Università degli studi di Padova
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in

Fisica

Cryo-electron microscopy structure of Photosystem II-LHCII supercomplex

RELATORE: Prof. Flavio Seno

CORRELATORE: Prof. Giuseppe Zanotti

LAUREANDO: Alessandro Grinzato

Anno accademico: 2015/2016

Al Dr Ermanno Grinzato

Contents

List of Figures

Sommario

Per molti anni la crio-microscopia elettronica (cryo -EM) è stata utilizzata unicamente per la determinazione di strutture biologiche di grandi complessi a bassa risoluzione (generalmente 10 Å o più). Grazie ai recenti sviluppi nella rivelazione degli elettroni e nel processamento delle immagini, la risoluzione ottenibile con la cryo-EM sta raggiungendo valori di risoluzione simili a quelli della cristallografia a raggi X (3-3.5 Å, nei casi migliori 2.5 Å) senza però avere la necessità di ottenre cristalli del campione e dunque aumentando significativamente il numero delle molecole di cui è possibile ottenere la struttura.
Questo progetto di tesi si propone di risolvere la struttura del fotosistema II in complesso con il light harvesting complex (PSII-LHCII) nel pisello (*Pisum sativum*), utilizzando immagini di cryo-EM.
Il PSII è un complesso della membrana tilacoidale normalmente attivo in forma dimerica e concorre nel processo di fotosintesi. I PSII delle piante contengono una serie di complessi perimetrali tra cui gli LHCII e le proteine di legame della clorofilla CP29, CP26 e CP24. Queste proteine circondano il PSII, assorbono l'energia luminosa e la trasmettono al centro di reazione.
Il complesso in esame, denominato C2S2M, è formato dal core (C2), che lega a sè tre trimeri LHCII, due di questi sono legati in modo relativamente forte (S), tramite una coppia di subunità monomeriche denominate lhcb5 (proteina CP26) e lhcb4 (proteina CP29), mentre un altro è legato più debolmente (M), tramite le subunità lhcb6 (proteina CP24) e Lhcb4. All'inizio di questo lavoro, in letteratura era riportato un modello tridimensionale del C2S2M2 ottenuto con una risoluzione di 30 Å. Mentre questo lavoro era in corso (giugno 2016) è apparsa la struttura del C2S2 (monomera, e senza le componenti accessorie) ad una risoluzione di 3.2 Å tramite cryo-EM.
La preparazione del campione e l'acquisizione delle immagini utilizzate in questo lavoro sono state eseguite dalla Dott. ssa Cristina Pagliano (Politecnico di Torino, DISAT -Dipartimento Scienza Applicata e Tecnologia).
Il campione è composto da dimeri di PSII-LHCII provenienti dal tilacoide del Pisello (*Pisum sativum*). Le immagini di crio-microscopia sono state acquisite con il microscopio a trasmissione di elettroni (TEM) FEI Titan Krios, associato ad un rivelatore a trasmissione di elettroni Falcon II. Le immagini acquisite al microscopio elettronico sono state processate con i software contenuti all'interno del pacchetto Scipion, da me installato sul cluster di calcolo HPC del CRIBI.
Preliminarmente, è stata calcolata la Contrast Transfer Function (CTF) di ciascuna micrografia, scartando tutte quelle che presentavano aberrazioni o uno scarso contrasto delle immagini rispetto al fondo. Delle micrografie rimanenti, sono state selezio-

nate manualmente le immagini del C2S2M che sono poi state classificate e allineate. Tale classificazione ha permesso di discriminare i complessi PSII-LHCII di tipo dimerico da quelli monomerici, oltre a consentire un'ulteriore cernita sulle particelle. Al termine della classificazione 2D, il processo di ricostruzione del modello si è concentrato sulle particelle in stato dimerico. Si è proceduto, quindi, alla determinazione di un modello iniziale e, tra i modelli ottenuti, è stato scelto quello maggiormente conforme a quanto precedentemente presente in letteratura. Il modello iniziale è stato quindi portato ad una risoluzione di 60 Å tramite opportuni filtri al fine di essere utilizzato per la classificazione 3D delle particelle. Al termine di quest'ultima fase sono state ottenute tre classi, rappresentanti rispettivamente lo stato conformazionale C2S2M, C2S2 e una classe di particelle da cui non è stato possibile estrarre una probabilità di densità di potenziale conforme con il modello iniziale.

Successivamente è stato effettuato il raffinamento della classe corrispondente allo stato C2S2M, eseguito suddividendo le particelle in due gruppi indipendenti tra loro, in modo da eliminare gli effetti bayesiani sulla ricostruzione del modello. Questa operazione ha portato ad ottenere una mappa di densità con risoluzione 16 Å; un successivo affinamento, ottenuto mascherando il fondo, ha portato ad una mappa di densità con risoluzione di 10.5 Å.

Utilizzando quest'ultima mappa ed estraendone la densità elettronica relativa a uno dei due supercomplessi componenti il dimero C2S2M è stato eseguito il docking dei modelli delle singole componenti del complesso già depositati nel Protein Data Bank. È stata anche estratta la mappa di densità elettronica relativa alla proteina CP24. In questo modo è stato possibile raffinarne il modello nello spazio reciproco. Tale raffinamento è stato eseguito iterativamente al fine di minimizzare il valore del fattore R cristallografico e dei parametri geometrici. In questo modo si è stati in grado di fornire un'ipotesi della struttura della proteina CP24 che compone la subunità monomerica Lhcb6. Infine, è stato eseguito un raffinamento nello spazio reale dell'intero modello.

In seguito alla pubblicazione della struttura completa del monomero PSII-LHCII in conformazione C2S2 (codice PDB 3JCU, EMD 6617), è stato eseguito un confronto tra i due modelli per verificare eventuali discrepanze, all'interno del limite di risoluzione.

Summary

For many years the cryo-electron microscopy (cryo-EM) was used only to determine low resolution (generally more than 10 Å) electron density maps of big proteins structure. Thanks to recent developments in electron detection and images processing, the resolution achievable with cryo-EM is reaching values similar to those obtained with X-ray crystallography (3-3.5 Å, maximum 2.5 Å) without the necessity to crystallize the sample. This thesis aims to solve the structure of photosystem II in complex with the light harvesting complex (PSII-LHCII) in pea (*Pisum sativum*), using cryo-EM images.

The PSII is a proteins complex of the thylakoid membrane involved in the photosynthesis normally active in his dimeric conformation. The PSII of plants contain a variety of peripheral complexes that include LHCII and chlorophyll binding proteins CP29, CP26 and CP24. These proteins surround the PSII, adsorb light energy and transmit it on to the reaction center.

In this thesis was used the complex called C2S2M, composed by the PSII core (C2), that binds three LHCII trimers, two of these are linked in a relative strong manner (S), via a pair of monomeric subunits called lhcb5 (protein CP26) and lhcb4 (protein CP29), while another is bound more weakly (M), via the monomeric subunits lhcb4 and lhcb6 (protein CP24). At the beginning of this thesis, the only model reported in literature was an electron density map of the C2S2M2 with a 30 Å resolution. during the last phase of this work (June 2016) the structure of C2S2 (monomer and without lhcb6 and S trimer) at 3.2 Å was published by Wei and colleagues.

Sample preparation and data harvesting with cryo-EM were performed by Dr. Cristina Pagliano (Politecnico di Torino, DISAT -Dipartimento Scienza Applicata e Tecnologia). The sample was composed by dimers of PSII-LHCII from the pea thylakoids (*Pisum sativum*). The images of cryo-microscopy were acquired with the transmission electron microscope (TEM), FEI Titan Krios, associated with a transmission electron detector Falcon II. The images acquired by electron microscopy were processed with the software contained in the Scipion package, that I installed on the HPC cluster of CRIBI.

Preliminary, the Contrast Transfer Function (CTF) of each micrograph was calculated and the micrographs whit aberration or poor image contrast were discarded. The images of C2S2M were manually selected from the remaining micrographs and then they were classified and aligned. This classification allowed to discriminate the dimer of PSII-LHCII complex from the monomer and to separate the good particles from the noisy one. At the end of the 2D calssification, the model reconstruction focused on the particles in the dimeric conformation. The initial model was determined choosing the one that most corresponds to the previous deposited model. The initial model was low pass filtered to a resolution of 60 Å using appropriate filters, in order to be used for the 3D classification of the particles. This last step divided the particles in three different classes, representing the conformational state C2S2M, C2S2 and a class of noisy particles respectively.

Subsequently, the class, that correspond to the C2S2M state, was refined dividing the particles in two independent groups in order to eliminate any Bayesian effects in the reconstruction. This refinement led to an electron density map with a resolution of 16Å. A subsequent refinement of the previous map led to an electron density map with a resolution of 10.5 Å, obtained by masking the background noise.

The models of the individual components of the C2S2M supercomplex, already deposited in the Protein Data Bank (PDB), were fitted in one of the monomer of the latter map. The electron density map corresponding to the protein CP24 was extracted. Then it was possible to refine the CP24 model in the reciprocal space. This refinement was performed iteratively in order to minimize the value of the crystallographic R factor. A hypothesis of the CP24 structure, that compose the lhcb6 subunits, is given. Finally, the entire model was refined in the real space.

After the publication of the complete structure of PSII-LHCII monomer in C2S2 conformation (PDB code 3JCU, EMD 6617) a comparison of the two models to check for any discrepancies, within the limit of resolution.

1 | Introduction

1.1 Electron microscopy

In agreement with classical optics, the resolving power of a microscope is linearly dependent on the radiation wavelength according to the Abbe Formula $d = 0.61\frac{\lambda}{NA}$ where d is the lateral resolution, λ is the beam wavelength and NA is the numerical aperture of the microscope. Therefore the fact that electrons wavelength is about ten of thousands time shorter than the photons suggests the possibility of using electron beams to obtain very high resolving power[1].

In principle an electron microscope operates as a normal optical microscope. However, since the normal optical devices do not deviate the electrons, they are replaced with magnetic lenses or electrostatic lenses that, acting on the electric charge of the electrons, resulting in its deviation.

An electron microscope is essentially composed by an electronic source of convenient intensity (generally an incandescent filament that emits electrons by the thermoelectric effect) and by a device that imparts strong accelerations to the electron beam emitted by subjecting them to a high voltage (from 20 to 300 kV). The accelerated electron beam passes through a capacitor (electrostatic or magnetic), hit the sample and is collected by a detector; all the process described above occurs in high vacuum ensured by a system of pumps.

The main difference between the various types of electron microscopes resides in how the electron beam is collected after it has interacted with the sample. In the Transmission electron microscope (TEM) the detector is positioned along the axis of the beam after the sample and it collects the transmitted electrons that carry information about the structure of the sample. The Scanning Electron Microscope (SET) produces images by probing the sample with a focused electron beam an collecting the backscattering electrons that carried information about the specimen surface. In the Reflection Electron Microscope (REM), as in TEM, an electron beam is incident to the sample surface, but in this case the reflected beam of elastically scattered electrons is detected.

1.1.1 Transmission electron microscopy

In TEM the electrons are accelerated and made to interact with a sample before the spatial distribution is detected. The interpretation of the obtained images is not as direct as it may seem, since there is not a unique correspondence between the electron counts and the presence of matter in the sample. Moreover the images obtained are two dimensional projections of three-dimensional structures.

Structure and functioning

The electrons generated by the electron gun are then accelerated in an electric field until they reach energies between 80 and 300 keV. The so accelerated electrons are directed along the column of the microscope, they are deflected and focused by some magnetic lenses, interact with the sample and then are projected on a florescent screen or detected by a sensor.

The lenses system of a TEM can be divided in four groups:

condensing lenses that focus the electron beam on the sample;

objective lenses that form the diffraction on the rear focal plane and the image of the sample in the image plane;

intermediate lenses that magnify the image;

projector lenses that project the image on the detector.

in addition to magnetic lenses there are quadrupoles, able to vary the beam direction (beam tilt), and pairs of quadrupoles, able to deflect the beam maintaining the direction of the beam (beam shift). In the TEM column there are also apertures, i.e. circular slits of various sizes that allow to block part of the beam allowing to select particular areas of the sample. The apertures in general, and in particular that of the condensing lens, have the function to select the most central part of the beam.

