattering

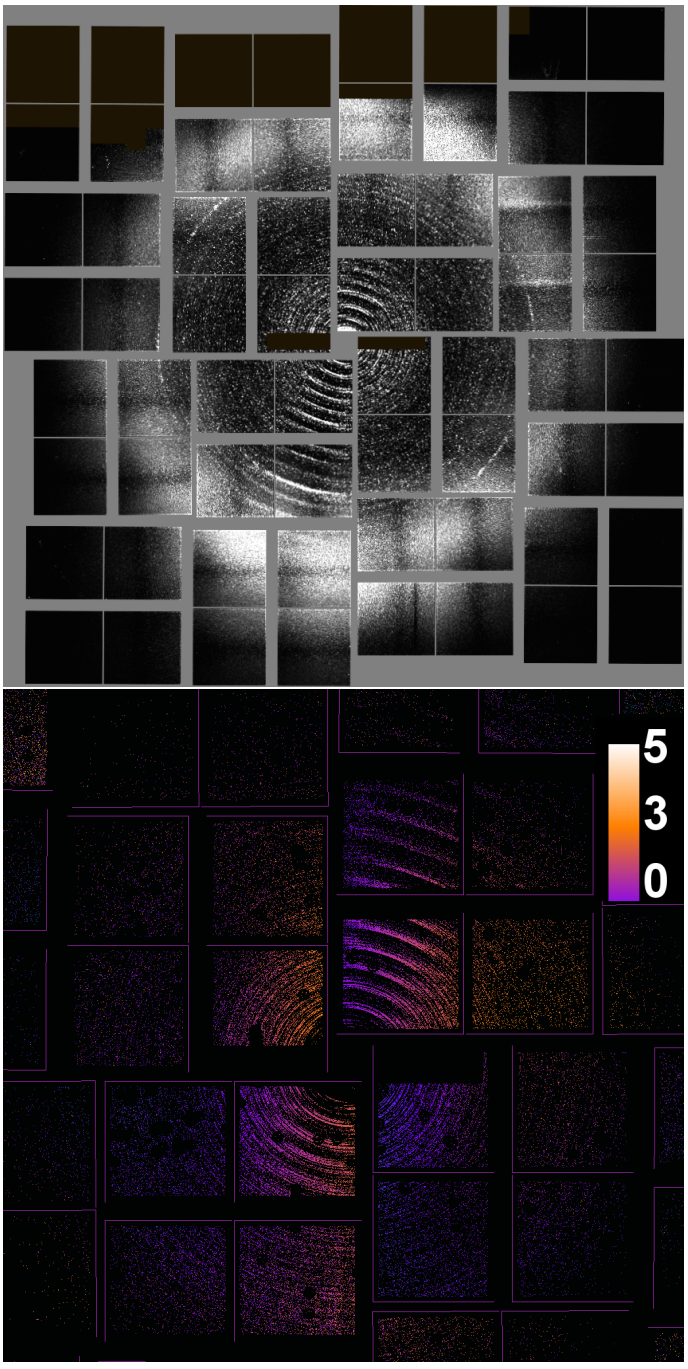

Figure 6.4: Top: example of virtual powder pattern obtained by summing $4,000$ diffraction patterns. Bottom: zoomed-in section of the average displacement map of the found peaks from the predicted Bragg peaks location, using the geometry correction algorithm. The average displacement is represented in number of pixels, as shown in the color bar.

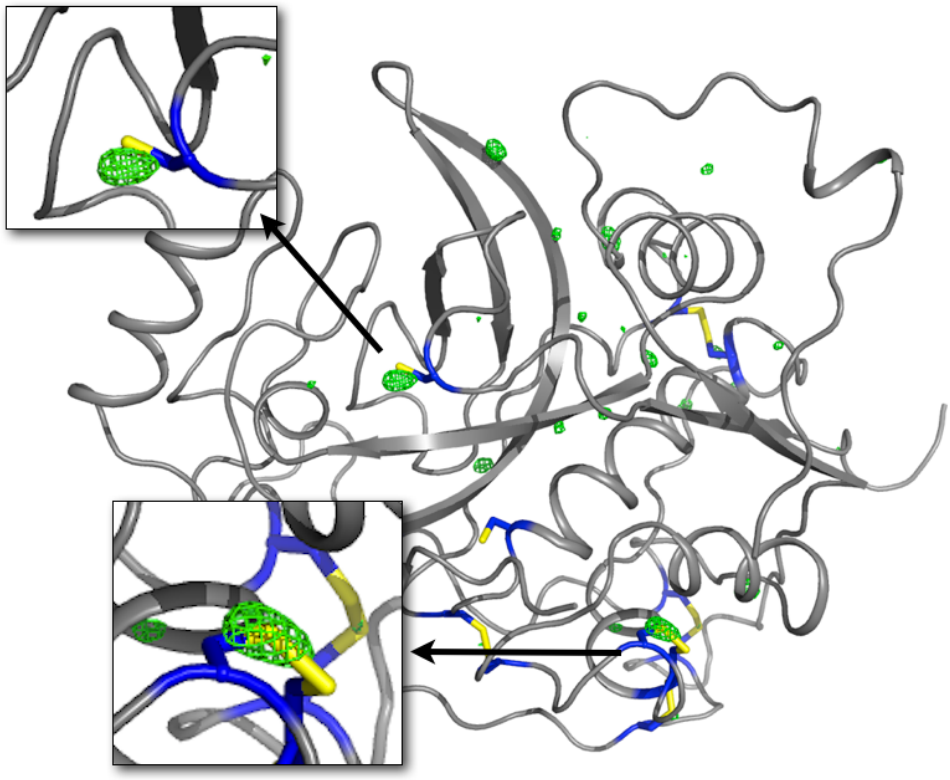

Figure 6.5: RIP map countoured at 4.5 sigma, superposed to the CatB model. Sulfur atoms are represented by yellow sticks.

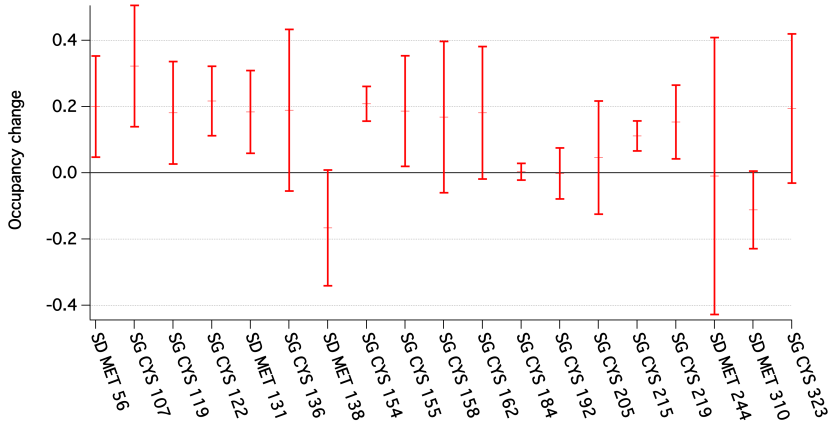

Figure 6.6: The difference in occupancy of the sulfur sites from the two datasets (low fluence - high fluence). The labels in the x axis are consistent with the ones used in the deposited structure.

power of the bleached sulfur species which can be thought as if a pristine sulfur presents a lower occupancy. In the absence of high intensity bleaching effects, the occupancy of the sulfur sites should be highest (in theory, close to unity), while the relative occupancy change between low and high fluence should be proportional to the number of lost electrons.

Several occupancy refinement attempts have been made using *REFMAC*5 [89], considering an incomplete occupancy of the S sites of the previously refined model, solved with molecular replacement. This refinement has been found to be in some cases insensitive or, to the contrary, strongly dependent on the initial occupancy set in the input model, depending on the site studied. For these reasons, 10 parallel refinement runs were carried out, where the initial occupancy differed from 0 (absence of S) to 1.0 (full occupancy), at 0.1 steps. The resulting occupancies from those runs were averaged and the standard deviation was calculated. This procedure was performed on both the collected datasets, and the results were expressed for the single sulfur sites. Figure 6.6 shows the average occupancy difference between low and high fluence, for each of the sulfur sites of the CatB (shown in the x axis). The image shows an average positive difference, consistent with an increased ionization at high fluence. Averaging all the contributions, the loss of occupancy is calculated as $0.12 \pm 0,15$, meaning an average ionization of $1.9 \pm 2.4e^{-}$.

6.5 Discussion

The simulations reported in the previous chapter suggests that the experiment was mostly limited by the low number of indexed diffraction patterns, and by the relatively low resolution of the datasets, which prevented any *de novo* phasing approach. Nonetheless, the difference signal from the phased difference map gave the first experimental evidence of the change in scattering strength of some of the sulfur sites. The occupancy refinement was found to be not accurate enough to provide a clear proof of ionization change, because of the relative low resolution of the data and the sensitivity of the refinement to other parameters, such as the B-factor or the constrains of the aminoacid sequence. In particular, different sulfur sites presented a different change in occupancy, which is not anticipated by the theory, and the convergence of the refinement did often depend on the initial conditions. The inaccuracy of the retrieved occupancies is reflected in the large uncertainties shown in figure 6.6, and on the large error associated to the calculated ionization change, which cannot rule out a zero contrast.

As will be shown in the next chapter, by using particular criteria for selecting the best diffraction patterns, the difference of the average scattering strength of the heavy atoms between the LF and HF sets can be increased. This selection reduces the number of usable patterns, so it requires initial datasets containing a high number of images: a condition that was not available in this experiment. Another criteria to classify patterns according to the high fluence effects can be to look at the intensity of the Bragg reflections which are more sensitive to the scattering strength of the heavy atoms. This approach is however limited by the reflection partiality, an important factor that is still not entirely understood.

As discussed previously, further improvement may come using extra diagnostics connected to the SFX experiment, that allow to have a shot-by-shot information about the real fluence impinging on the crystal, such as a beam intensity monitor after the interaction region, to measure the transmitted beam intensity, a time of flight spectrometer to collect the information about the ion species created after the interaction between the FEL pulse and the crystal, or a simultaneous measurement of the fluorescence.

Chapter 7

HI-RIP experiment using a high-Z atomic species

Due to their large interaction cross section, heavy atoms are particularly affected by photoionization. The cross section is the highest when the X-ray wavelength is close to an absorption edge. As a rule of thumb, the bleaching effects due to the high X-ray intensity will be the highest if the dataset is collected in resonant condition with the absorption edge of that heavy element. Despite the fact that the ionization can reduce the out of phase contrast, the collection of a low fluence dataset can allow the use of standard anomalous phasing methods, while the combination of a low and high fluence sets can be used for an HI-RIP approach. Such an experiment was performed at LCLS during the last shift of the LA06 beamtime, using gadolinium atoms bound to chicken egg white lysozyme molecules. Here I show experimentally that the photoionization effect allows retrieval of the Gd positions from the difference between two datasets of a Gd derivative of lysozyme microcrystals, collected at high and low X-ray fluences, using a single wavelength just above the L III absorption edge of gadolinium.