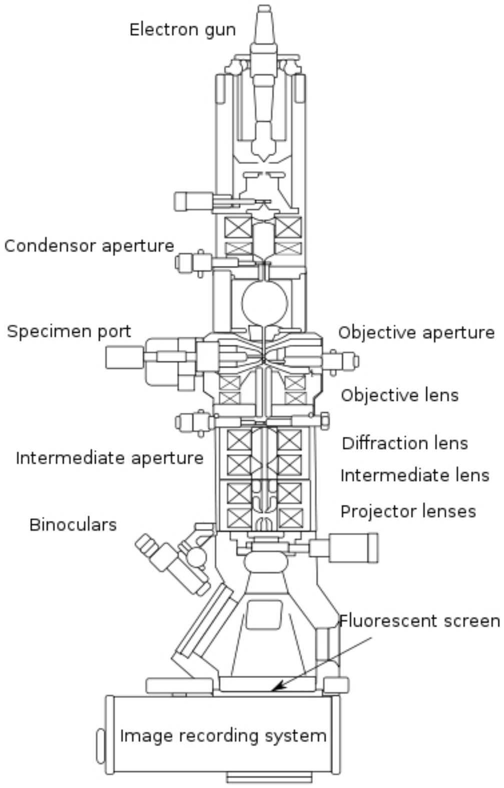


Figure 1.1: schematic diagram of a TEM

The magnetic lenses used in TEM are coils of copper wire(figure 1.2). This lenses generate a magnetic field that interacts with the electron beam. Magnetic lenses could be considered as ideal lenses, so the image formation could be described using the same principles of classical optics. The effect of the electric and magnetic field on the electron beam is well described by the Lorentz force: in the presence of an electric field $\vec{E}$ and a magnetic field $\vec{B}$, the Lorentz force is $\vec{F} = -e(\vec{E}+\vec{v}\times\vec{B})$ where e and $\vec{v}$ are the charge and the velocity of the electron. The intensity of the current in the coil determines the strength of the lenses, expressed by the focal length $f \propto \frac{V}{Ni}$ where V is the acceleration voltage, N is the number of windings an i is the current intensity. The electrons travel initially parallel to the lens axis until the interaction of their axial velocity with the radial component of the magnetic field produces a helical motion. This helical path of the electron beam generates a rotation of the image that is a typical characteristic of the magnetic lenses that distinguishes them from optical lenses.

In an magnetic lens the action of the magnetic field grows with the distance from the lens axis, so a magnetic lens acts as a weakly perturb the electron with small inclination with respect to the direction of the field.

In practice, magnetic lenses with small focal lengths are obtained by concentrating the magnetic field with pole pieces: the cooper coil is placed inside an iron box and only a small gap is left between two pole pieces. In this way the magnetic field is weak on axis and increases in strength toward the sides of the pole pieces, so the more the electrons travel off axis, the more strongly they are deflected.

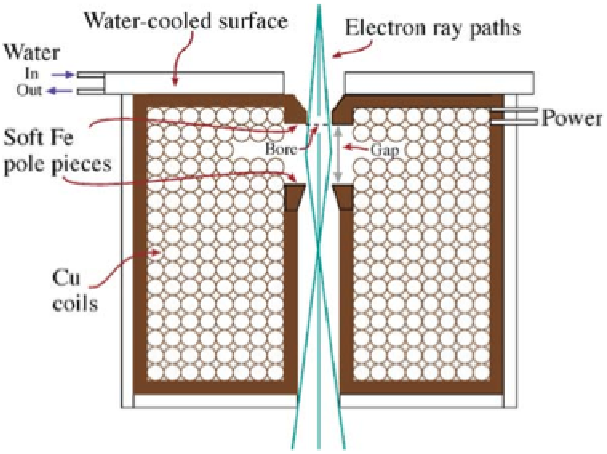


Figure 1.2: Schematic diagram of an magnetic lens. The soft-iron pole-pieces sit in the hole down the middle of the lens and are surrounded by the copper coils through which the current runs[2].

As in the case of the optical microscope, lenses introduce various types of aberrations and each has a different effect on the final image. Briefly, the most common TEM aberrations are astigmatism, spherical and chromatic aberration.

Astigmatism

Astigmatism occurs when electrons go through a non uniform magnetic field. This defect happens because it is impossible to make the iron around the coil perfectly cylindrical. The iron may also have microstructural defects that cause local variations in the magnetic field strength. Even if these difficulties were overcome, the apertures we introduce before the lens may disturb the field if they are not precisely centered around the lens axis. Furthermore, if the apertures are not clean, the contamination charges up and deflects the beam. So there are a variety of contributions to astigmatism, which globally distorts the image by an amount $R_a = \beta\Delta f$, where Δf is the maximum difference in focus induced by the astigmatism.

Spherical aberration

Spherical aberration reflects the inability of the lens to focus all the incident rays from a point source to another point. This defect is caused by the magnetic fields of the lenses that act unevenly on the rays off axis.
In paraxial conditions the rays are close to the lens axis and form only small angles with it: more an electron is far from the lens axis, less is its focal distance. Therefore, a point-like object form a disc image in the Gaussian image plane. The radius R_s of this disc, called aberration disc, depends on the angular opening β according to the expression $R_s = C_s\beta^3$ where C_s is the spherical aberration coefficient that, for a typical TEM, is 1÷3 mm[2]. For a magnetic lens placed behind the objective lens the aperture of the incident beam is smaller than the opening of the beam incident on the objective lens by a factor given by the magnification of this lens. This explains why only the aberrations of the objective lens are taken into account when the limit of resolution of a microscope is evaluated.

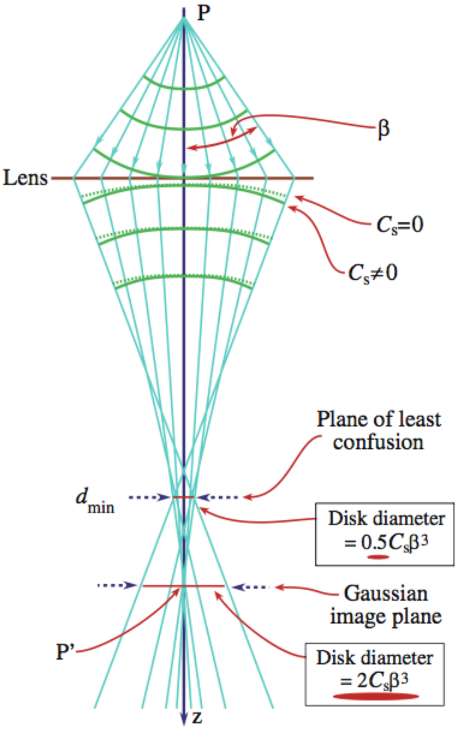


Figure 1.3: spherical aberration[2]

Chromatic aberration

The focal distance of a magnetic lens vary with the energy of the electrons; lenses strongly retain electrons with low energy, and then from a point in the object plane a disk in the Gaussian image plane is obtained. The R_C of this disc radius is given by the expression $R_c = C_c \frac{\Delta E}{E} \beta$, where C_c is the chromatic aberration coefficient, ΔE is the variation of the electron energy from its mean value E , and β is the opening angle of the lens. The chromatic aberration coefficient C_c of a magnetic lens usually has a numerical value slightly lower than that of the focal length[2]. This aberration degrades the image when the electrons in the beam cease to be monoenergetic and this can occur when electrons are generated from the electron gun with a spread of energies; the acceleration voltages or currents in the coils fluctuate over time; the electron beam loses energy through collisions, passing through a sample. In more modern microscopes the stability of the accelerating voltage and the current in the lens are well controlled, so they are no more influential in chromatic aberrations, when compared with the loss of energy associated to the electrons transmitted through the sample. The inelastic scattering of high energy electrons by the plasmonic excitations is a classic effect for which the electrons lose 10÷20 eV; this effect is even more important in the case of thick samples. So, to minimize chromatic aberrations in TEM, it is necessary to use samples that are as thin as possible[2].

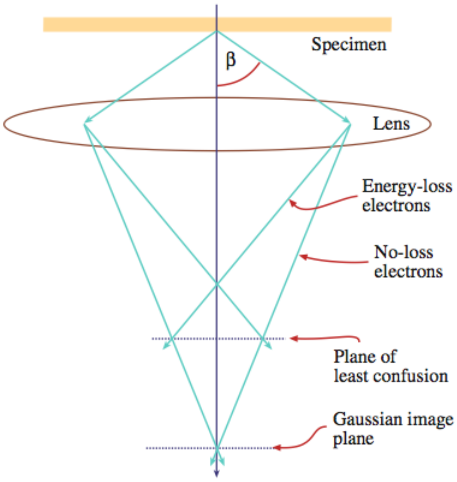


Figure 1.4: cromatic aberration[2]

Resolution

As in any optical system, the theoretical resolution (i.e. the resolution without any aberrations) of a TEM follows the Rayleigh criterion for which two points cannot be distinguished if their distance is less then their Airy radius $R_A = 1.22\frac{\lambda}{\beta}$, were λ is the beam wavelength and β the angular aperture. TEM resolution is mainly affected by the spherical aberration. The resolution of the object in given by the combination of the Rayleigh criterion and the aberration $R = (R_A^2 + R_s^2)^{\frac{1}{2}}$ that

could be approximated to $R(\beta) \approx \left[\left(\frac{\lambda}{\beta}\right)^2 + (C_s\beta^3)^2\right]^{\frac{1}{2}}$.
Since the two terms vary differently with the aperture collection angle β we find that $\frac{\lambda}{\beta} = AC_s\beta^3$ so $\beta = A\left(\frac{\lambda}{C_s}\right)^{\frac{1}{4}}$, where A depends on various terms included in the definition of resolution. For a TEM with a C_s=3mm $R = 0.9(C_s\lambda^3)^{\frac{1}{4}}$ [2].

1.1.2 Detector

Electrons detectors play a key role in microscope resolution. The first electronic microscopes were equipped with film[1], subsequently, with the development of technology, films have been replaced by Charge-Coupled Device detector (CCD) that, although charaterize by a lower resolution[3, 4], ensured a much more rapid acquisition and a greater ease in the analysis of images[5]. The latest generation microscopes are gradually abandoning the CCD in favor of the Direct Electron Detector[6]. A quantitative description of a detector performance is provided by the Detection Quantum Efficiency (DQE), defined as the ration between the Signal to Noise Ratio (SNR) of the output over the SNR of the input signal, $DEQ = \frac{SNR_out}{SNR_in}$. An ideal detector has a $DQE = 1$, a real one has $DQE < 1$ [2].

CCD

The CCD consists of an integrated circuit formed by a line, or by a grid, of semiconductor elements (photosites) capable of accumulating an electric charge proportional to the intensity of the electromagnetic radiation that hits them. These elements are coupled so that each of them, stressed by an electrical pulse, transfers its charge to another adjacent element. When a timed pulsed sequence is send to the device, it gives in output an electrical signal thanks to which it is possible to reconstruct the matrix of pixels in the image projected on the surface of the CCD itself (figure 1.5). In a TEM the incident electron beam hits a scintillator, generating light, which is partially captured by fiber optics, and directed onto the cooled CCD.

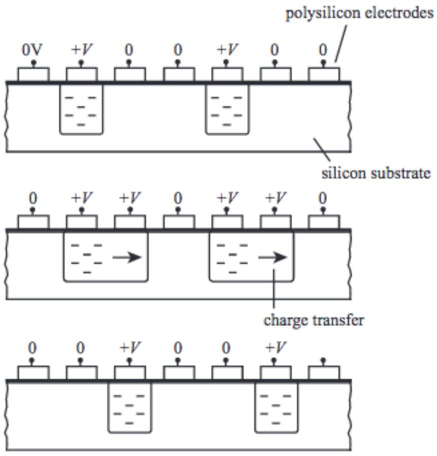


Figure 1.5: Schematic diagram of a three-phase CCD that show the process of charge integration and readout[5].

When a CCD is used in an electron microscope the three main sources of noise are: dark current, readout noise and spurious events due to X-rays or cosmic rays. Dark current happens when electrons in the silicon crystal lattice possess a thermal energy that allows them to jump spontaneously across the band gap into the conduction band, becoming a free electron. The generation of free electrons without any illumination falling on the CCD detector results in a dark image, which needs to be subtracted from the acquired image. Because dark current generation is strongly temperature-dependent, its effect can be reduced by cooling the CCD. Another disadvantage of the CCD cameras is their Point Spread Function: a 200 kV electron has quite a large interaction volume in the scintillator, thus generating light in a region far exceeding an individual pixel on the CCD chip. Also, the two-step conversion, from electrons to light and from light to charge, leads to a reducthion of the achievable signal to noise [6]. Anyway, when cooled, a CCD have a good DQE ($>$0.5)[2].