7.1 Materials and methods

Rod-shaped microcrystals ($\leq 1\times \leq 1\times \leq 2\,\mu\mathrm{m}^3$) of chicken egg-white lysozyme (SIGMA, Schnelldorf, Germany) were grown as described in [52] and stored in a stabilization solution consisting of 8% NaCl in 0.1 M sodium acetate buffer, pH 4.0. At least 30 minutes prior to data collection, 100 mM gadoteridol (Gd^{3+}:10-

(2-hydroxypropyl)-1,4,7,10-tetraazacyclododecane-1,4,7-triacetic acid) was added to the crystal suspension. This compound contains a Gd atom, and two gadoteridol molecules can be incorporated per asymmetric unit [90]. Before injection, the crystals were left to settle at the bottom of a 15 ml Greiner tube after which the supernatant was removed until the volume of packed crystals was a third of the total volume. Then, the crystals were resuspended by gentle agitation and injected into the 200 nm focus of the CXI instrument at the LCLS using a liquid jet of 4 μm diameter running at 25 μl min^{-1}. A rotational anti-settling device [91] equipped with a thermostat kept the crystal suspension homogeneous by mixed and at 20°C.

SFX diffraction snapshots were collected at 120 Hz using the CSPAD, which was placed 11.5 cm from the interaction region. Lysozyme microcrystals were hit stochastically by 40 fs duration, 8.48 keV X-ray pulses. Two different data sets were collected over two 12 hr shifts: a first "low fluence" (LF) dataset was recorded with the X-ray beam attenuated to 1.73% of its full intensity. A second "high fluence" (HF) dataset was then collected with the unattenuated beam. To protect the detector from the damage due to the high intensities of some of the diffracted beams, a 240 μm thick flat Si attenuator was placed behind the interaction region. The average FEL pulse energy during the experiment was 1.6 mJ: assuming a beamline transmission of 30% and a perfect Gaussian spot of 0.2 μm FWHM, the estimated peak X-ray fluence in the interaction region is $7.8 \cdot 10^{12}$ photons/μm^2 for the unattenuated beam and $1.3 \cdot 10^{11}$ photons/μm^2 for the low fluence dataset, resulting in average doses of 1.27 GGy and 220 MGy, respectively. The photoabsorption cross section for neutral Gd at 8.48 keV is $1.04 \cdot 10^{-5}$ μm, and as such the saturation X-ray fluence for Gd (at which every Gd is photoionised once) is $1/(1.04 \cdot 10^{-5}\,\mu\text{m}) = 9.6 \cdot 10^{10}$ photons/μm^2. That is, every Gd atom could be photoionised once on average during the duration of a low fluence pulse, but high fluence pulses were up to 82 times higher than the Gd saturation fluence. The detector geometry was first calibrated using the virtual powder pattern method, followed by a detector geometry refinement, as described in 6.3.1. For the high fluence dataset, the edges of the detector were masked, to cover the possible scattered signal from the edge of the Si attenuator, limiting the highest resolution to about 2 Å. A total of 983, 180 crystal diffraction patterns were identified using *Cheetah*, with an average hit rate of about 43%. 592, 362 of the hits were successfully indexed using *CrystFEL*. The unit cell parameters were determined utilizing a subset of the collected data. Subsequent indexing was performed comparing the resulted unit cell parameters to the determined ones, allowing a tolerance of 10% in axis length and 2° in angle. The final Monte Carlo integration resulted in two datasets (see tables 7.1, 7.2, and 7.3 for the statistics of the single sets) which were truncated to a resolution of 2.1 Å. The low fluence dataset, which

was not limited in resolution by the applied mask at the detector edges, shows that the observed diffraction was limited geometrically by the detector, as reported in table 7.2. The resolution-dependent attenuation of the Si attenuator was corrected in the HF data set after the Monte Carlo integration process by multiplying each reflection's intensity by the calculated attenuation factor, γ, at the corresponding scattering angle 2θ, as:

$$\gamma = e^{-\frac{\Delta l}{\mu}}$$

with μ the attenuation length of the Si and Δl the angle-dependent thickness, calculated as:

$$\Delta l = \frac{240\ \mu m}{\cos(2\theta)}\ .$$

Structure factors were calculated for the HF and LF data sets using *CCP4 Truncate* [92] with default options. Scaling between datasets was performed with *CCP4 Scaleit*, treating the high fluence data as native and the low fluence as derivative, since the ionized Gd atoms, with fewer electrons, can be considered as lighter elements. To visualize the difference signal of the Gd atoms, a $Fo - Fc$ difference density map was calculated using the lysozyme phases obtained by molecular replacement using the data collected at low photon flux. The phasing was performed to 1.9 Å with *PHASER* [74], followed by few cycles of model building in COOT [76] and REFMAC5 [74], to an $Rfactor$ of 19.9% ($Rfree = 22.2\%$). As search model, the structure of Gd-derivatized lysozyme (Protein Data Bank code 1h87 [90]) was used after the removal of the gadolinium ions. The final model was then checked with *Molprobity* [93]. The difference Fourier map, displayed in figure 7.1, shows two high peaks at the Gd locations. One peak is higher than the other (9.0 σ vs. 6.2 σ), perhaps due to the higher occupancy of the site [90].

7.2 Data analysis

7.2.1 Theoretical considerations

The X-ray ionization dynamics involving various charge states of heavy atoms at high-fluence X-ray beam can be calculated using the XATOM toolkit. Since the HF peak fluence is much higher than the saturation fluence, one may expect that highly charged ions are formed during the X-ray pulse via photoionization. Furthermore, every single photoionization event would knock out $2 - 12$ electrons from the same

Space group	$P4_32_12$
Unit cell parameters	$a = b = 79.2 \pm 0.7$ Å, $c = 39.4 \pm 0.4$ Å
Resolution	$56.0 - 2.0$ Å
Indexed images (low fluence)	$218,598$
Indexed images (high fluence)	$373,764$
Indexed images (high fluence, best)	$121,917$
Completeness* (%)	100 (100)
SFX multiplicity*	4792 (1216)

* Considering Friedel mates as individual measurements

Table 7.1: Data statistics.

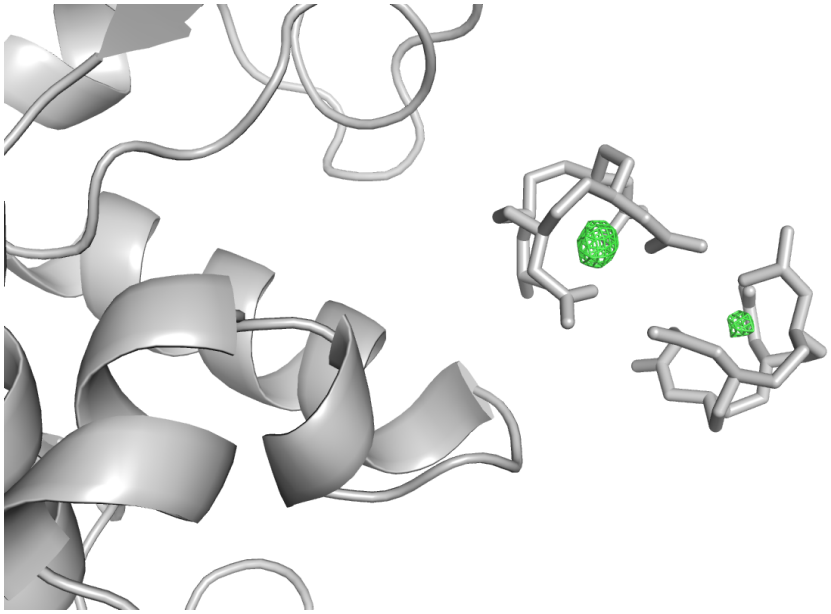

Figure 7.1: Phased difference $Fo - Fc$ Fourier map, superposed to the Gd-lysozyme model. Data to 1.9 Å, contoured at 4 σ.

	LF (218,598 **patterns**)		**HF** (373,764 **patterns**)	
Resolution ($\AA$)	**Redundancy**	$I/\sigma(I)$	**Redundancy**	$I/\sigma(I)$
6.48	3469	30.53	5728	40.47
3.43	3016	25.52	5105	32.70
2.87	3003	22.12	5204	28.99
2.56	2800	19.06	4797	25.25
2.35	3055	17.06	5018	22.28
2.20	2796	13.89	4551	18.42
2.08	1965	9.31	1261	7.58
1.98	1253	6.13	88	1.98

Table 7.2: Quality of the datasets used.

atom via Auger cascade. On the other hand, even though the LF peak fluence is close to the saturation fluence, most of the LF X-ray beam regions do not saturate single photoionization, and most Gd atoms remain intact. In order to compare with experimental results, the effective scattering strength of the heavy atom has been calculated, averaging weighted by the spatial and temporal pulse shape, as:

$$f_{eff} = \sqrt{\frac{\int d^3x \int dt \mathcal{F}(\mathbf{x}) g(t) |\tilde{f}(\mathbf{Q}, \mathcal{F}, \omega, t)|}{\int d^3x \int dt \mathcal{F}(\mathbf{x}) g(t)}} \,,$$

where $\mathcal{F}(\mathbf{x})$ is the X-ray fluence at a given position, $g(t)$ is the temporal pulse shape, $\mathbf{Q}$ the photon momentum transfer, and ω is the X-ray wavelength. The "dynamical" form factor (as defined in [48]) is given by:

$$\tilde{f}(\mathbf{Q}, \mathcal{F}, \omega, t) = \sum_q P_q(\mathcal{F}, \omega, t) \left[f_q^0(\mathbf{Q}) + f_q'(\omega) + f_q''(\omega) \right] \,, \tag{7.1}$$

where P_q is the time-dependent population of the charge state q and f_q^0 (f_q' and f_q'') are normal (anomalous) atomic form factors for the ground configuration of the charge state q. This analysis must take into account the spatial profile of the beam at the interaction region, assumed to be Gaussian with a FWHM of 0.2 μm on a broad

pedestal of much lower fluence which extends much further [94]. The focused part of the beam is considerably smaller than the average width along the crystals' shortest side, 1 μm. When this fully intersects the crystal, the fluence in that intersection volume may be more than 80 times the saturation fluence for Gd, conditions that one would expect to create highly-charged ions from direct photoionisation alone. However, the low-fluence part of the beam may intersect a much larger volume of crystal and contribute to the diffraction signal under lower-ionising conditions. The relative contributions to the total scattered signal from the high and low regions of the beam are given by the ratio of integrated photon counts in those regions (assuming a constant crystal thickness). Although this beam characterization has not been carried out, it was found that the ratio of low and high fluence regions of the focus at another beamline of LCLS did indeed contain comparable numbers of photons [94], much similar to the experimental evidences reported in section 4.2.2 and scheteched in figure 4.4 (for the 1 μm focusing optics).

The spatial beam profile for the LF data collection was the same as for HF and it is expected that the beam-weighted average corresponds to only a proportion of Gd atoms that are singly photoionised, even though the fluence at the focus center is almost equal to the Gd saturation fluence. In the absence of a low fluence pedestal, assuming a perfect Gaussian spot (0.2 μm FWHM) and considering a flat-top temporal shape profile (40 fs), the effective scattering strength of Gd in the forward direction is calculated as 57.4e^- for the LF case and 32.4e^- for the HF case. This sets the highest contrast achievable between two datasets to be 25.0e^- per Gd. This effective scattering strength does not show strong dependence strongly on the temporal fluctuations of the X-ray pulse, but it is sensitive to its spatial fluence distribution. For example, if the spatial distributions is modeled by a double Gaussian shape (50% in the central region and 50% in a broad background with only 0.6 μm FWHM), the effective scattering strength increases to 58.6e^- for the LF case and 43.6e^- for the HF case, providing a difference of around 15 electrons. The cascade of collisional ionization leads to a much greater ionization of not only Gd atoms, but all atomic species in the sample, and can reduce the contrast of the heavy atom ionization. The highest-energy photoelectrons are from the light atoms (which have low binding energies). For example the photoelectron energy from carbon atoms is 8.2 keV, which can generate almost 400 collisional ionization's within a time of 100 fs [87, 44]. The L-shell photoelectrons of Gd are no greater than 1.2 keV (L III) which may produce 50 collisional ionization's, but the Gd Auger electrons are of high energy. Although the absorption cross section of C is about 144 times lower than Gd, and so the production of photoelectrons per atom is less than for Gd, there are many more C atoms than Gd in the sample. This is the case for all the light

elements of the sample, and in general the overall generation of the electron cascades scales with the X-ray energy deposited on average per atom, which is proportional to the dose. For the HF dose of 1.27 GGy we expect that the collisional ionization rate saturates after about 10 fs, and that by the end of the X-ray pulse, also the light elements can be moderately ionized. At LF instead, collisional effects are negligible, so most of the light element will still be pristine.

The total number of free electrons created increases with time, and is therefore lower with shorter pulses. As described in section 2.3.2, the effect of Bragg termination, where the diffraction signal is gated due to the onset of disorder in the crystal due to random atomic displacement or random ionization, gives rise to a shorter effective pulse duration for the measurement (the later part of the pulse is filtered out of the measurement by selecting just Bragg peaks). I expect that this limits the average "ionization background" experienced at LF and HF and which acts to reduce the contrast of the specific Gd photoionisation. Plasma code simulations predict that at HF the Bragg signal is terminated at about 20 fs [23], which will limit the average ionization of a few percent, while at LF Bragg termination effects are only expected for time much longer than the pulse duration.

7.2.2 Estimation of the average ionization

In order to estimate the relative number of electrons making up the difference between the two datasets at the Gd positions, two separate molecular replacement runs were performed with *REFMAC5* to 2.1 Å, using the previously refined model from which the two Gd ions and the indole group of a tryptophan residue (48 electrons in total) had been removed, together with the structure factors at low and high fluence previously scaled with Scaleit. The indole group was chosen for its stable, aromatic structure, easy to fit into an electron density map. No significant change in the B factors (global and local around the omitted regions) was observed in the two separately refined structures. This finding is important for a quantitative comparison of the electron densities of the omitted parts. $Fo-Fc$ maps were generated with *FFT* [95] around the two missing regions, and these positive difference electron densities were volume-integrated with an *ad hoc* script. The average occupancy of the Gd sites was calculated by merging 6 different X-ray measurements from macrocrystals crystallized in different conditions and exposed to different radiation sources, in order to mimic the possible anisomorphism of the SFX data. The structure from this hybrid set was processed in a similar way as described above and the occupancy refinement was performed with *Phenix Refine* [75], yielding occupancies of 0.82 and

0.76 for the two Gd sites. The ratio between the integrated densities around the Gd and the Trp, multiplied by the number of missing electrons at the Trp location, give an estimate of the effective scattering strength of the two Gd ions. Considering the average occupancy of the two sites, I found that the difference between the two datasets was around $8.8e^-$ per Gd. By repeating the same procedure with other Trp present in the protein, it was possible to get an estimation of the error to associate to the number of electrons, which is around 22%.

Another piece of qualitative evidence for the ionization caused by the FEL radiation comes from the refinement of the atomic form factor (f' and f'' refinement). This was performed with *Phenix Refine*, starting from the DANO values and the phases from the best refined model. 20 cycles of alternated real space and f'/f'' refinement of the two Gd were performed for the LF and HF data. Figure 7.2 displays the resulting scattering strength of the single Gd ion as a function of the refinement cycle, suggesting that the degree of ionization is higher for the HF set, with a difference of about 5 electrons. This value is most probably underestimated, because of the limitation of the refinement process. The refinement, indeed, does not take into account the different conformations of the gadoteridol, which can be clearly seen from the $Fo - Fc$ map, but cannot be clearly distinguished. The partial occupancy of the Gd sites, optimized and left fixed during the refinement process, does not fully model the possible rearrangement of the molecule, and this can reduce the retrieved relative contrast between the two fluences.

7.2.3 Sorting of the datasets

Due to the stochastic nature of the FEL operation, and the uncertain position, size and shape of the focus, the nominal "high fluence" dataset is aggregated from a mixture of different fluences and therefore a mixture of doses. A similar but less dramatic result applies to the low fluence dataset, since the fluence is not high enough to cause a significant change of the scattering factors. In order to optimize the difference signal the single wavelength high-intensity phasing method, the difference between the X-ray fluences must be the highest possible. To achieve this I sorted the indexed diffraction snapshots to select only the patterns with the highest fluence. The narrow size distribution of the lysozyme microcrystals means that the observed diffracted intensity should be proportional to the fluence impinging on the crystal except for the consideration of the beam's spatial profile as discussed above. The number and average integrated intensity of peaks detected in the patterns was used, combined with readings from a pulse intensity monitor located upstream of the

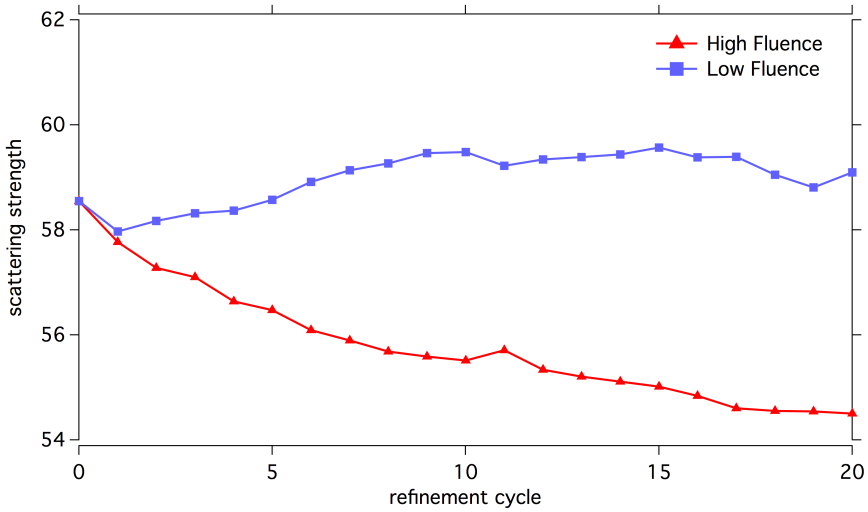

Figure 7.2: The resulting effective scattering strength of the single Gd ion at the end of each refinement cycle, indicating a clear deviation from the expected scattering signal at high fluence, due to ionization effects.

focussing mirror, to find the snapshots corresponding to the highest dose. These values are represented as a scatter plot in figure 7.3, showing a correlation between the number and the average peak intensity to the beam energy. In particular, bright diffraction patterns are mostly found for high intensity X-ray pulses, and often present a large number of Bragg spots. Furthermore, using a criterion of a high number of Bragg peaks also selects the highest-resolution patterns, as shown in the inset. A subset of $121,917$ indexed images was selected from the stream of indexed images at HF, by requiring a pulse intensity higher than 1 mJ (as recorded from the pulse intensity monitor), an average peak intensity (expressed as the sum of all integrated peaks intensity divided by the number of them) greater than 4000 counts and more than 40 found peaks in the pattern. This selection did not compromise the data quality, as shown in table 7.3). The previous analysis was repeated showing a higher ionization degree of the Gds, corresponding to $12e^{-}$, consistent with the difference Fourier map that also showed peaks at slightly higher sigma levels (9.2 σ and 6.3 σ).

7.2.4 Phasing approaches

SAD phasing was performed with *Phenix Autosolve* [75] and was accomplished in a straightforward manner for both X-ray fluences. Interestingly, the LF data had

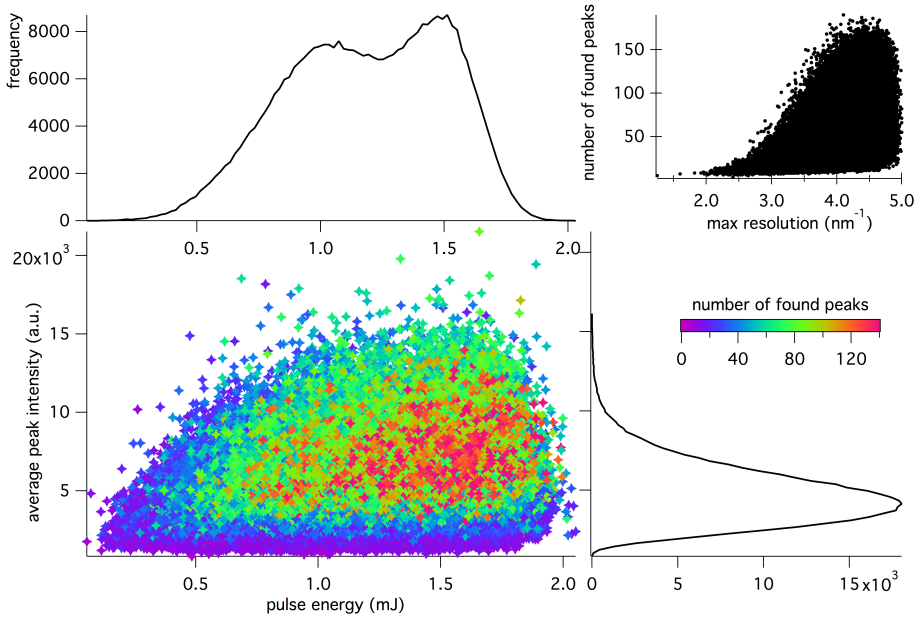

Figure 7.3: Scatter plot of the average intensity of found peaks against the pulse energy, for the high fluence dataset. Each point corresponds to a single indexed diffraction pattern. The colors refer to the number of Bragg peaks found in the pattern. The black curves are the projected histograms of the values of the corresponding axis. Inset: Discrete scatter plot of the number of found peaks versus the highest resolution found.