Direct Electron Detector

To overcome the disadvantage of the double conversion step that occurs in the CCD (from electrons to photons and vice versa) a new generation of active pixels sensors based on Complementary Metal-Oxide Semiconductor CMOS tecnology were develloped. They are capable of detecting the electrons directly. When an incident beam passes through a thin CMOS layer ($150 \mu m$[6]) leaves an ionization trail. The electron (or hole) generated in the semiconductor is accelerated to an adjacent contact creating a signal. This greatly increases the SNR of the incoming signal, and therfore the DQE, compared to CCD and photographic films [3, 7]. Furthermore the small thickness of the CMOS layer minimizes the lateral charge spread of the ion trail resulting in higher spatial resolution with respect to that of CCD camera. a direct electron detector has also an high frame rate, with no dead time between frame. This high frame rate delivers intrinsic dose fractionation during image acquisition, which can be exploited for beam induced motion correction[8, 9], damage compensation, and other image processing techniques[10, 11] .

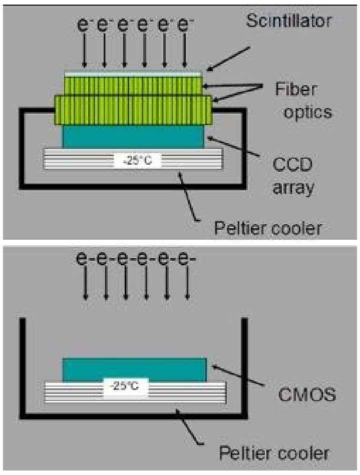


Figure 1.6: Comparison between a direct electron detector (bottom) and a CCD (top)

1.1.3 Cryo-electron microscopy

For many years the cryo-electron microscopy (cryo -EM) was used only for the determination of biological low resolution structures of large complexes (generally up to 10 Å). Thanks to recent developments in the electron detection and in images processing, the resolution achievable with cryo-EM is reaching values similar to those reached with the X-ray crystallography (3-3.5 Å, at best 2.5 Å), significantly increasing the number of molecules of which it is possible to obtain the structure [12].
This resolution revolution of the TEM field is due to the great developments of technologies and softwares used for the determination of the structures[13, 14].

Sample preparation

There are several techniques for the TEM sample preparation depending on the type of specimen analyzed and on the scope of the analysis. Since biological sample are easyly damaged by the collision with the electrons, only a low energy electron beam must be use in cryo-EM. This leads to a low contrast of the biological sample. In order to improve contrast the sample can be stained; staining is usually done with heavy metal salts commonly containing molybdenum, uranium, or tungsten. The limit of the negative staining is related to the fact that the stain shows only the surface features, proving few information of the internal structure. The resolution is believed to be limited by the dimension of the stain particles resulting in a 3D final map of low resolution [15]. For the 3D reconstruction a common disadvantage can come from the flattening induced by the staining dehydration resulting in a geometrical distortion. Cryo-EM is a method for the sample preparation that preserves the specimen through a fast freezing in a thin layer of amorphous or vitreous ice. This technique avoids dehydration or support adsorption, resulting in a close to native state of the target molecule [12]. Another important parameter in the Cryo-EM sample preparation is the thickness of the ice: a very thick ice results in a low SNR that prevents the sub-nanometer resolution[16, 17].

Immaging and data collection

The grid with the sample is inserted in the TEM to proceed to the acquisition of the micrographs. The quality of the micrographs and the acquisition apparatus play a key role in achieveing a good resolution. Microscopes like Titan Krios working at 300 kV increases the acceleration of the electrons and therefore the SNR[18]. Moreover, as seen in section 1.1.2, the new direct electron detector allow to correct the beam induced motion of the sample, increasing the accuracy of the subsequent analysis processes[8].

Single particles analysis

A significant part of EM procedure is the image processing, in which a key tool is the development of good software[19] that can reduce the operator influence on the process of reconstruction[20]. Single particles analysis reconstruction methods were developed by several laboratories, in particular by Joachim Frank and coworkers. They deal with single particle of the biological sample that assume random or multiple orientations on the grid [13]. The singe particles reconstruction workflow can be divided in four main steps to generate a final 3D volume: preprocessing of the micrographs, particle picking, 2D classification and 3D reconstruction.

Micrographs processing The first step is the evaluation of the quality of the the collected micrographs. All the aberration introduced by the microscope are characterized by the Contrast Transfer Function (CTF) . The CTF correction must be done on each image in order to reject micrographs with strongly asymmetric rings, (astigmatism) or that fade in a particular direction[21].

Particles picking The quality of the particles is a critical point for the final reconstruction because the inclusion of a huge number of bad particles may preclude the structure determination [19]. Particles can be selected in a manual, semi or fully automated manner. When the particles are localized, they are windowed (boxed) and assembled into a stack.

2D classification 2D classification groups the particles in different orientation set. This reveals the presence of invalid particles, image artifacts or empty fields that need to be removed, or it highlights the presence of few views that will conduce to an unsuccessful 3D analysis. It also verifies the presence of high quality classes with high SNR necessary for the computational determination of the 3D structure[19].

3D recostruction If there is no reasonable template or guesses for the structure, the average of the classes evaluated in the previous step must be used to generate an initial model. This first model must contain the main features of the 3D object at low resolution. This 3D model is then refined with a 3D projection-matching procedure that modifies the orientation parameters of single particle images (projections) to achieve a better match with the re-projections computed from the approximation of the structure[19]. An indicator of the map goodness is the Fourier Shell Correlation (FSC). It indicates the level of the SNR as a function of the spatial frequency and the resolution of the map.

$$FSC(r_i) = \frac{SNR(r_i)+\frac{2\sqrt{SNR(r_i)}}{\sqrt{n(r_i)}}+\frac{1}{\sqrt{n(r_i)}}}{SNR(r_i)+\frac{2\sqrt{SNR(r_i)}}{\sqrt{n(r_i)}}+1}$$

where r_i is the radius of a shell and $n(r_i)$ is the number of voxels (3D pixels) contaeined in a given shell of radius r_i[22]. Resolution is however depends on the cut-off level of the FSC curve:

the conventional FSC cut-off is 0.5, but, recently, a new resolution estimation method, called gold standard FSC, uses a 0.143 cut-off. This cut-off value is selected based by relating EM results to those in X-ray crystallography[23, 24, 25]. It is important to notice the the quality of an EM map is described by the whole FSC curve[26] and there are maps with the same nominal resolution that differs significantly in the overall quality or in their local resolution [27].

1.2 PSII–LHCII supercomplex

In this thesis cryo-EM and single particles analysis were used to obtain the structure of the Photosystem II-Light Harvesting Complex supercomplex (PSII-LHCII) of *pisum sativum.*

PSII-LHCII supercomplex, the first part of the photosynthetic machinery, is a large membrane proteins complex located in the thylakoid membranes of cyanobacteria and chloroplasts of plants[28].

Chloroplasts are semi-autonomous organelles comprising two envelope membranes, an aqueous matrix known as stroma, and internal membranes called thylakoids. All of the light-harvesting and energy-transducing functions are located in the thylakoids, which form a physically continuous membrane system that encloses an aqueous compartment, the thylakoid lumen. With few exceptions, thylakoids are differentiated into stacked grana and non-stacked stroma membrane regions. The most prominent effect of membrane stacking is the physical segregation of most Photosystem II to stacked grana membranes[29]. Therefore, PSII-LHCII may function as a dimer in the thylakoid membrane[30].

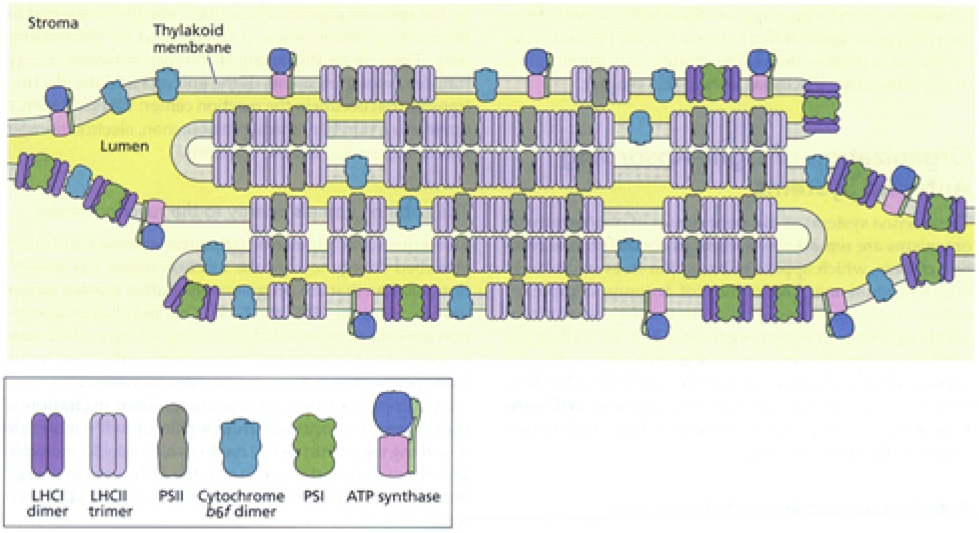


Figure 1.7: distribution of the proteins complex on staked and unstaked thylakoid membrane (source: Taiz L., Zeiger E., 2010)

In plants or green algae the PSII core is associated with its Light Harvesting antenna Complex (LHCII) to form the PSII-LHCII supercomplex[31].

1.2.1 PSII

In plants and green algae the PSII core complex is mainly embedded in the stacked regions of the thylakoid membranes, where it is organized as a dimer[30], each monomer consisting of several proteins including:

- D1 and D2 proteins that compose the photochemical Reaction Center (RC);
- CP47 and CP43 proteins that act as inner antenna proteins;
- The extrinsic polypeptide subunits (Psbs) that form the Oxygen Evolving Complex (OEC) on the lumenal side of the membrane and play a role in nonphotochemical quenching (NPQ)[32];
- Several Low Molecular mass sub units (LMM) that that make up more than half of the entire complex and have a stabilizing and cofactors-binding role.

Nowaday the highest resolution structure available for PSII core complex has been obtained by X-ray diffraction[33], where it is possible to see that PSII is composed of two halves related by twofold rotational symmetry with its axis normal to the thylakoid membrane[34] .

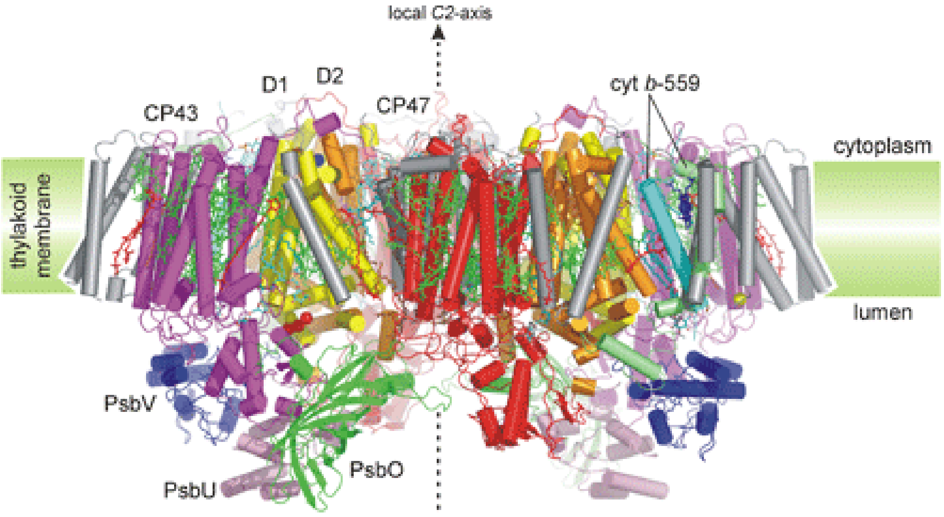


Figure 1.8: photosystem II in the tylakoid membrane[34]

1.2.2 LHCII

The most abundant PSII-associated LHCII complex, called *major*, consists of homo or hetero trimers of Lhcb1, Lhcb2 and Lhcb3 polypeptides [35] , whose high resolution structures have been solved by X-ray crystallography[36, 37]. LHCII have three membrane spanning regions connected by both stromal and lumenally exposed loops and bind a total of 14 chlorophyll (Chl) molecules (8 Chl a and 6 Chl b) plus 4 carotenoid molecules[38]. In addition, there are three *minor* LHCII antenna polypeptides, termed Lhcb4 (CP29), Lhcb5 (CP26) and Lhcb6 (CP24)[39], which usually occur in monomeric form.To date, among the minor LHCII antenna proteins, the three dimensional structure is available at high resolution only for Lhcb4[40, 41] and Lhcb5[41].