	Rsplit (%)		
Resolution ($\AA$)	**LF**	**HF**	**HF, best**
6.48	3.04	2.55	4.23
3.43	3.78	2.97	4.98
2.87	4.32	3.25	5.11
2.56	4.46	3.73	5.93
2.35	5.36	4.24	6.53
2.20	6.74	4.85	7.19
2.08	10.51	12.82	19.46
1.98	86.13	112.7	127.2

Table 7.3: Rsplit in resolution bins.

a slightly lower R factor than the high fluence data (0.342 vs. 0.316), even though the former had a lower Rsplit (see table 7.3). To further check the effect of the high pulse intensity on the structure factors, three datasets were created from the low and high fluence data, each containing 73,791 indexed patterns from the best diffracting crystals, defined as shown in table 7.4. The best results from SAD phasing were still obtained with the lowest fluence, while the strongest diffracting patterns at high fluence gave the worst solution. As a comparison, the same datasets where phased with molecular replacement: in this case the results show a much less significant degradation of the quality of the electron density map with the fluence (the resulting metrics are reported in table 7.5). These results can be considered a further indication of the ionization dynamics effect on the anomalous signal of the heavy atoms [49]. The experimental data at different fluences can be considered to a first approximation as a radiation induced phasing (RIP/RIPAS) dataset, using the HF data as the "damaged" set and the LF as the "undamaged". Several phasing attempts were carried out, without success. In the RIPAS approach, the phase information only came from the huge anomalous signal from the heavy atoms, which made SAD phasing straightforward, while any possible isomorphous difference contributed only destructively to the phasing solution, making the final result worse than the SAD approach alone (the final R factor was 0.3349). This might be caused by a wrong

	Original dataset	# of Bragg peaks	Average Bragg counts	pulse energy (mJ)
Low fluence	LF	> 40	> 1000	> 1.0
Medium fluence	HF	> 40	3000 - 5000	> 1.0
High fluence	HF	> 40	> 5000	> 1.0

Table 7.4: Criteria for the selection of the best diffraction patterns. Each set contained $73,791$ patterns.

cross scaling procedure, due to a possible Bragg termination which could have the same effect as non-isomorphism.

7.2.5 Discussion

I have shown the contrast in the effective scattering strengths between low- and high-fluence data. The theoretical model of an isolated Gd atom, as implemented in the XATOM toolkit, predicts an effective ionization between 15 and 25 electrons, whereas the experimental analysis for the Gd derivative of lysozyme shows an average charge state between +8.8 and +12. This discrepancy can be due to multiple reasons, involving both the assumptions made in the theory, unknown experimental parameters, and the inability of standard software to account for the change of the scattering factors due to fast ionization processes. The theoretical model is based on isolated-atom calculations. Charge rearrangement and local plasma formation that might occur in a molecular environment are not included in the model. The electron transfer from neighboring atoms to the highly charged heavy atom will increase the effective scattering strength of the heavy atom, reducing the LF/HF contrast. Similarly, collisional ionization processes are not accounted for, and can potentially reduce the contrast of the heavy atom ionization. To estimate the effect of collisional processes, simulations were performed using a non-local thermal equilibrium plasma code – CRETIN [96]. The simulations were done similar to those described in [97]. This approach has the advantage that it considers the plasma environment, including effects such as continuum lowering and ionization by secondary electrons. Unfortunately, the atomic model of Gd within CRETIN does not include a sufficiently precise description of the atomic levels. For a qualitative analysis of

	low fluence	**medium fluence**	**high fluence**
FOM (Solve)	0.516	0.513	0.544
R factor (solve)	0.3377	0.3433	0.3552
Map-model CC	0.85	0.80	0.61
# residues	126	117	76
Rwork (SAD)	0.2263	0.2658	0.4267
Rfree (SAD)	0.2639	0.3065	0.4567
Rfactor (REFMAC5)	0.2075	0.2247	0.2443
Rfree (REFMAC5)	0.2289	0.2466	0.2643

Table 7.5: Results from SAD phasing (Phenix pipeline) and molecular replacement (REFMAC5) using three subsets of the indexed patterns from the HF and LF sets.

how the plasma environment effects the ionization, I considered a system containing Fe instead of Gd. During the X-ray exposure, secondary ionization will generate a large number of free electrons, which will increase the ionization of all the atoms in the system. This reduces the difference in ionization between the LF and the HF experiments. To visualize the effect of the secondary collision ionizations from the electrons different plasma simulations performed with and without considering the collisional ionizations. Figure 7.4 shows that in the absence of collisions, the average ionization of Fe at low fluence is underestimated by a factor of three, while for the high fluence case, the ionization is saturated even when the secondary effects are disregarded. We assume that this discussion holds for Gd as well, which would to some extent explain why we observed a smaller difference in ionization in the HF and the LF case compared to what was estimated from the atomic model. Another effect that could furthermore reduce the ionization contrast is the turning of the Bragg signal due to the loss of coherent scattering [46]. In the HF case we expect that only the first 20 fs of the pulse will actually contribute to the Bragg signal, so relaxation effects taking place after that time range do not contribute to the total ionization.

The effect of ionization-induced fluctuations is disregarded in standard crystallographic software. During the intense X-ray pulse, the form factors of heavy atoms are stochastically and dramatically changed through strong ionization, assuming the form similar to equation 7.1. Using this equation and assuming that only heavy atoms scatter anomalously and undergo ionization dynamics independently, the scattering intensity can be written as:

with the introduction of a time-averaged form factor $\bar{f} = \int dtg(t)\tilde{f}(t)$. F_P^0 is the molecular form factor for the protein without Gd atoms and N_{Gd} is the number of Gd atoms in a crystal. The dependences on $\mathbf{Q}$, $\mathcal{F}$, and ω are omitted for simplicity. In conventional X-ray crystallography, only the first part of the equation is used to fit to the scattering intensity measurement, while the fluctuations defined by:

$$V_1 = \int dtg(t) \left[\sum_q P_q(t)|f_q|^2 - |\sum_q P_q(t)f_q|^2 \right] ,$$

$$V_2 = \int dtg(t)|\tilde{f}(t)|^2 - |\int dtg(t)|\tilde{f}(t)|^2$$

are not taken into account. If we assume the same beam properties as before, the calculated standard deviation given by $\sqrt{V_2}$ for a Gd atom is $5.9e^-$ for the LF case and $10.6e^-$ for the HF case. As a result, the effective scattering strength would be overestimated by the standard crystallographic software because it neglects a large

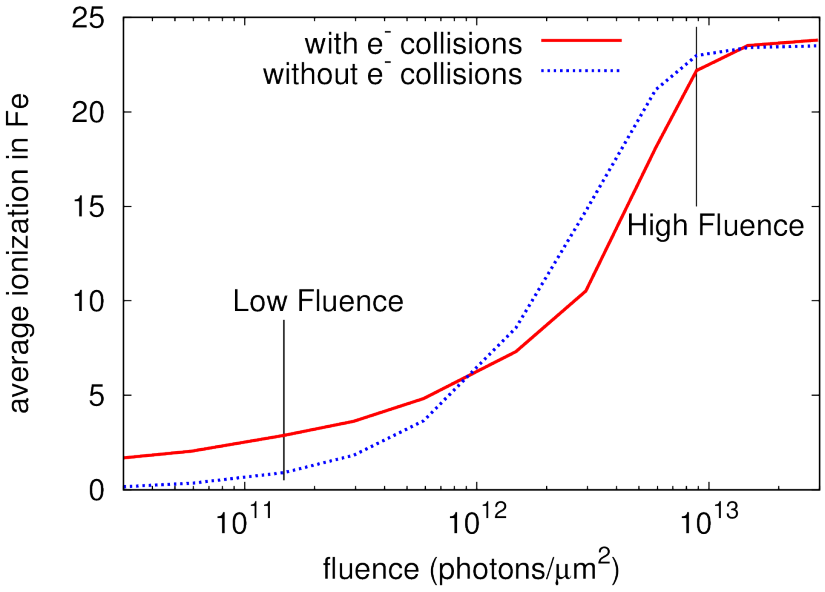

Figure 7.4: Average ionization of Fe atoms contained in a protein system, at the end of a 40 fs X-ray pulse, as a function of the X-ray fluence, simulated using a plasma physics code. The red line represents the results of simulations in which collisional ionization effects have been modeled. Courtesy of Nicusor Timneanu.

contribution from V_2.

Experimentally, to analyze the scattering signal and electronic damage at high X-ray intensity, it is important to know the X-ray fluence hitting individual atoms. If some X-ray beam parameters are unknown, a proper volume averaging cannot be performed. The calculations of the effective scattering strength show a strong dependence on the interaction volume geometry, suggesting the need for a calibration of the X-ray beam profile. Another experimental issue is the position dependence of the X-ray fluence across the microcrystal, as already discussed in the previous chapters.

Another problem in treating our data with standard crystallographic software lies in the scaling procedure of SFX data exposed to very high X-ray fluence. This is because the ionization mechanisms of the light atoms, which result in an overall decrease in scattering strength of the molecule, may not be fully corrected for. Similarly, Bragg termination effects may introduce changes in the scattering factors and their resolution dependence which are not compensated by the Wilson type scaling procedures.

7.3 Tailoring the crystal size to compensate for an imperfect FEL beam

The shot-by-shot size and shape of the X-ray FEL beam at the interaction region is one one of the main unknown that can prevent the successfulness of a HIP experiment. In particular, a non-uniform photon density impinging on a crystal will cause a mix of diffracted intensities to be summed on the single image. As a result, the integrated Bragg intensity will contain a mix of fluences which are intrinsically impossible to disentangle. Knowing the real beam intensity profile, however, can be of great help, since the crystal size could be tailored to minimize the contribution of the low fluence region of the beam, while the recorded data could be sorted *a posteriori* to get rid of low fluence hits, similarly as described in this chapter.

Here I describe two extreme example of crystal shapes, which corresponds to the CatB and the granulovirus (assumed for simplicity as spherical crystals). The X-ray beam is assumed as an inner circular region at high photon density and a concentric annulus at lower photon density. The central beam has a diameter of $0.2\,\mu$m, while the low flux tail extends to $1\,\mu$m, and the two areas contain the same number of photons. This beam travels along the $\hat{y}$ direction of a cartesian system, and, for the CatB case, it interacts with a crystal modeled as a cylinder of 1 μm

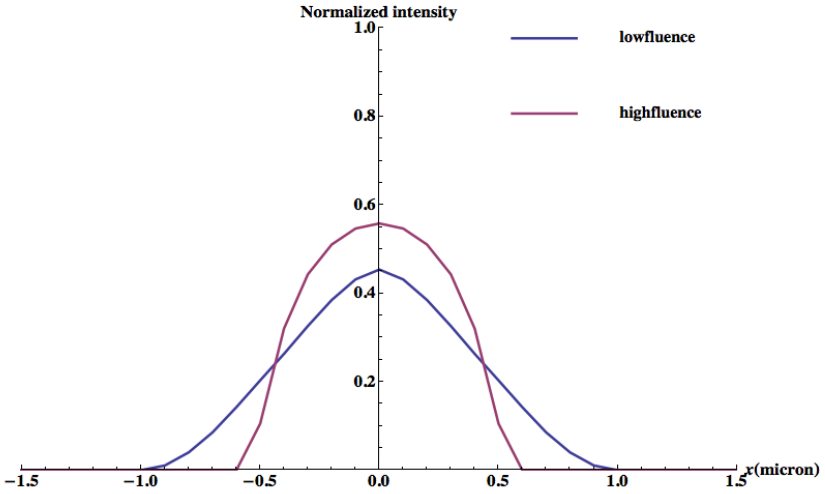

Figure 7.5: Interaction between an imperfect X-ray beam and a cylindrical crystal (CatB case).

of diameter and several microns in length. Assuming that the center of the CatB crystal is at the origin of the reference system and that its long axis lays along the $\hat{z}$ direction (exactly perpendicular to the FEL beam), the scattered signal obtained by scanning the X-ray beam along $\hat{x}$ varies as in figure 7.5. Here, the scattered signal is divided into the two regions of the beam (central at high fluence and external at low fluence). Even at $x = 0$, i.e. at the perfect center of the crystal, the high fluence region contributes only 55% to the scattering signal.

The second case is represented by the granulovirus particles, represented as spherical crystal of 0.5 μm of diameter. Figure 7.6 shows the results obtained, using the same reference systems, and positioning the crystal at the center of the cartesian axes. A clear improvement of the high intensity signal level is evident, which now reaches 87% at $x = 0$. This improvement comes from the fact that the crystal size is compatible to the width of the "hot" X-ray beam region.

In this last case considered, the maximum high fluence contribution also coincides with the highest Bragg signal on the detector, so a similar procedure as the one used in chapter 7 to select the best patterns can be utilized. As a contrary effect, only a very small percentage of shots will intersect the crystal close to its center. Moreover, smaller crystals usually requires a more concentrated solution to obtain the same hit rate.

The Gd-lysozyme crystals described in this chapter falls into the first case, since their average dimensions are bigger than the nominal high fluence profile. Nev-

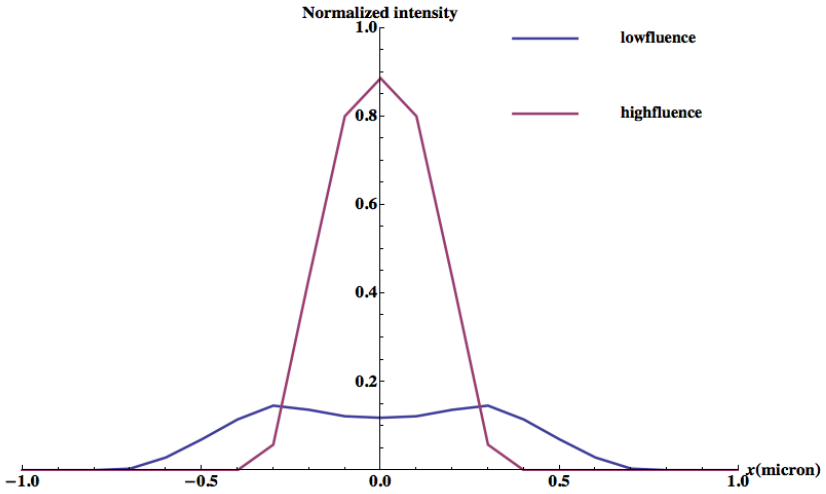

Figure 7.6: Interaction between an imperfect X-ray beam and a spherical crystal (granulovirus case).

ertheless, lysozyme can be still considered one of the best sample to use for future experiments, aimed to develop high intensity phasing at FELs, due to the large availability of the protein and the easiness to grow tailor-sized crystals, with a relatively low size distribution.

Chapter 8

Conclusions and outlook

8.1 Conclusions

In the thesis I showed that the pulse brightness of XFELs allows us to collect diffraction patterns from much smaller crystals than been examined at synchrotron radiation facilities. The best example of this is the GV structure, presented in chapter 4, which demonstrated the possibility to solve protein structures from crystals which consisting of only few thousands of unit cells. The experimental analysis showed that the great majority of diffraction patterns collected does not use the highest dose delivered by the beam, but only a portion of it, generated from the tail of the X-ray beam. Furthermore, the collected data were limited in resolution by the experimental geometry. These facts indicate the possibility of collecting patterns from much smaller crystals, only a few hundreds time larger than the single molecule. The retrieved 2 Å structure does not show effects due to radiation damage, or clear Bragg termination effects, meaning that even higher fluence XFEL beams will be beneficial for crystallography.

While the ionic motion initiated by the Coulomb scattering is relevant only after tens of femtoseconds, the electronic damage that initiate it can result in a loss of scattering power during the diffraction. This ultrafast damage process is inevitable at high fluences, since it is produced almost instantaneously and increases within the first femtoseconds due to relaxation and collisional processes. In chapter 7 I demonstrated that this damage effect frustrates standard experimental phasing, which rely on the out of phase contrast, bleached at high intensity. Nevertheless, I showed that a novel phasing approach could be possible, which exploits the loss of scattering strength of heavy atom species. In chapter 5 I demonstrated that a standard RIP approach can in theory be used to determine the ionized substructures as well as

determine high quality phases, by simulating two serial femtosecond crystallography experiments at different X-ray fluences, under experimental conditions that mimic those available at the LCLS.

The analysis of experimental data reported in chapter 6 and 7 showed that this approach suffers by the lack of information about the exact pulse fluence hitting the crystal on a single shot, and the amount of crystalline material exposed to the radiation. The large number of data recorded in the Gd-lysozyme experiment allowed to select smaller sets of data according to the average scattering intensity, which has shown to increase the ionization contrast between high and low fluence. In the case of an imperfect beam profile, presenting low fluence tails, the integrated Bragg intensity will contain a mix of fluences which are intrinsically impossible to disentangle in the case of crystals with not uniform size. These effects can be avoided growing crystals of size comparable with or smaller than the expected FWHM of the high fluence region, while a more precise sorting of the collected diffraction patterns could be possible during the data analysis step.

Extra diagnostic tools, such as a beam intensity monitor after the interaction region, a time of flight spectrometer, or a simultaneous measurement of the fluorescence, can improve the sorting of the recorded pattern as a function of strength of the interaction between the FEL pulse and the crystal, as well as helping during the alignment of the liquid jet to the X-ray beam.

Plasma code simulations have shown that the collisional processes initiated by the photoionization events can provoke moderate ionization of light atoms, acting as a global damage effects which reduces the high fluence contrast on the heavy atoms. This damage mechanism develops and saturates within the first 10 fs, and it can be partially overcome by using shorter pulses. To the contrary, other relaxation effects such as Auger decay and fluorescence have a beneficial bleaching effect and have lifetimes that depend on the atomic species, so the optimal pulse length will change for different heavy atoms, but a rule of thumb should not exceed 10 fs.

8.2 Outlook

8.2.1 Experimental determination of the atomic form factors at high X-ray intensity

So far, the generalized version of anomalous diraction at high X-ray intensity has only been expressed mathematically, but it has never been tested experimentally. The determination of the anomalous coefficients is often a key ingredient in stan-

dard anomalous phasing methods and it is even more important in the FEL case, where those coefficients experience a dramatic, fluence-dependent change. A direct measurement could, in addition, allow a quantitative comparison to the theory. By performing a scattering experiment in combination with transmission and fluorescence measurements, utilizing simple atomic and molecular crystalline systems, in form of solid crystalline targets, it should be possible to retrieve those high intensity coefficients from the equations describe below. Systems such as Fe or Cu films, in particular, can be well simulated by plasma or atomic physics codes and are often found natively or artificially bound to biologically interesting proteins.

As described in details in section 2.4, in the specific case of a Fe solid target with a thickness of 200 nm, exposed to pulse of a $5 \cdot 10$ photons/μm^2, the expected variation of the transmission is around 6%. Experimentally, the transmission can be measured in a single shot with an accuracy of about 2%, and one could achieve much better than 0.1% error by averaging thousands of shots and by binning shots by the incident pulse energy. The high intensity anomalous coefficient c can be also determined via fluorescence measurements, using the equations shown in section 2.4. Fluorescence signal can be determined using a detector covered with a solid filter made of two consecutive materials, in order to subtract the elastic scattering or other unwanted signals. For example, a thin polyimid filter with two separated regions covered with a layer of V and Co can be used to extract the fluorescence signal of Fe, as sketched in figure 8.1. By integrating the signal from a large detector area, then, it could be possible to achieve enough single-shot accuracy to sort the diffraction patterns according to the calculated fluorescence, which is proportional to the volume of the crystal exposed to the radiation and to the intensity of the X-ray beam.

The other high intensity coefficients can be extracted from the scattering intensity of the crystalline compounds, which can be recorded in parallel to the transmission and the fluorescence measurements.

8.2.2 Exploit UV radiation induced damage to understand the mechanism of disulphide bond breakage

An interesting application of time-resolved protein nanocrystallography is related to the study of radiation-induced damage, such as the one provoked by an UV radiation (described in Section 1.3 and studied for example in [98]). The mechanisms of disulphide bond breaking, in particular, are now studied in association to cancer-related proteins [99] and new therapies are being considered which target these

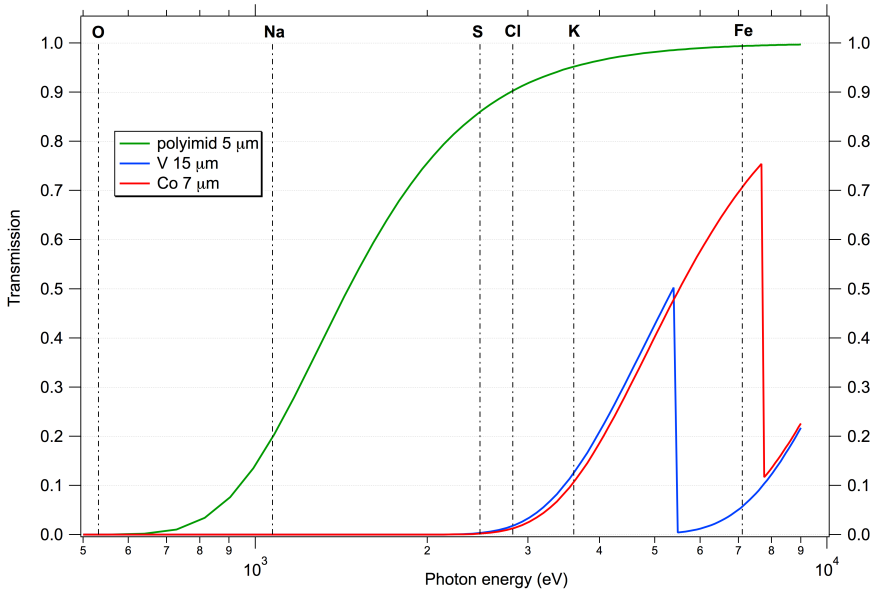

Figure 8.1: Transmission of thin layers of polyimid, V, and Co, as a function of the X-ray energy. The fluorescence energy of few selected elements is shown with dashed lines.

particular bonds, called allosteric. While inside the cells the cleavage of allosteric disulphide bonds is controlled chemically, it is known that UV light can induce the reduction of disulphide bridges. This reduction can be accomplished through direct photoionization or through indirect electronic transfer processes (such as electron generations from aromatic protein residues), on a sub-microsecond timescale, as shown in figure 8.2. The combination of a UV pumping laser and the SFX technique may allow the determination of transient states between the UV excitation and the disulphide breakage, allowing to elucidate the mechanisms inside these allosteric proteins.