PSII core complex can be bound to a variable number of LHCII to form different types of PSII-LHCII supercomplexes, named accoding to their composition[42]. The dimeric PSII core complex (C2) strongly binds two copies of the monomeric Lhcb4 and Lhcb5 and two LHCII strong bounded trimers(S trimer) in order to form the C2S2 supercomplex, which can be regarded as a basic building block of PSII in vivo. There are, also, larger PSII-LHCII supercomplexes, containing two extra copies of the monomeric Lhcb6 with one or two additional LHCII trimers (M trimer) moderately bound to the dimeric PSII core complex via Lhcb4 and Lhcb6. This supercomplexes are known as C2S2M1-2 and have been found to represent the basic organization of the PSII in *Arabidopsis thaliana* thylakoid membranes.

Occasionally even larger supercomplexes have been observed in isolated spinach thylakoids fragments, with one or two additional LHCII trimers (L trimer) even more loosely bound to the dimeric PSII core complex via Lhcb6, and are known as C2S2M2L1-2 [43].

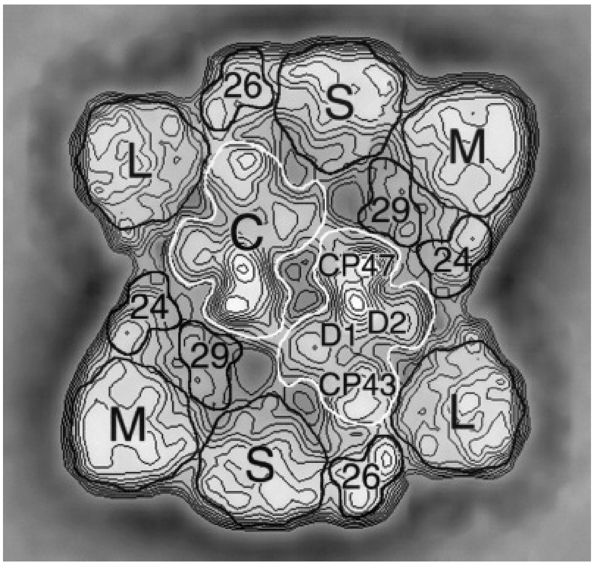


Figure 1.9: top view of a constructed image of a complete PSII-LHCII supercomplex. PSII boundaries are depicted in white, whereas those of the various antenna complexes are depicted in black. C is the PSII core complex; S is strongly bound LHCII; M is moderately bound LHCII; L is loosely bound LHCII; 29 is protein CP29; 26 is protein CP26; 24 is protein CP24. The right-hand PSII core complex shows the locations of its four largest subunits. Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure.

All the structures of the PSII-LHCII supercomplex were obtained by cryo-EM; at the beginning of this work, in literature was reported a three-dimensional model of C2S2M2 obtained with a resolution of 30 Å[38] and a 17 Å model of C2S2 [44]. Very recently (June 2016) a new structure of C2S2 in monomer conformation at a resolution of 3.2 Å[41] was deposited. The difference in resolution between the C2S2M2 and C2S2 structure is due to the greater difficulty in obtaining a homogeneous sample of the C2S2M2 conformation.

1.2.3 PSII-LHCII role in photosynthesis

Photosynthesis is the process by which plants, algae and photosynthetic bacteria use light energy to drive the synthesis of organic compounds. The photosynthetic process results also in the release of molecular oxygen and the removal of carbon dioxide from the atmosphere that is used to synthesize carbohydrates (oxygenic photosynthesis).

Plant photosynthesis is driven by visible light that is absorbed by pigment molecules (mainly chlorophyll (Chl) and carotenoids) bounded to LHCIIs. The latter surround the RC that serve as an antenna. Photosynthesis is initiated by the absorption of a photon by an antenna molecule, which occurs in about a fs and causes the transition of a Chl from the electronic ground state to an excited state. Within $10\div 13$ s the excited state decays by vibrational relaxation to the first excited singlet state. This transition is guided by the structure of the protein. Owing to the proximity of other antenna molecules with the same or similar energy states, the excited state energy has a high probability of being transferred by resonance energy transfer to a neighbor. The probability of transfer is dependent on the distance between the transition dipoles of the donor and acceptor molecules ($\propto \frac{1}{D^6}$), the relative orientation of the transition dipoles, and the overlap of the emission spectrum of the donor molecule with the absorption spectrum of the acceptor molecule[45].
Light energy that arrives to the PSII RC is used to drive two chemical reactions: the water oxidation and the reduction of plastoquinone[46]. Here a special form of Chl, P680 (where the number 680 indicates the maximum absorption wavelength $\lambda = 680$nm) is excited to the P680* state that is a strong reducing agent. In the time scale of picosecond (ps) P680* reduces a pheophytin molecule (Pheo) to form the radical pair P680$^+$ Pheo$^-$. Subsequent electron transfer steps prevent the primary charge separation from recombining; this is accomplished by transferring the electron within 200 ps from pheophytin to a plastoquinone molecule (Q$_A$) that is permanently bound to PSII. Q$_A$ normally is a two-electron acceptor, but in PSII works as one-electron acceptor at the Q$_A^-$ site. The electron on Q$_A^-$ is then transferred to another plastoquinone that is loosely bound at the Q$_B^-$ site. Q$_B^-$ is a two-electron acceptor second photochemical turnover that reduces Q$_B^-$ to Q$_B^{2-}$, which is then protonated to plastoquinol PQH$_2$ and released from the Q$_B^-$ binding site of PSII into the lipid bilayer, where it is subsequently oxidized by photosystem I (PSI). The full reduction of plastoquinone requires the addition of two electrons and two protons, i.e. the addition of two hydrogen atoms. Because the Q$_B^-$ site is near the outer aqueous phase, the protons added to plastoquinone during its reduction are taken from the outside of the membrane.
Photosystem II is the only known protein complex that can oxidize water, resulting in the release of O$_2$ into the atmosphere. Energetically, water is a poor electron donor, the redox midpoint potential of water being +0.82 V. In photosystem II this reaction is driven by P680$^+$ (the midpoint potential of P680$^+$ is estimated to be +1.2 V at pH 7). It is known that P680$^+$ oxidizes a tyrosine on the D1 protein and that Manganese (Mn) plays a key role in water oxidation. Four photochemical turnovers of the RC provide a cluster of four Mn ions and one Ca ion (Mn$_4$Ca^{10+}), with a total of four oxidizing equivalents that are used to oxidize two water molecules to

form dioxygen. Each oxidation state generated within the oxygen-evolving complex (OEC) is represented as an intermediate of the S-state cycle (fig. 1.10.b) and removes one electron. The net reaction results in the release of one O_2 molecule, the deposition of four protons into the inner water phase, and the transfer of four electrons to the Q_B^- site (producing two reduced plastoquinone molecules)[28, 30, 46].

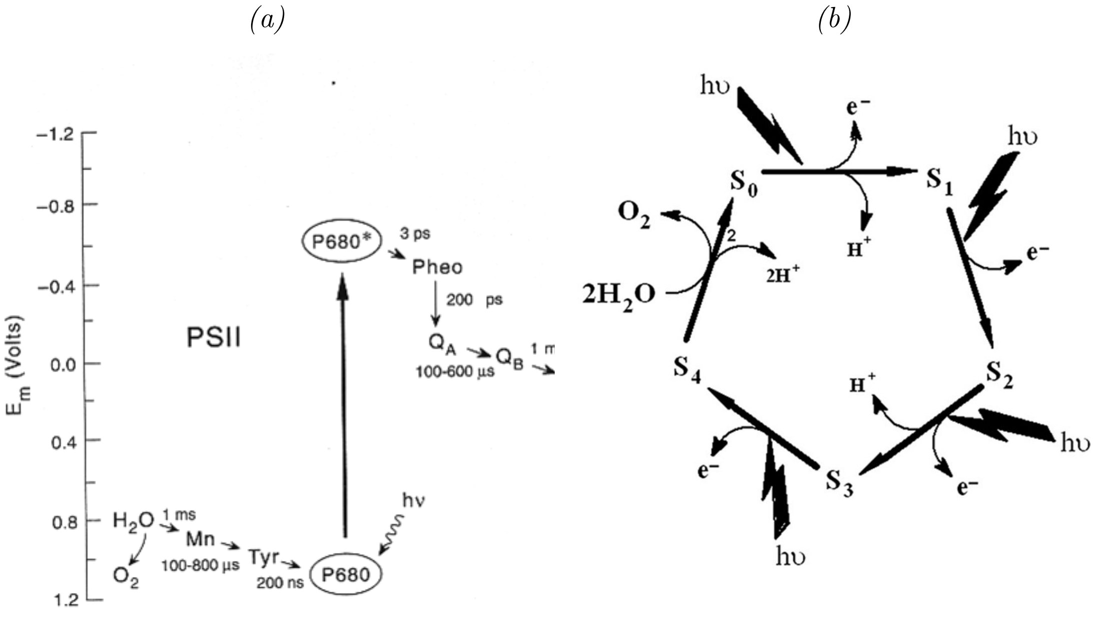


Figure 1.10: (a) Diagram of energy level for light induced electron transfer in PSII. (b) S-state cycle shows how the absorption of four photons of light (hν) drives the splitting of two water molecules and the formation of O_2 through a consecutive series of five intermediates (S0, S1, S2, S3, and S4). Protons (H^+) are released into the lumen during this cycle, except for the S1 to S2 transition.

2 | Materials and method

2.1 Sample preparation and data harvesting

Sample preparation and data harvesting with cryo-EM were performed by Dr. Cristina Pagliano (Politecnico di Torino, DISAT -Dipartimento Scienza Applicata e Tecnologia). The sample was prepared in her laboratory, while the data were measured at Max-Planck institute of biochemistry (Munich- Germany).

2.1.1 Sample preparation

Stacked pea thylakoid membranes were extracted from plants at the end of the daily dark phase of growth [47], and finally suspended and stored in 25 mM MES pH 6.0, 10 mM NaCl, 5 mM $MgCl_2$ and 2 M glycine betaine.
Stacked thylakoid membranes, at a chlorophyll (Chl) concentration of 1 mg mL-1, were treated with 50 mM n-dodecyl-α -D-maltoside or n- dodecyl-β-D-maltoside (α-DDM or β-DDM) for 1 min at 4 °C in the dark [48].
Phenylmethylsulphonylfluoride (500 μ M) was present during the solubilization to inhibit protease activity. After centrifugation, at 21000 g for 10 min at 4 °C, 700 μ L of supernatant was added to the top of a linear sucrose gradient, previously prepared by a freezing and thawing cycle applied to ultracentrifuge tubes filled with a buffer made of 0.65 M sucrose, 25 mM MES pH 5.7, 10 mM NaCl, 5 mM $CaCl_2$ and 0.03% (w/v) DDM. Centrifugation was carried out at 100000 g for 12 h at 4 °C (Surespin 630 rotor, Thermo Scientific).
The sucrose band containing PSII-LHCII supercomplexes was carefully harvested using a syringe and, if necessary, concentrated by membrane filtration via Amicon Ultra 100 kDa cut-off devices (Millipore) at a Chl concentration around 1 mg mL-1 and then flash frozen for storage at -80 °C.
Optimal separation of the thylakoid membrane protein complexes and isolated PSII-LHCII supercomplexes was obtained [49], using large pore blue native polyacrylamide gel electrophoresis (lpBN-PAGE).

2.1.2 Data harvesting

Sucrose gradient band containing α PSII-LHCIIsc was carefully harvested using a syringe and concentrated by membrane filtration via Amicon Ultra 100 kDa cut-off devices (Millipore) at a Chl concentration around 1 mg mL-1 and then flash frozen for storage at -80 °C until use. Thawed samples were used for cryo-grid preparation in a Vibrobot system (humidity 100%, 22 °C). Four microliter droplets of sample were applied to Lacey carbon grids previously rendered hydrophilic in a plasma cleaner and let stand to adhere for 60 s. After a short washing step, using four-microliter of distilled water, excess suspension was removed by blotting once for 4 s with filter paper, subsequently the grid was plunged into liquid ethane for vitrification.
TEM was performed using a FEI Titan Krios TEM [18] operated at 300 keV. A total of 7800 images were automatically recorded at 59000x on a Falcon II direct electron detector[50], leading to a final pixel size of 1.4 Å at the specimen level.