8.3 Future perspectives

By increasing the pulse irradiance of an FEL source (either by increasing the pulse energy or by tighter focusing), it will be eventually possible to collect data from nanocrystals made of just a few tens of unit cells, or even single molecules. Under these extreme fluence regimes, the atoms in the sample will experience severe electronic damage. Pulse lengths shorter than the typical lifetimes of relaxation effects

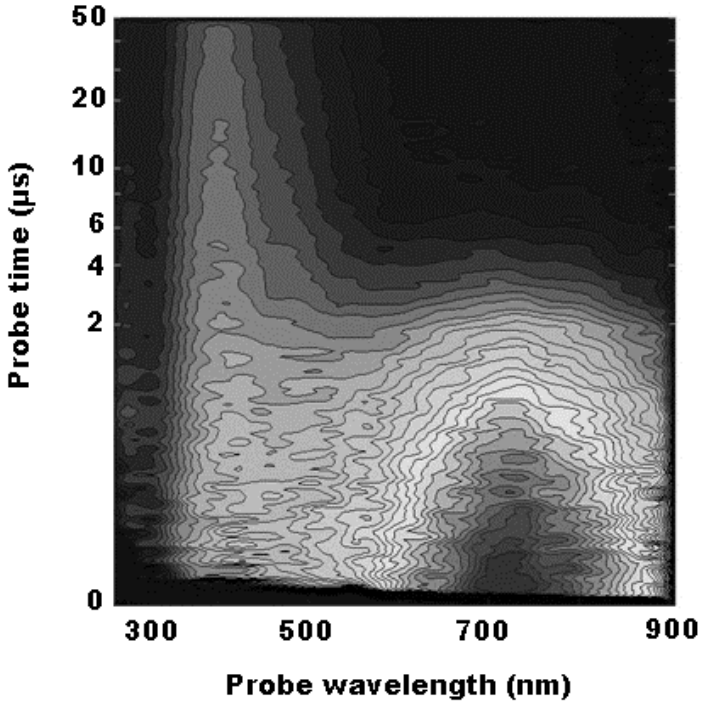

Figure 8.2: Disulphide bond reduction induced with UV light, observed with time-resolved absorption spectroscopy. Black regions correspond to zero absorption, white to a strong absorption. © 2012 M. T. Neves-Petersen, S. Petersen, G. P. Gajula. Originally published in [98] under CC BY 3.0 license. Available from: http://dx.doi.org/10.5772/37947.

or other secondary process can partially mitigate radiation damage. Nevertheless, photoionization is an instantaneous process (at least, below sub-femtoseconds), and there is still much to be done in order to achieve sub-femtosecond pulses with pulse compression techniques, without reducing the number of electrons in the bunch (and consequently the intensity of the radiation). For these reasons, the recorded diffraction patterns will always be influence by electronic damage. In order to optimize the data collection strategy and to create more accurate models able to explain the electron dynamics at high X-ray intensity, it is necessary to explore the effects of high fluence, by means of simple experiments, such as those explained in the previous section. Current SFX experiments can as well take great advantage of the extra diagnostic tools and of accurate theoretical models, in particular during the sorting of the recorded pattern as a function of strength of the interaction between the FEL pulse and the crystal, which can improve the statistics and the quality of the final dataset.

Chapter 9

Appendix:

9.1 Lorentz space-time and frequency-wavenumber transformations

Lorentz space-time transformations provide relationships between spatial and temporal scales in two frame of reference S and S' when the relative speed between the two approaches that of light.

Assuming (x, y, z, t) the coordinates for the frame S and (x', y', z', t') for S', which moves at velocity v with respect of S in the z, z' direction, then:

$$\begin{aligned} z &= \gamma(z' + \beta ct') \\ t &= \gamma(t' + \tfrac{\beta z'}{c}) \\ y &= y' \\ x &= x' \end{aligned}$$

are the Lorentz space-time transformations between the two sets of coordinates, with $\beta = v/c$ and $\gamma = 1/\sqrt{1-\beta^2}$. The Lorentz transformations reduce to the Galilean transformations as $v/c \to 0$ (indeed $\beta \to 0$ and $\gamma \to 1$, so $t = t'$ and $z = z' + vt'$). From those relations, it is possible to calculate the relationships between frequency ω and wavelength λ in the two coordinate systems. The phase factor ϕ of a propagating wave in the frames S and S' can be expressed, respectively, as:

$$\begin{aligned} \phi &= \omega t - k_z z - k_x x - k_y y \\ \phi' &= \omega' t' - k'_z z' - k'_x x' - k'_y y' \ . \end{aligned}$$

Utilizing the equality $\phi = \phi'$ and the Lorentz transformations, the relationships between frequency and wavelength can be obtained:

$$\omega = \gamma(\omega' + \beta c k'_z)$$
$$k_z = \gamma(k'_z + \tfrac{\beta\omega'}{c})$$
$$k_x = k'_x$$
$$k_y = k'_y \, .$$

It is also convenient to express the frequency shift (also called Doppler shift) in terms of the angle θ between $\mathbf{k}$ and $\mathbf{k}_z$ since $|\mathbf{k}| = \omega/c$, then $kz = \omega/c\cos(\theta)$, so:

$$\omega = \omega'\gamma(1 + \beta\cos(\theta')) \, .$$

Similarly, the angular transformations between the two frames of reference can be written as:

$$\cos(\theta) = \tfrac{\cos(\theta')+\beta}{1+\beta\cos(\theta')}$$
$$\sin(\theta) = \tfrac{\sin(\theta')}{\gamma(1+\beta\cos(\theta'))} \, .$$

By combining the previous two equations, the formula for the tangent of the angle can be expressed:

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

this formula is convenient for illustrating the folding of the radiation cone characteristic of synchrotron radiation (also called "searchlight effect"): for relativistic electrons, for which $\gamma \gg 1$, even if in their frame of reference S' the radiation cone is broad (say $0 < \theta' < \pi/4$), in the laboratory frame the radiation will be delimited in a cone of half angle of order $1/2\gamma$.

9.2 Semi-classical model for bound electrons

In the semi-classical model an atom is represented by a massive positively charged nucleus, surrounded by several electrons held at discrete binding energies, and an impinging electromagnetic wave is described by an electric field E_i of frequency ω. Each bound electron is forced by the incident field to a simple harmonic motion, while the positively charge nucleus acts as a restoring force. The electron response depends on the closeness of the incident wave frequency to the resonant frequency ω_s, that is, on $\omega - \omega_s$, and the equation of motion for each of the bound electrons can be written as:

$$\frac{md^2\mathbf{x}}{dt^2} + m\gamma\frac{d\mathbf{x}}{dt} + m\omega_s^2\mathbf{x} = -e(\mathbf{E_i} + \mathbf{v} \times B_i) \,, \tag{9.1}$$

with m the electron mass, γ the dissipative frequency (assumed $\gamma \ll \omega$), and e the elementary charge. The Lorentz force $-e(\mathbf{E_i} + \mathbf{v} \times B_i)$ can be neglected for non-relativistic oscillation velocities v, and for a driving electric field of the form $\mathbf{E} = \mathbf{E}_i exp(-i\omega t)$ the displacement $\mathbf{x}$ will have the same time dependence, so the equation 9.1 becomes:

$$m(-i\omega)^2\mathbf{x} + m\gamma(-i\omega)\mathbf{x} + m\omega_s^2\mathbf{x} = -e\mathbf{E_i} \,.$$

The harmonic displacement is then given by:

$$\mathbf{x} = \frac{1}{\omega^2 - \omega_s^2 + i\gamma\omega}\frac{e\mathbf{E}_i}{m} \,.$$

From this solution one can derive the semi-classical scattering cross section for a bound electron (see Chapter 2 of reference [36] for detailed definitions and examples):

$$\sigma = \frac{8\pi}{3}r_e^2\frac{\omega^4}{(\omega^2 - \omega_s^2)^2 + (\gamma\omega)^2} \,,$$

where r_e is the classical electron radius. The cross section shows a strong resonance at $\omega \simeq \omega_s$, while for very large frequencies σ approaches the cross section of a free electron (Thomson's result), so the bound electrons scatter as if they were free. Well below the resonant frequency, instead, the cross section has a very strong λ^{-4} dependence, as in the Rayleight formula.

9.3 Construction of the Patterson map from Fourier synthesis

The Patterson map can be constructed from a Fourier synthesis using the experimental intensities, $|\mathbf{F_h}|^2$, as Fourier coefficients. The convolution of two function is defined by a convolution integral of the form:

$$Conv(\mathbf{u}) = f(\mathbf{r}) \otimes g(\mathbf{r}) = \int_R f(\mathbf{r})g(\mathbf{r} + \mathbf{u})d\mathbf{r} \,.$$

If one sets $f(\mathbf{r}) = \rho(\mathbf{r})$ and $g(\mathbf{r}) = \rho(-\mathbf{r})$, the convolution is equivalent to the Patterson function, defined in equation 1.8. From the Fourier convolution theorem:

$$\Im[f(\mathbf{r}) \otimes g(\mathbf{r})] = \Im[f(\mathbf{r})] \cdot \Im[g(\mathbf{r})]$$

(where the Fourier transform operation is denoted by $\Im$) and by recalling that the Fourier transform of the electron density is the complex structure factor $\mathbf{F_h}$, follows that the Fourier transform of the Patterson function is:

$$\Im[P(\mathbf{u})] = \Im[\rho(\mathbf{r}) \otimes \rho(-\mathbf{r})] = \Im[\rho(\mathbf{r})] \cdot \Im[\rho(-\mathbf{r})] = \mathbf{F_h} \cdot \mathbf{F_{-h}} = |\mathbf{F_h}|^2.$$

Inverting the Fourier transform, one can obtain:

$$P(\mathbf{u}) = \Im^{-1}[|\mathbf{F_h}|^2] \, ,$$

which, for a discrete summation over the observed Bragg reflections of the form $\mathbf{h} = (h, k, l)$, can be rewritten as:

$$P(u, v, w) \sim \sum_h \sum_k \sum_l |\mathbf{F_h}|^2 \cos 2\pi(hu + kv + lw)] \, .$$

9.4 Iterative substructure determination

Experimental phasing of macromolecules often requires the presence of marker atoms, such as metals or sulfurs (called, in general, "heavy atoms"), whose determination is one of the required steps for structure solution programs. In order to determine the heavy atom substructure, Patterson seeding methods [100] are often used, which consider the strongest peaks of the interatomic distance vectors map (the Patterson function) as potential candidates of two- (heavy) atom fragment. A large number of random positions in the unit cell are tested for one of these peaks, and the position of the two atoms is used to generate a full-symmetry Patterson superposition function. The solution that minimizes this function is then adopted as the seed used to obtain further heavy atom locations. The resolution to which the data are automatically truncated for this substructure determination step is given by the resolution to which significant anomalous (or isomorphous) differences are observed.