2.2 3D map reconstruction

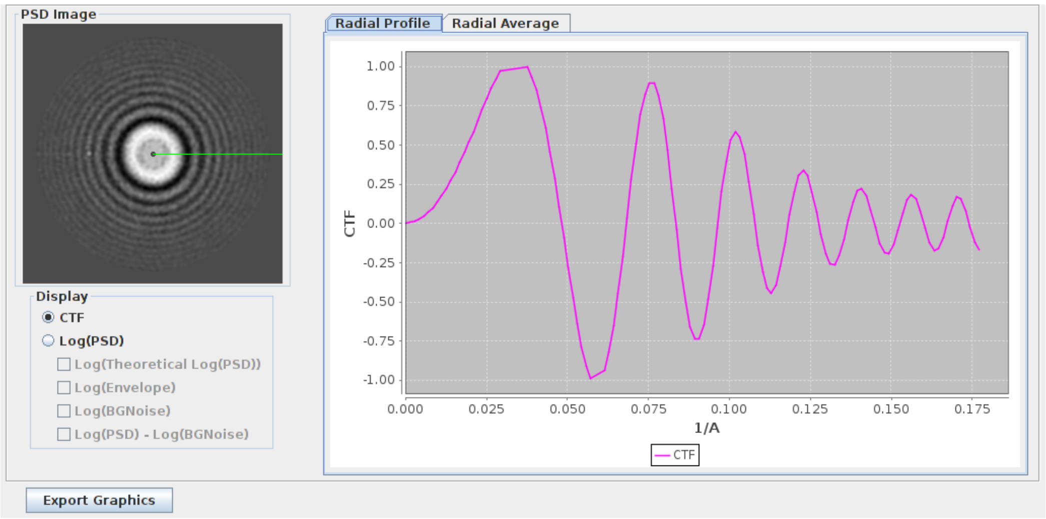
The collected micrographs were analyzed with Scipion[51], a framework for cryo electron mycroscopy image processing that integrates several software packages such as Xmipp[52], EMAN2[53] and RELION[54]. In this work Xmipp was used for micrographs Contrast Transfer Function (CTF) estimation, particles extraction and final model masking, EMAN2 for the manual particles picking and the evaluation of the initial volume, RELION for particles classification and refining (fig 2.12). Because of the computational requirement of the software, Scipion was istalled in the HPC cluster of *Centro di ricerca interdipartimentale per le biotecnologie innovative* (CRIBI) (Appendix D).

2.2.1 CTF estimation and particles picking

The effect of microscope aberrations is described in Fourier space by the contrast transfer function (CTF) (Appendix A.1). The CTF of the 7800 collected micrographs was estimated with the *Xmipp CTF estimation* protocol that is based on a parametric Power Spectrum Density (PSD) estimation, using the Auto-Regressive Moving-Average (ARMA)[55] models. These models assume that the value of each pixel can be approximated by the values of its neighboring ones, plus a correlated noise term[56]. CTF was not corrected in this step because CTF correction is more accurate if made inside the *RELION 2D classification* protocol on each particles. Once the CTF of each micrographs was estimated, micrographs that had the resolution rings with significant aberration such as astigmatism or low SNR (Fig. 2.1) were discarded.

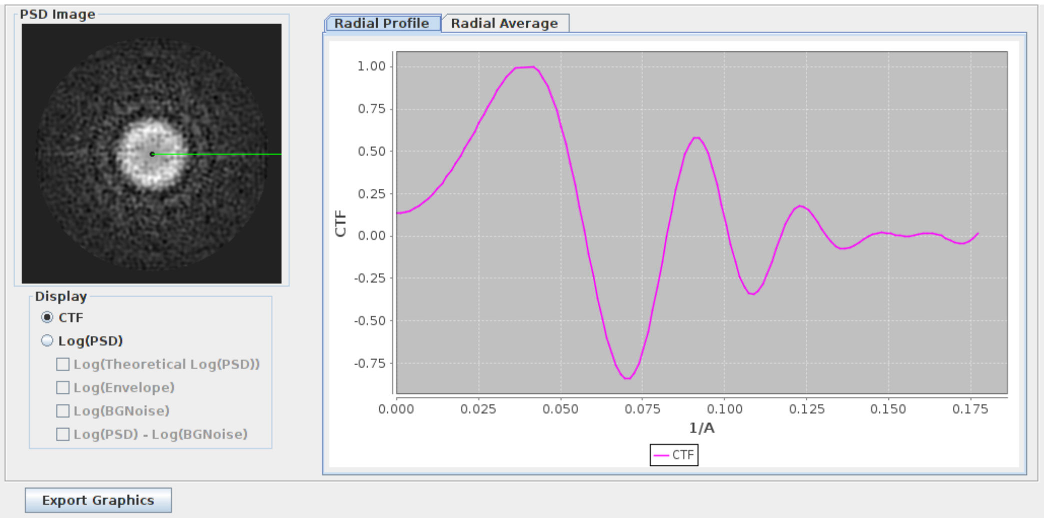
From the remaining 6336 micrographs 29748 particles were manually picked using *eman2-boxer* protocols with a box size of 300 pixels (420 Å). This was one of the crucial part of the work, we had to pay attention not to pick particle that have in their box other particles, ice or the carbon grid.

(a)



Display
● CTF
○ Log(PSD)
□ Log(Theoretical Log(PSD))
□ Log(Envelope)
□ Log(BGNoise)
□ Log(PSD) - Log(BGNoise)

(b)



Display

● CTF

○ Log(PSD)

□ Log(Theoretical Log(PSD))

□ Log(Envelope)

□ Log(BGNoise)

□ Log(PSD) - Log(BGNoise)

Figure 2.1: PSD image and CTF profile of a micrograph without significant aberration (a) and with a low SNR (b).

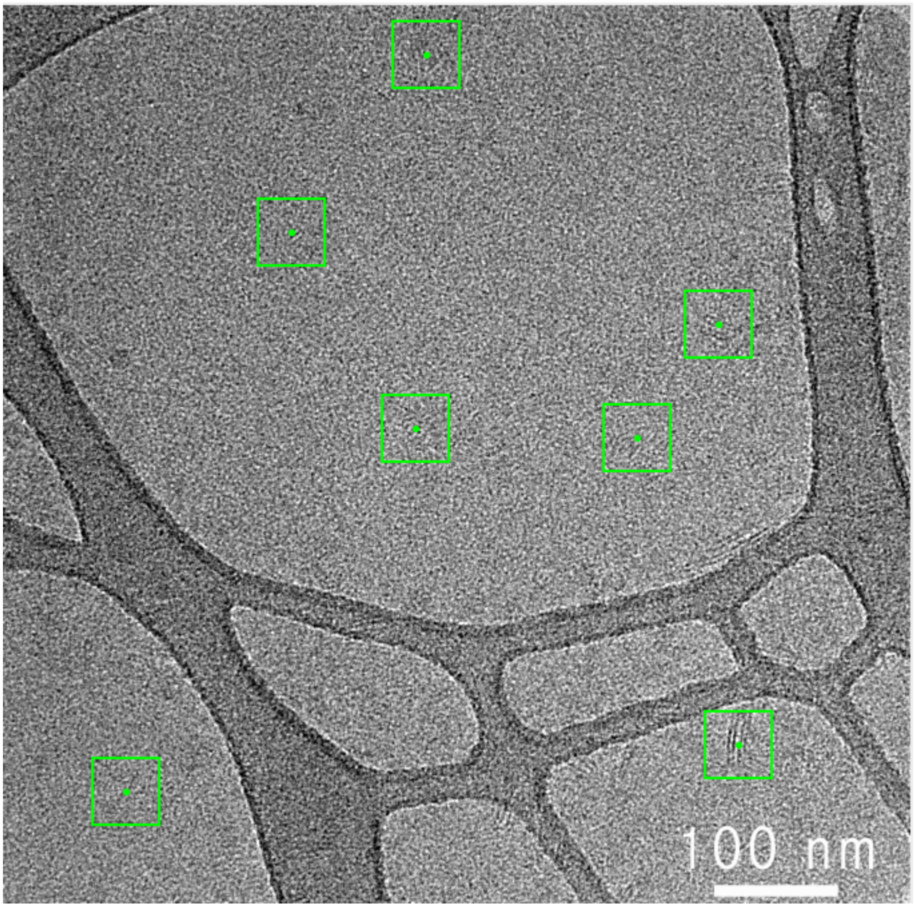


Figure 2.2: image of a micrograph from which the particles used for the analysis (green boxes) are selected.

2.2.2 Particles preprocess

Before the reconstruction process could start, the picked particles must be normalized and cleaned from artifacts or dead/hot pixels effects.
The *RELION preprocess particles* protocols was used to normalize the particles calculating the standard deviation and the average values outside a circle of 140 pixels (196 Å) radius with its center in the middle of the particles box. This value was chosen to be sure to select only the background of each box. Normalization is a fundamental step because different particles could have different background value (for example particle belonging to different micrographs, or to micrograph with an ice inhomogeneous thickness) and, if compared without normalization, can led to a wrong interpretation of the images and thus to a wrong model.
In order to remove artifacts, dust, ice anisotropy or dead/hot pixels of the detector from the image background the dust removal options was used. Dust removal replace all pixels with values above 3.5 times the standard deviation of the noise with a random values from a Gaussian distribution. *RELION preprocess particles* was also used to invert contrast, because the particles picked are black over a withe background but RELION need withe over black particles to perform the next steps.

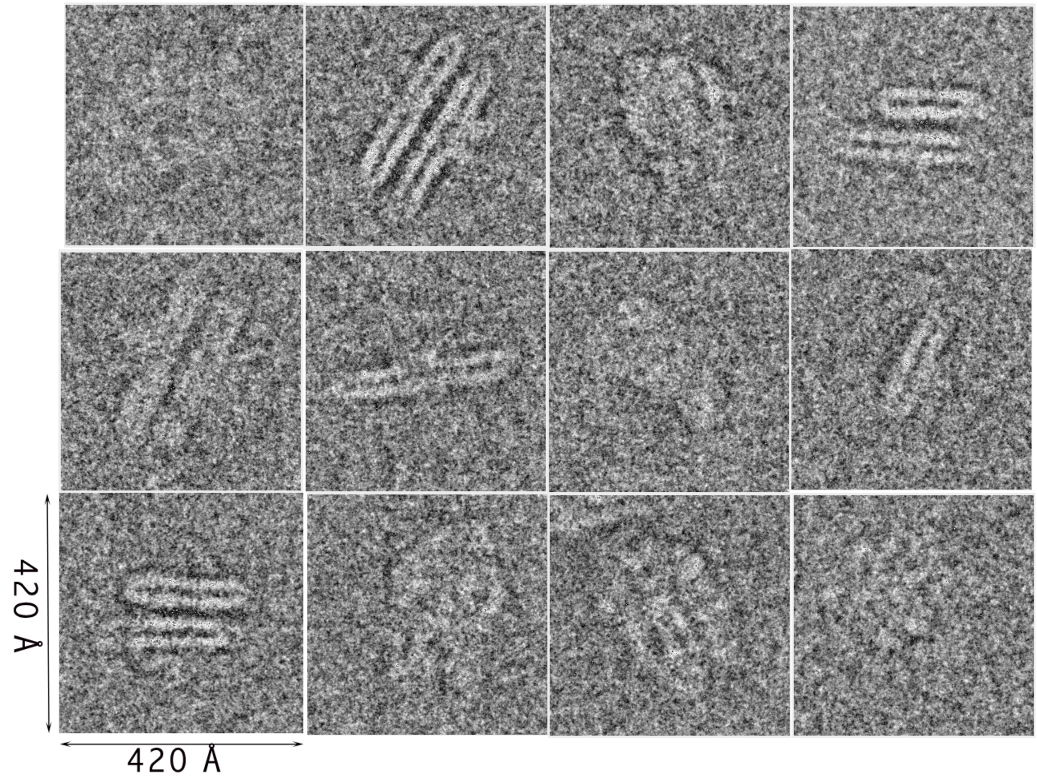


Figure 2.3: A sample of the preprocessed particles

2.2.3 2D classification and alignment

The *RELION 2D classification* protocol[57] has the three-fold purpose of separating the 2D projections of the particles in homogeneous conformational groups, align the particles along a same axis and to eliminate bad/junk particles in the data set. The protocol was performed in several steps in order to separate the monomers from the dimers and discard noise (background, ice, low contrast particles) from good particles. At the end of each step particles were grouped as monomer side view (fig 2.4.a), dimer side view (fig 2.4.b), top view(fig 2.4.c) and noise (fig 2.4.d), each group was again subjected to the 2D classification and divided in the previous group. This operation was repeated iteratively until any change in the groups it was no longer observed. Finally, the top view particles were added to the monomer and dimer side view and at the end of this procedure 20218 particles, divided into 89 classes, that correspond to the top and side view of the dimeric conformation, were selected.