The correlation coefficient (CC) between the observed and the calculated normalized structure factors ($CC(all)$), and the CC based on the reflections not used in the substructure determination ($CC(weak)$) are often used as quality metrics of the results.

9.5 Molecular replacement

Molecular replacement is the process of solving the phase problem for an unknown structure, in the case which a similar structure is known. By placing the atomic model for the known structure in the unit cell of the unknown structure in such a way as to best reproduce the observed structure factors. Once placed, phases can be calculated and, in combination with the observed structure factors, used to start the process of interpreting the electron density map. The initial location process is divided into a three dimensional rotational and translational search, which uses a correlation function between the observed and calculated Patterson maps. In this thesis, this search has been performed automatically using *Molrep* [101]. Phasing and automated refinements are instead estimated using maximum likelihood methods implemented in *Phaser* [74] or in *REFMAC5* [89]. A general rule for this phasing method to work is that the sequence identity of the homologous model to the unknown structure has to be greater than 30%.

9.6 Primary functions of Cheetah

As explained in section 3.3.1, *Cheetah* is a multi-purpose software created to address the issue of pre-processing big data sets in serial X-ray diffraction, in a fast and efficient way. The primary functions of the software are here described in details:

Detector corrections

Detector artifacts, such as saturated pixels, can be identified and flagged by applying a simple intensity threshold. X-ray-free dark frames (typically collected at the beginning and during an experiment) can be used to determine the detector offsets and can be subtracted. Similarly, common mode offsets for the single detector panel can be estimated and corrected. Each pixel can be also scaled according to the gain calibration map, if provided. Finally, it is possible to apply algorithms to determine and to mask out bad pixels, with the possibility of creating a "bad pixel mask" in the form of a binary image. All these functions can be individually turned on or off manually, and are normally used as a standard part of the data analysis.

Subtraction of the photon background

Serial X-ray diffraction experiments have constantly changing background signals, due to source fluctuation or to differences in the sample. The background subtraction in *Cheetah* can be performed in different ways, depending on the shot-to-shot variation of its signal. When the photon background is relatively constant, a running background subtraction can be used, which estimate the background signal from the blank frames in between the hits. In this case, a pixel-wise median is periodically calculated from the saved non-hit frames, and this median filter is subtracted from the diffraction patterns. For experiments performed using a liquid jet, however, the scattering from the solution can vary significantly from shot to shot. In this case, the running background method should be avoided or used with great precaution. In the case of a crystalline sample, it is convenient to use a local background subtraction across the image. For each pixel, the background is estimated as the median of all the pixel values contained in a square of side $2n + 1$, centered at the pixel location. If the size of the box is sufficiently larger than the average dimension of a Bragg peak, the blind median is a good estimation of the local background, which vary on a relatively long pixel scale. As a rule of thumb, the area inside which the median is calculated has to be three times larger than the area of any Bragg peaks.

Image analysis

Diffraction images containing possible Bragg peaks are selected on the basis of the minimum and number of peaks identified in the pattern. Peaks in the intensity are recognized as connected clusters of pixels above a given threshold value; these clusters have to contain no more than a n_{min} pixels, and fewer than n_{max}, in order to reject single-pixel outliers or too diffuse peaks. The user can also specify more sophisticated algorithms for peak searching, adopting, for example, the intensity ratio between the intensity of the peak and the local background, determined from the intensity in the region around the peak.

At the end of the peak search procedure, the images labeled as "hits" are saved, both as cleaned (i.e. background-corrected) and raw images, together with a list containing the coordinates of the centroids of the found peaks and their total intensities.

9.7 Monte Carlo integration of intensities

Once each diffraction pattern have been indexed and the peak intensities in the predicted locations have been determined, the final intensity have to be calculated for each of the symmetrically unique reflection. As the crystals can have different sizes, shapes, orientations and qualities, a Monte Carlo method of integration of the intensity over these quantities close to the diffraction condition can be used. For m different crystallites, the integrated experimental intensity of a particular Bragg reflection hkl of wave vector $\overline{\Delta k}$ can be calculated as:

$$I_{hkl}(m) = \sum_{n=1}^{m} \sum_{\{j\}_{m,hkl}} I'_n(\overline{\Delta k_j}) \ ,$$

where $\{j\}_{m,hkl}$ is the set of pixels in the patterns composing the integration region of the reflection hkl and $I'_n(\overline{\Delta k_j})$ is the diffracted intensity of the reflection in the nth pattern, after background and polarization factor corrections:

$$I'_n(\overline{\Delta k_j}) = \frac{I_n(\overline{\Delta k_j}) - I_{bg}(\overline{\Delta k_j})}{P(k_{0,j})\Delta\Omega_j} \ .$$

Here $P(k_{0,j})$ describes the polarization factor for an incoming radiation of wave vector k_0 and $\Delta\Omega$ is the solid angle subtended by a detector pixel. The final value for the experimental intensity assigned to the reflection hkl is then the average of the integrated intensity over all the diffracted intensities from equivalent reflections.

The convergence of the Monte Carlo approach with respect to the integration domain size has been investigated by Kirian *et al.* [63], showing the existence of an optimal integration volume which depends on the crystallite sizes. The integration and merging process implemented in CrystFEL, however, are only minimally sensitive to the size of the integration domain, since the integrated intensity is taken as the average of the total peak intensity [102]. The effects on the convergence due to reflection partialities have been evaluated by White *et al.* [102].

Bibliography

[1] Feynman, R. P.; Leighton, R. B.; Sands, M. *The Feynman lectures in physics, Mainly Electromagnetis and Matter*; Addison and Wesley, 1963; Vol. 2.

[2] Giacovazzo, C.; Monaco, H.; Viterbo, D.; Scordari, F.; Gilli, G.; Zanotti, G.; Catti, M. *Fundamentals of Crystallography*; Oxford University Press, 1993.

[3] Ashcroft, N. W.; Mermin, N. D. *SolidStatePhysics*; Thomson Learning, 1978.

[4] Friedrich, W.; Knipping, P.; Laue, M. *Annalen der Physik* **1913**, *346*, 971–988.

[5] Bragg, W. L. In *Proceedings of the Cambridge Philosophical Society*; p 4.

[6] McPherson, A.; Kuznetsov, Y. G.; Malkin, A.; Plomp, M. *Journal of structural biology* **2003**, *142*, 32–46.

[7] Rossmann, M. G.; van Beek, C. G. *Acta Crystallographica Section D: Biological Crystallography* **1999**, *55*, 1631–1640.

[8] White, T. A. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2014**, *369*, 20130330.

[9] Warren, B. E. *X-ray Diffraction*; Courier Dover Publications, 1969.

[10] Rupp, B. *Biomolecular crystallography: principles, practice, and application to structural biology*; Garland Publishing: New York, NY, USA, 2010.

[11] Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. *Nature* **1958**, *181*, 662–6.

[12] Dickerson, R.; Kendrew, J. t.; Strandberg, B. *Acta Crystallographica* **1961**, *14*, 1188–1195.

[13] Cullis, A. F.; Muirhead, H.; Perutz, M.; Rossmann, M.; North, A. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1962**, *265*, 161–187.

[14] Blake, C. C. F.; Fenn, R. H.; North, A. C.; Phillips, D. C.; Poljak, R. J. *Nature* **1962**, *196*, 1173–6.

[15] Coster, D.; Knol, K.; Prins, J. *Zeitschrift für Physik* **1930**, *63*, 345–369.

[16] Bijvoet, J.; Peerdeman, A.; Van Bommel, A. *Nature* **1951**, *168*, 271–272.

[17] Wang, B.-C. *Methods in enzymology* **1985**, *115*, 90.

[18] Hendrickson, W. A.; Teeter, M. M. *Nature* **1981**, *290*, 107–113.

[19] Blake, C. C. F.; Phillips, D. C. *International Atomic Energy Agency Symposium* **1962**, 183–191.

[20] Henderson, R. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **1990**, *241*, 6–8.

[21] Henke, B. L.; Gullikson, E. M.; Davis, J. C. *Atomic data and nuclear data tables* **1993**, *54*, 181–342.

[22] Paithankar, K. S.; Owen, R. L.; Garman, E. F. *J Synchrotron Radiat* **2009**, *16*, 152–62.

[23] Chapman, H. N.; Caleman, C.; Timneanu, N. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2014**, *369*, 20130313.

[24] Caleman, C.; Timneanu, N.; Martin, A.; Aquila, A.; Barty, A.; Schott, H.; White, T.; Chapman, H. *In preparation.*

[25] Garman, E. F. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 339–351.

[26] von Sonntag, C. *The chemical basis of radiation biology*; Taylor & Francis London, 1987.

[27] Burmeister, W. P. *Acta Crystallographica Section D: Biological Crystallography* **2000**, *56*, 328–341.

[28] Weik, M.; Bergès, J.; Raves, M. L.; Gros, P.; McSweeney, S.; Silman, I.; Sussman, J. L.; Houée-Levin, C.; Ravelli, R. B. G. *J Synchrotron Radiat* **2002**, *9*, 342–6.

[29] Ramagopal, U. A.; Dauter, Z.; Thirumuruhan, R.; Fedorov, E.; Almo, S. C. *Acta Crystallographica Section D: Biological Crystallography* **2005**, *61*, 1289–1298.

[30] Weik, M.; Ravelli, R. B.; Kryger, G.; McSweeney, S.; Raves, M. L.; Harel, M.; Gros, P.; Silman, I.; Kroon, J.; Sussman, J. L. *Proceedings of the National Academy of Sciences* **2000**, *97*, 623–628.

[31] Ravelli, R. B.; Leiros, H.-K. S.; Pan, B.; Caffrey, M.; McSweeney, S. *Structure* **2003**, *11*, 217–224.

[32] Sheldrick, G. M. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 479–485.

[33] Howell, P.; Smith, G. *Journal of applied crystallography* **1992**, *25*, 81–86.