As for the the preprocess protocol, the particle mask diameter is set to 140 pixels to be sure not to cut-off any real signal and, at the same time, try to minimize the noise around the particles. The most important parameters to take care of in the *2D classification* protocol are the number of classes (K) and the regulation parameter T (Appendix B).

For any classification step K was set as to have approximately 100-200 particles per class but never a values greater than 200, due to the excessive execution time that such a value of K would have required. So, for example, the first classification step with 29748 particles were performed with K=200 and an intermediate step with 4589 particles were performed with K=45.

The regulation parameter T takes account of the correlations between Fourier components in the signal. In agreement with the literature[25] a value of T=2 was adopted.

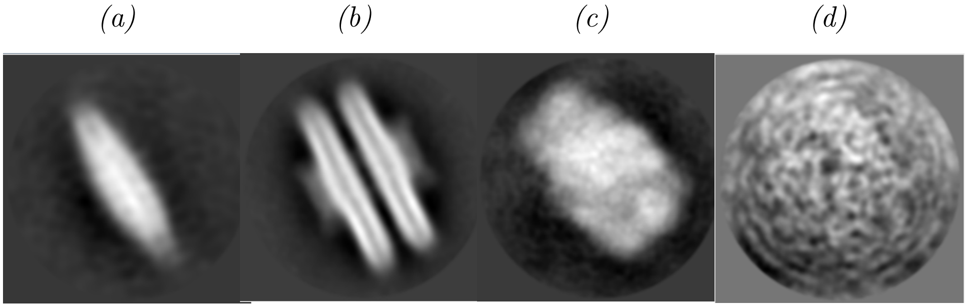


Figure 2.4: Average of the classes, evaluated with the RELION 2D classification protocol, of a particular orientation for: (a) monomer side view, (b) dimer side view, (c) top view, (d) noise particles. The particles are enclosed in a circle of 140 pixels (196 Å) radius, while the side of the box is 300 pixels (420 Å).

2.2.4 Initial volume

Because there was not an available map to use as a template and the number of particles was not sufficient to calculate the reference map from a set of independent data, the averages of the 89 classes were used to create the initial reference model for the 3D classification. The initial map was evaluated with the protocol *EMAN2 initial model* [53] that treats the averages of the classes found in the previous steps as particles to refine a randomized initial blob. This operation is carried out iteratively for a number of cycles set by the operator (in our case 15, to have a good relationship between the time taken by the process and the quality of the final model) and each step gives a series of increasingly refined maps.

The protocol *EMAN2 initial model* gives the opportunity to impose a defined symmetry to the map, however, it was not imposed any symmetry to be sure that the initial model was as faithful as possible to the data. Many structures, and ours is no exception, have a number of local minima in the energy space that generates a number of incorrect structures. To figure out which map to use as initial model, the 2D projections of each map generated with *EMAN2-initial model* were compared with a set of the experimental particles. Among all the possible solutions provided by the software, the one with the best agreement with the data and simultaneously the closest to the density maps already published [38] was chosen.

The initial model was low pas filtered to 60 Å to avoid any errors propagation in the following 3D classification.

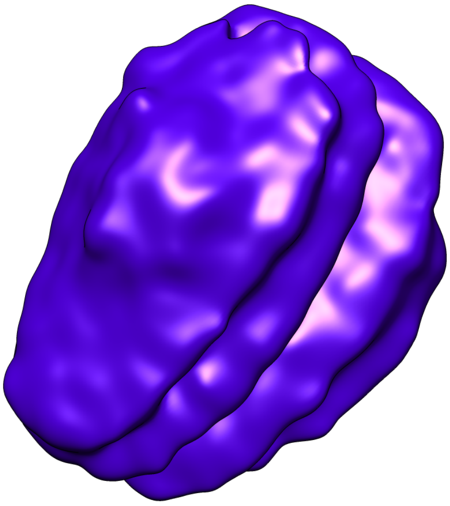


Figure 2.5: Map of the initial model produced by EMAN2 with a resolution of 60 Å.

2.2.5 3D classification

The protocol *RELION 3D classification* has the purpose to detect possible structural heterogeneity. To do so the protocol iteratively repeats two steps: alignment and maximization.

In the Alignment step, computer-generated projections of the structure are compared with the experimental images, resulting in information about the relative orientations of the images. In the second step the experimental images are combined with the prior information into a smooth 3D reconstruction[58].

As for the 2D classification, the most important parameters to take care are the number of classes (K) and the regulation parameter T. In 3D classification each classes must to have at least 5000÷10000 particles, so the 20218 were divided into 3 classes (K=3) with T=4 as suggested by the literature [54]. 3D classification was performed in several steps increasing the accuracy of the research angle from one step to another to be sure to take into account also small conformational differences. Even in 3D classification, to be sure not to introduce artifacts, no symmetry was imposed.

The 3D classification gives 3 different classes of 10497, 6104 and 3617. They represent respectively the conformational state C_2S_2M (fig 2.7.a), C_2S_2(fig 2.7.b) and a class composed by noisy particles(fig 2.7.c).

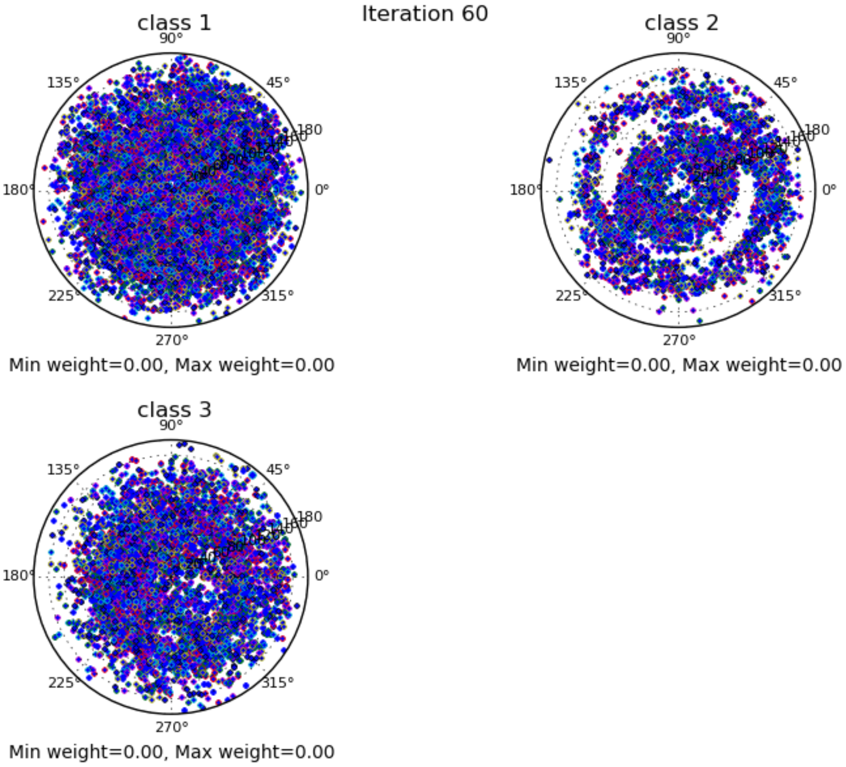


Figure 2.6: Angular distribution of the three class C_2S_2M (class 1), C_2S_2(class 3) and noise (class 2). It is possible to notice that the first class has the more homogeneous angolar distribution while the other two, while the other two, and in particular the class 2, do not cover most of the projections directions.

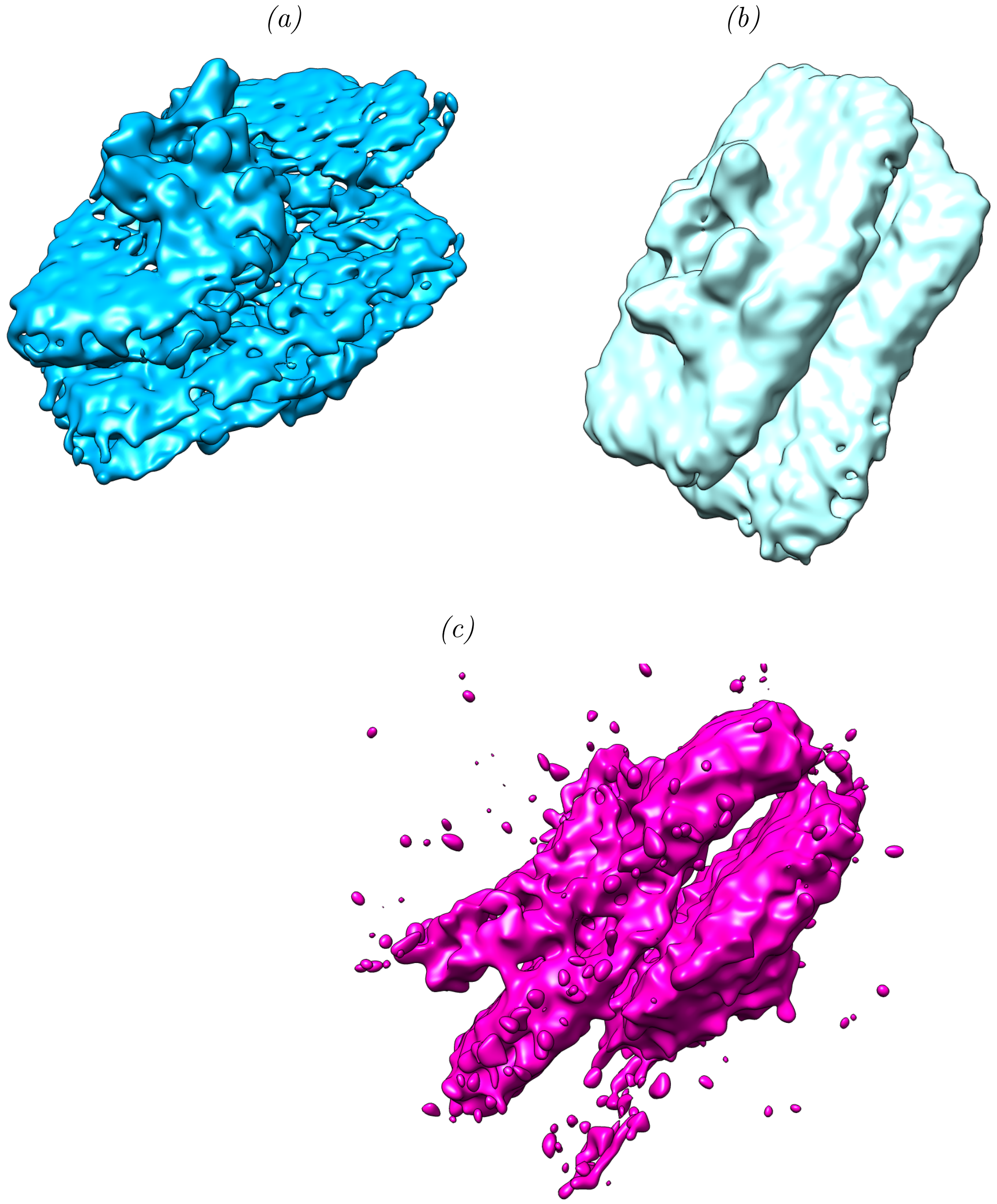


Figure 2.7: The map of C_2S_2M (a), C_2S_2(b) and a class composed by noisy particles(c).

2.2.6 Refining

The selected C2S2M map was refined using the protocol *RELION 3D auto-refine*. This protocol splits the data into two halves and refines independent reconstructions against each half-set. The Fourier shell correlation (FSC) between the two independent reconstructions then yields a reliable resolution estimate, so that the iterative build up of noise can be prevented[58, 25] giving a refined map with 16 Å resolution.