[34] Kabsch, W. *Journal of Applied Crystallography* **1988**, *21*, 67–72.

[35] Nanao, M. H.; Sheldrick, G. M.; Ravelli, R. B. *Acta Crystallographica Section D: Biological Crystallography* **2005**, *61*, 1227–1237.

[36] Attwood, D. *Soft x-rays and extreme ultraviolet radiation: principles and applications*; Cambridge university press, 1999.

[37] Saldin, E.; Schneidmiller, E.; Yurkov, M. *Proceedings of FEL 2006, BESSY, Berlin, Germany* **2006**.

[38] Saldin, E.; Schneidmiller, E.; Yurkov, M. *Optics Communications* **2008**, *281*, 4727–4734.

[39] Bonifacio, R.; de Salvo Souza, L.; Pierini, P.; Scharlemann, E. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1990**, *296*, 787–790.

[40] Yu, L. H. *Physical Review A* **1991**, *44*, 5178.

[41] *https://www.elettra.trieste.it/lightsources/fermi/beam-parameters-copy.html*; 2013.

[42] Young, L.; Kanter, E.; Krässig, B.; Li, Y.; March, A.; Pratt, S.; Santra, R.; Southworth, S.; Rohringer, N.; DiMauro, L.; et al. *Nature* **2010**, *466*, 56–61.

[43] Neutze, R.; Wouts, R.; van der Spoel, D.; Weckert, E.; Hajdu, J. *Nature* **2000**, *406*, 752–757.

[44] Caleman, C.; Bergh, M.; Scott, H. A.; Spence, J. C.; Chapman, H. N.; Tîmneanu, N. *Journal of Modern Optics* **2011**, *58*, 1486–1497.

[45] Hau-Riege, S. P.; London, R. A.; Szoke, A. *Physical Review E* **2004**, *69*, 051906.

[46] Barty, A.; et al. *Nat Photonics* **2012**, *6*, 35–40.

[47] Son, S.-K.; Young, L.; Santra, R. *Physical Review A* **2011**, *83*, 033402.

[48] Son, S.-K.; Chapman, H. N.; Santra, R. *Phys Rev Lett* **2011**, *107*, 218102.

[49] Son, S.-K.; Chapman, H. N.; Santra, R. *Journal of Physics B: Atomic, Molecular and Optical Physics* **2013**, *46*, 164015.

[50] Als-Nielsen, J.; McMorrow, D. *Elements of modern X-ray physics*; John Wiley & Sons, 2011.

[51] Chapman, H. N.; et al. *Nature* **2011**, *470*, 73–7.

[52] Boutet, S.; et al. *Science* **2012**, *337*, 362–4.

[53] Redecke, L.; et al. *Science* **2013**, *339*, 227–30.

[54] Barends, T. R.; Foucar, L.; Botha, S.; Doak, R. B.; Shoeman, R. L.; Nass, K.; Koglin, J. E.; Williams, G. J.; Boutet, S.; Messerschmidt, M.; et al. *Nature* **2014**, *505*, 244–247.

[55] DePonte, D.; Weierstall, U.; Schmidt, K.; Warner, J.; Starodub, D.; Spence, J.; Doak, R. *Journal of Physics D: Applied Physics* **2008**, *41*, 195505.

[56] Weierstall, U.; et al. *Nat Commun* **2014**, *5*, 3309.

[57] Philipp, H. T.; Koerner, L. J.; Hromalik, M. S.; Tate, M. W.; Gruner, S. M. *Nuclear Science, IEEE Transactions on* **2010**, *57*, 3795–3799.

[58] Barty, A.; Kirian, R. A.; Maia, F. R.; Hantke, M.; Yoon, C. H.; White, T. A.; Chapman, H. *Journal of Applied Crystallography* **2014**, *47*, 1118–1131.

[59] White, T. A.; Kirian, R. A.; Martin, A. V.; Aquila, A.; Nass, K.; Barty, A.; Chapman, H. N. *Journal of Applied Crystallography* **2012**, *45*, 335–341.

[60] Powell, H. R. *Acta Crystallographica Section D: Biological Crystallography* **1999**, *55*, 1690–1695.

[61] Powell, H. R.; Johnson, O.; Leslie, A. G. *Acta Crystallographica Section D: Biological Crystallography* **2013**, *69*, 1195–1203.

[62] Duisenberg, A. J. *Journal of applied crystallography* **1992**, *25*, 92–96.

[63] Kirian, R. A.; Wang, X.; Weierstall, U.; Schmidt, K. E.; Spence, J. C.; Hunter, M.; Fromme, P.; White, T.; Chapman, H. N.; Holton, J. *Optics express* **2010**, *18*, 5713–5723.

[64] Kim, J.; Cox, J. A.; Chen, J.; Kärtner, F. X. *Nature Photonics* **2008**, *2*, 733–736.

[65] Byrd, J.; Doolittle, L.; Huang, G.; Staples, J. W.; Wilcox, R.; Arthur, J.; Frisch, J.; White, W.; et al. In *Proc. of the 2010 International Particle Accelerator Conference*.

[66] Aquila, A.; et al. *Opt Express* **2012**, *20*, 2706–16.

[67] Kupitz, C.; Basu, S.; Grotjohann, I.; Fromme, R.; Zatsepin, N. A.; Rendek, K. N.; Hunter, M. S.; Shoeman, R. L.; White, T. A.; Wang, D.; et al. *Nature* **2014**.

[68] Coulibaly, F.; Chiu, E.; Ikeda, K.; Gutmann, S.; Haebel, P. W.; Schulze-Briese, C.; Mori, H.; Metcalf, P. *Nature* **2007**, *446*, 97–101.

[69] Coulibaly, F.; Chiu, E.; Gutmann, S.; Rajendran, C.; Haebel, P. W.; Ikeda, K.; Mori, H.; Ward, V. K.; Schulze-Briese, C.; Metcalf, P. *Proceedings of the National Academy of Sciences* **2009**, *106*, 22205–22210.

[70] Gati, C.; Bourenkov, G.; Klinge, M.; Rehders, D.; Stellato, F.; Oberthur, D.; Yefanov, O.; Sommer, B. P.; Mogk, S.; Duszenko, M.; et al. *IUCrJ* **2014**, *1*, 87–94.

[71] Rohrmann, G. *Journal of General Virology* **1986**, *67*, 1499–1513.

[72] Lacey, L. A.; Wennmann, J. T.; G., K. R.; A., J. J. *Manual of techniques in invertebrate pathology (2nd edition)*; Academic Press, 2012.

[73] Brehm, W.; Diederichs, K. *Acta Crystallographica Section D: Biological Crystallography* **2013**, *70*, 101–109.

[74] McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *Journal of applied crystallography* **2007**, *40*, 658–674.

[75] Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; et al. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 213–221.

[76] Emsley, P.; Lohkamp, B.; Scott, W.; Cowtan, K. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 486–501.

[77] Joosten, R. P.; Joosten, K.; Murshudov, G. N.; Perrakis, A. *Acta Crystallographica Section D: Biological Crystallography* **2012**, *68*, 484–496.

[78] Ravelli, R. B.; Nanao, M. H.; Lovering, A.; White, S.; McSweeney, S. *Journal of synchrotron radiation* **2005**, *12*, 276–284.

[79] Evans, G.; Polentarutti, M.; Djinovic Carugo, K.; Bricogne, G. *Acta Crystallographica Section D: Biological Crystallography* **2003**, *59*, 1429–1434.

[80] Ten Eyck, L. F. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **1977**, *33*, 486–492.

[81] Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G.; McCoy, A.; et al. *Acta Crystallographica Section D: Biological Crystallography* **2011**, *67*, 235–242.

[82] *http://bl831.als.lbl.gov/∼jamesh/mlfsom/ano_sfall.com*; 2013.

[83] Holton, J. M.; Classen, S.; Frankel, K. A.; Tainer, J. A. *FEBS Journal* **2014**, *281*, 4046–4060.

[84] de Sanctis, D.; Nanao, M. H. *Acta Crystallographica Section D: Biological Crystallography* **2012**, *68*, 1152–1162.

[85] Thorn, A.; Sheldrick, G. M. *Journal of applied crystallography* **2011**, *44*, 1285–1287.

[86] Persson, P.; Lunell, S.; Szöke, A.; Ziaja, B.; Hajdu, J. *Protein Science* **2001**, *10*, 2480–2484.

[87] Caleman, C.; Ortiz, C.; Marklund, E.; Bultmark, F.; Gabrysch, M.; Parak, F.; Hajdu, J.; Klintenberg, M.; Timneanu, N. *EPL (Europhysics Letters)* **2009**, *85*, 18005.

[88] Mackey, Z. B.; O'Brien, T. C.; Greenbaum, D. C.; Blank, R. B.; McKerrow, J. H. *Journal of Biological Chemistry* **2004**, *279*, 48426–48433.

[89] Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. *Acta Crystallographica Section D: Biological Crystallography* **1997**, *53*, 240–255.

[90] Girard, E.; Chantalat, L.; Vicat, J.; Kahn, R. *Acta Crystallographica Section D: Biological Crystallography* **2001**, *58*, 1–9.

[91] Lomb, L.; Steinbrener, J.; Bari, S.; Beisel, D.; Berndt, D.; Kieser, C.; Lukat, M.; Neef, N.; Shoeman, R. L. *Journal of applied crystallography* **2012**, *45*, 674–678.

[92] French, S.; Wilson, K. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **1978**, *34*, 517–525.

[93] Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. *Acta Crystallographica Section D: Biological Crystallography* **2009**, *66*, 12–21.

[94] Murphy, B.; Osipov, T.; Jurek, Z.; Fang, L.; Son, S.-K.; Mucke, M.; Eland, J.; Zhaunerchyk, V.; Feifel, R.; Avaldi, L.; et al. *Nature communications* **2014**, *5*.

[95] Immirzi, A. *Crystallographic Computing Techniques (Ahmed, FR, ed.), p 399*, 1966.

[96] Scott, H. A. *Journal of Quantitative Spectroscopy and Radiative Transfer* **2001**, *71*, 689–701.

[97] Caleman, C.; et al. *submitted* **2014**.

[98] Neves-Petersen, M. T.; Gajula, G. P.; Petersen, S. B. *Molecular Photochemistry—Various Aspects* **2012**, 125–158.

[99] Hogg, P. J. *Nature Reviews Cancer* **2013**, *13*, 425–431.

[100] Schneider, T. R.; Sheldrick, G. M. *Acta Crystallographica Section D: Biological Crystallography* **2002**, *58*, 1772–1779.

[101] Vagin, A.; Teplyakov, A. *Journal of applied crystallography* **1997**, *30*, 1022–1025.

[102] White, T. A.; Barty, A.; Stellato, F.; Holton, J. M.; Kirian, R. A.; Zatsepin, N. A.; Chapman, H. N. *Acta Crystallogr D Biol Crystallogr* **2013**, *69*, 1231–40.