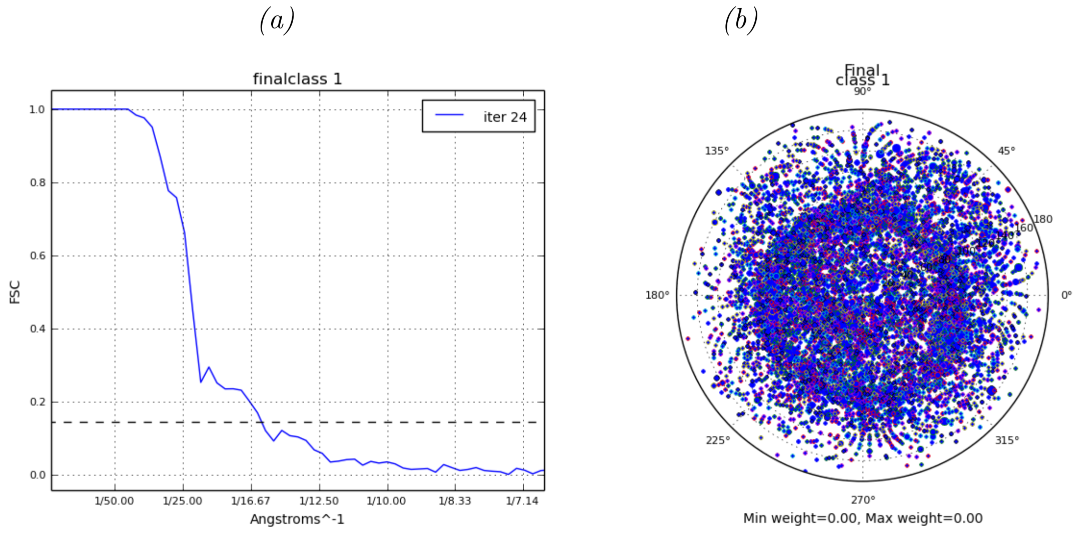


Figure 2.8: The FSC (a) intercepted at value 0.143 and the angular distribution of the refined map(b)

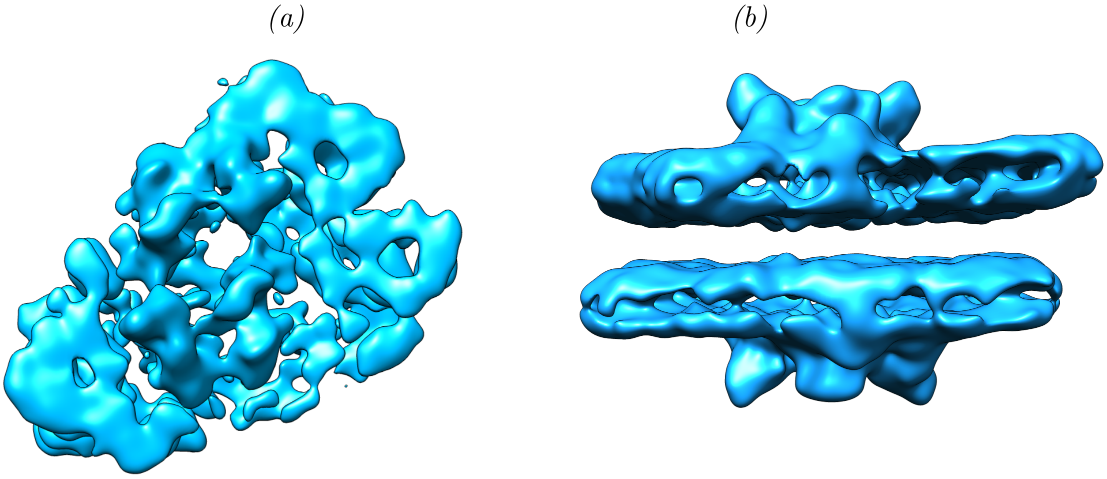


Figure 2.9: Top (a) and side view (b) of the refined map at the angular distribution of the refined map with a threshold of 0.14

2.2.7 Post processing

The refining step could underestimate the real resolution of the reconstructed map because, to prevent overfitting, no masks or filters have been applied[25].
Therefore, to determine the real resolution a tailored soft mask was applied to the map. The use of an optimal mask is important, as it serves to select the region of interest in the average image of all particles to reduce the influence of the background noise[59]. The protocol *RELION post-process* was used to calculate the FSC between the two masked map. To measure the inflating effect that the mask may have on the FSC curves the high-resolution noise substitution method was use. This method is based on the following steps:

1. randomize the phases of the two unfiltered maps before masking;
2. mask these two scrambled maps with the same mask
3. calculate a FSC curve between them;
4. assess the masking-effects on the FSC by analysing the resolution-dependent difference between the masked and unmasked FSC curves
5. output a masking-effect-corrected FSC curve (called rlnFourierShellCorrelationCorrected)

The post processing gives the final 10.5 Åresolution map.

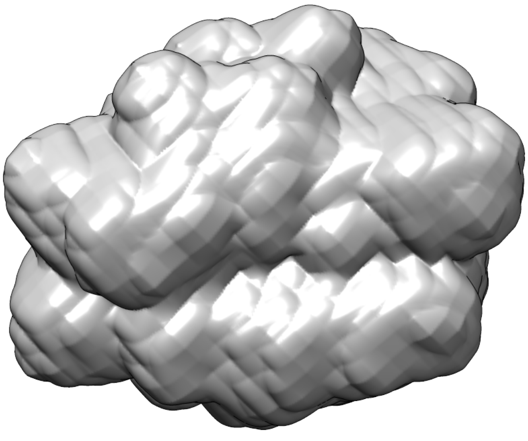


Figure 2.10: The soft mask used in the post processing protocol.

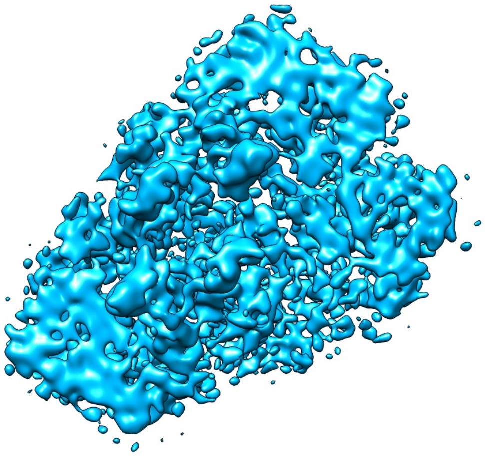


Figure 2.11: Top of the 10.5 Å postprocessed map with a threshold of 0.055.

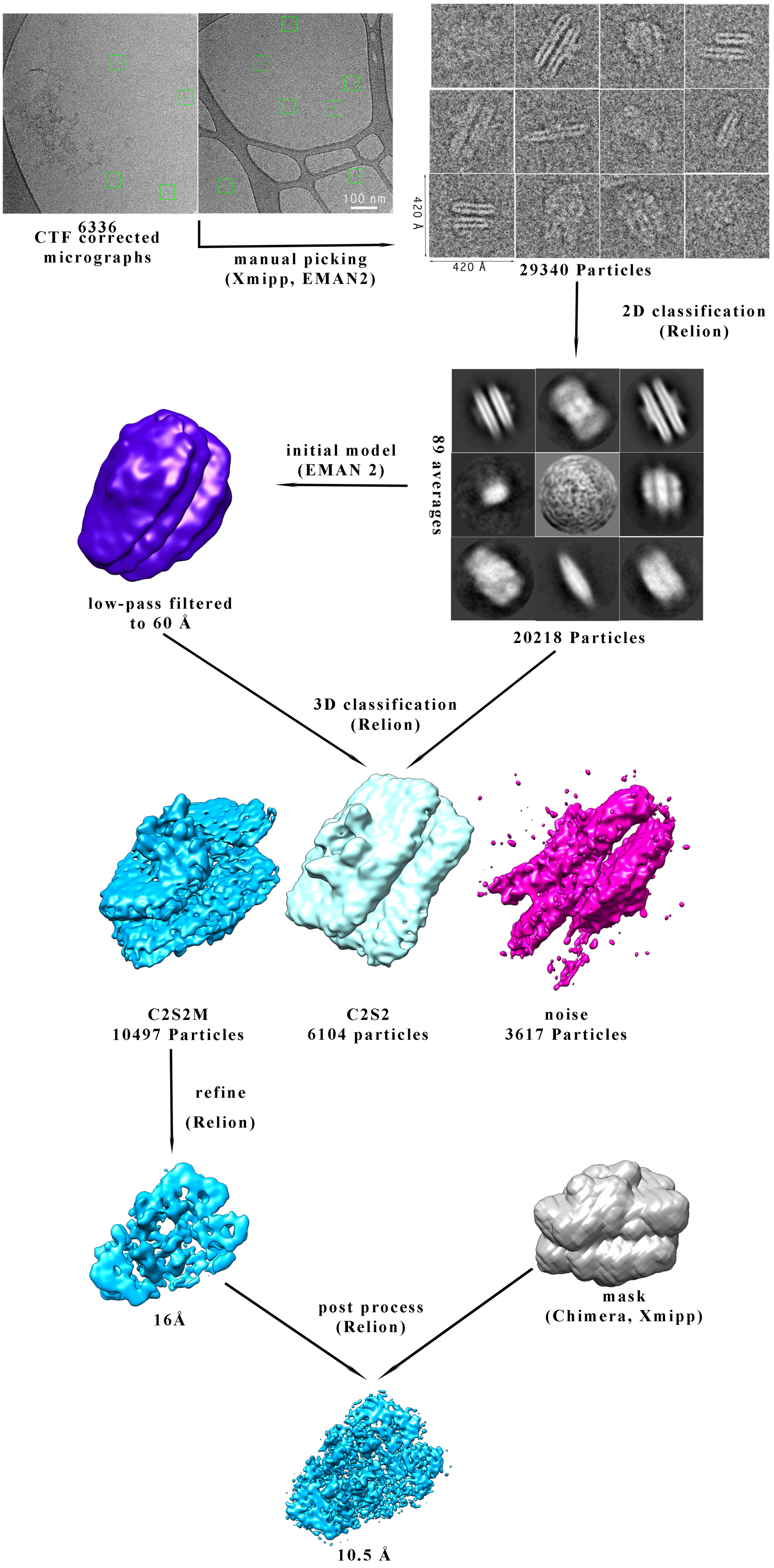


Figure 2.12: The 3D map reconstruction workflow followed in th with the software used in bracket, each step also indicates the number of micrographs or particles that take part in the process.

2.3 Structure reconstruction and refinement

Model building was performed only in one of the two monomers of the post processed map, where the PDB structures of the proteins that compose the supercomplex PSII-LHCII were fitted using UCSF Chimera[60] (fig 2.13).
The proteins structure derive from:

- **LHCII** from the crystallographic structure of pea LHCII (PDB code 2BHW)[36].
- **PSII, lhcb4 and lhcb5** were extracted with PyMOLE (www.pymol.org) from the spinach C2S2 PSII-LHCII supercomplex structure (PDB code 3JCU)[41].
- **lhcb6** was modeled with SWISS-MODEL[61] because at the moment there is no crystal structures available.

The model of the protein CP24 was separately refined against his density map, extracted with Chimera, in the reciprocal space, while the global model of C2S2M supercomplex was refined in real space by Phenix[62].

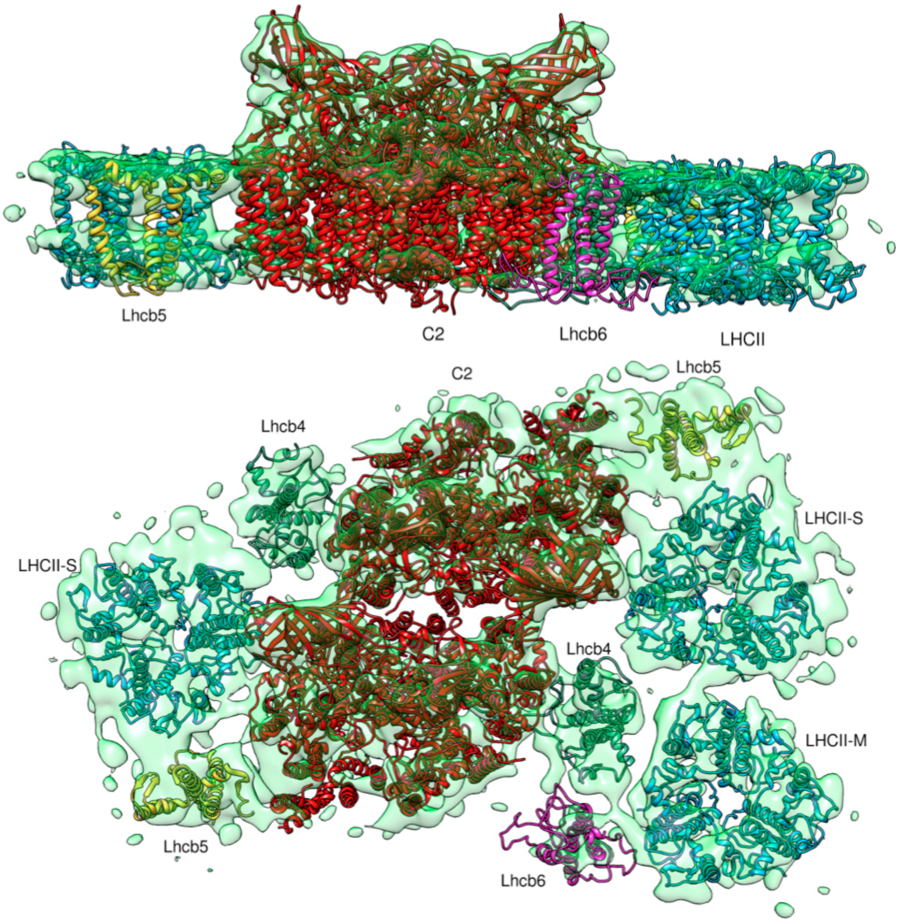


Figure 2.13: top and side view of one of the two monomer composing the C2S2M (transparent light green): PSII (part of 3JCU) in red, LHCII (2BHW) in light blue, lhcb5 (part of 3JCU) in yellow, lhcb4 (part of 3JCU) in green and lhcb6 in purple.

2.3.1 Reciprocal space refinement

The initial CP24 structure model was created with SWISS-MODEL, an automated protein structure homology-modelling server. A sequence of 206 amino acids of CP24 protein from *Aradinobsis tantalia* was used as target for the modeling and, from the model given, the one with the best sequence identity (42.62%) was chosen (fig 2.13.a). The obtained first model was docked in the post processed map (Section 2.2.7) using UCSF Chimera and the section of the map corresponding to the location of the CP24 protein was extracted (fig 2.14 b). To proceed with the reciprocal space refinement the extracted map was converted into structure factors with the Phenix *phenix.map_to_structure_factors* command that performs the Fourier transform of the map (Appendix C). Differently from structure factors obtained through X-ray diffraction, those calculated have both the module and the phase, moreover there are no limits to the resolution. In addiction the origin ($\mathbf{F}(0,0,0)$) can be calculate. However, the resolution should be limited in agreement to that obtained at the end of the reconstruction process. For our map the resolution limit was set to 12 Å in agreement with the one given by the post process protocol (fig 2.15).
Because of the limited resolution of the map, the PDB structure of CP24 was only subjected to a rigid body refinement on the α-helices (residue 90-125, 141-167, 214-250, 252-256), while the loops were left free to move. After every refinement with Phenix, the model has been adapted manually with *Coot* to improve the local agreement of the amino acids with the map and reduce Ramachandran, C-beta and rotamers outliers.
The refinement with Phenix and following adaptation with *Coot* has been repeated until a minimum of crystallographic R factor was reached.

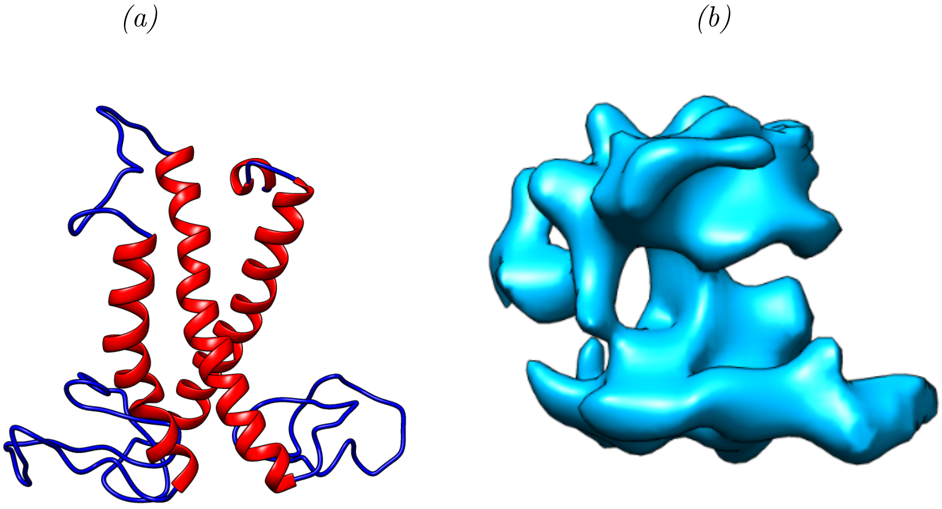


Figure 2.14: The CP24 starting model with the fixed residues (α-helices) in red and the free to move loops in blue (a) and the the correspondence electron density map(b).

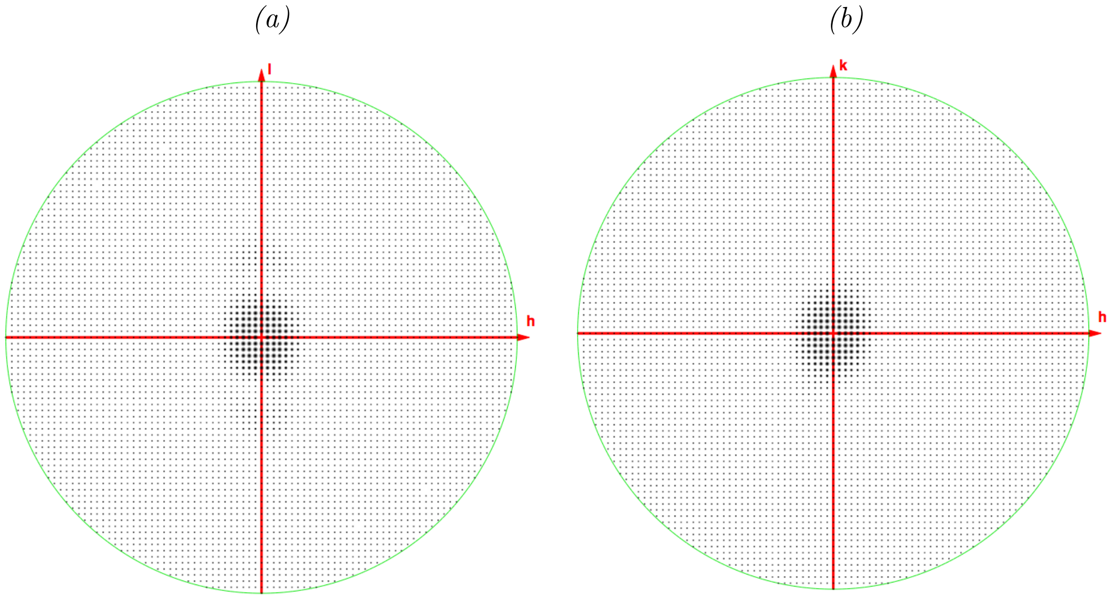


Figure 2.15: The reciprocal space h0l (a) and 0kl (b) generated by Phenix from the CP24 electron density map.

2.3.2 Real space refinement

The Phenix real space refinement program refines a model into an electron microscopy map to obtain a model that fits the map as good as possible posessing a meaningful geometry (no validation outliers, such as Ramachandran plot or rotamer outliers).

Each refinement cycle is composed by the following step[63]:

1. **Rigid body refinement**:Real space analogue of automatic multizone rigid body refinement with large convergence radius.
2. **Model idealization**: i.e. a model repacking to match the secondary structure of the model with minimal deviation from the starting model.
3. **Morphing**: a large rigid body translational shift to match the map.
4. **Weight calculation**: a fast algorithm of optimal refinement target weight calculation to guarantee best model to map fit and requested deviations from ideal stereochemistry.
5. **Minimization**: a gradient driven minimization of combined map with the restrains target (secondary structures, Ramachandran plot, CB deviation and rotamer restraints).
6. **Refine Non Crystallographic Symmetry (NCS) operators**: choose the NCS that best fits the map.
7. **Simulated Annealing**: real space simulated annealing refinement with slow cooling protocol starting at 5000 K.

8. **Rotamer fitting**: a second fit by rotamer enumeration that chooses a valid rotamer that fits map best and restrain to its current conformation.

The minimization step also produces in output the refined model and the refinement trajectory file that could be used to monitor how model changes in each cycle.

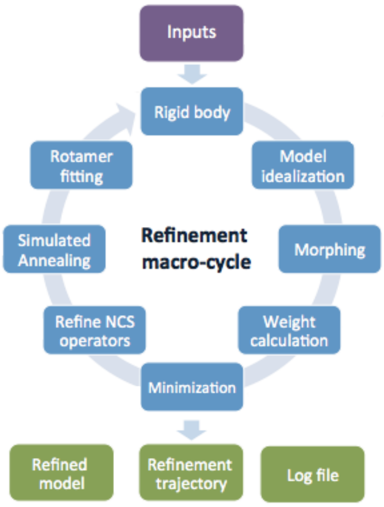


Figure 2.16: Real space refinement macro cycle[63]

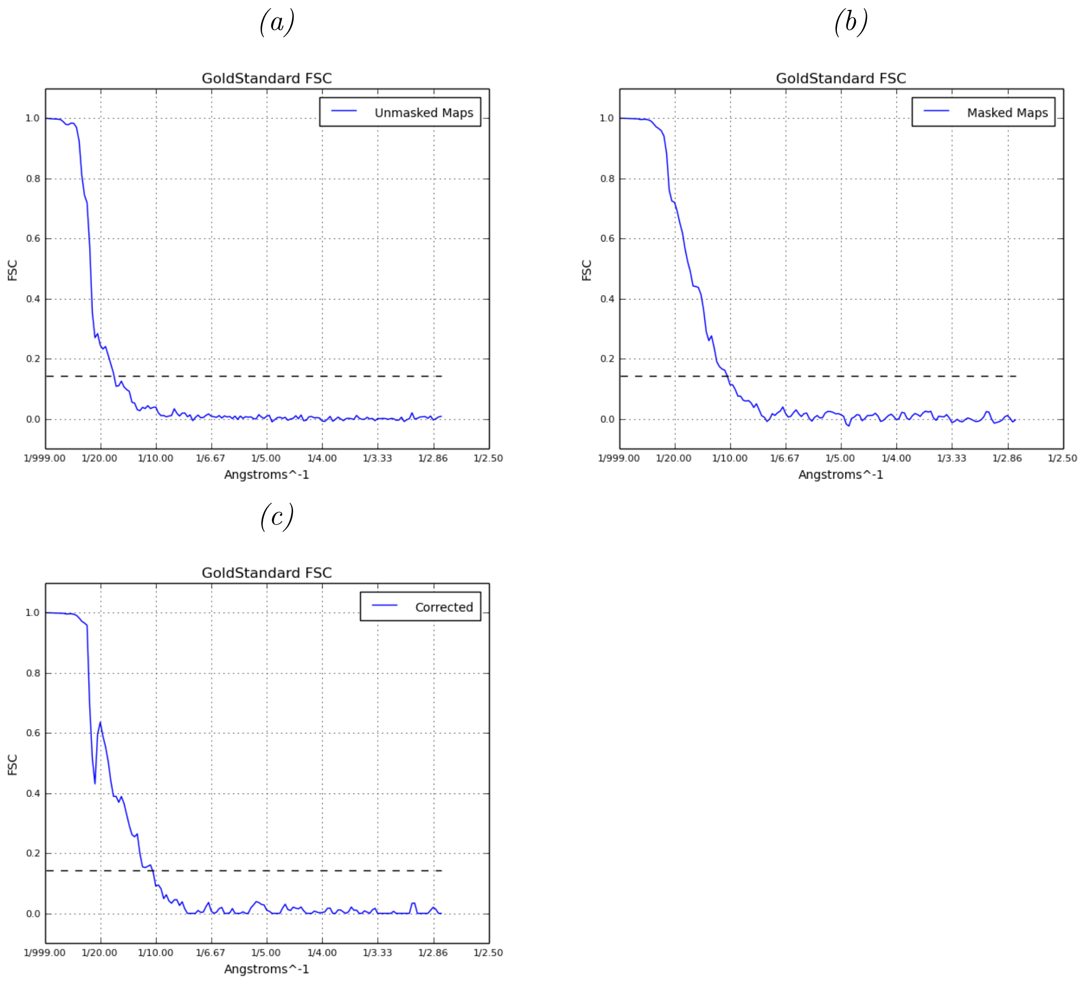
The structure of protein CP24 refined in the reciprocal space has been added to the PSII-LHCII structure that was subsequently refined in real space. The PSII-LHCII supercomplex model was run for 5 cycles with a rigid body refinement in which each chain forms an independent body.

3 | Results and discussion

3.1 Electron density map

3.1.1 Final resolution

The final electron density map was obtained with the *RELION postprocess* protocol, masking the refined map. In figure 3.1 a and b it is possible to see how the masking procedure improves the resolution of the reconstructed map evaluated with the gold standard Fourier Shell Correlation (FSC) method with the 0.143 cutoff (Appendix B). Despite the better resolution, the masked map can not be regarded as the final map because the mask used may introduce some bayas that, in turn, can lead to overfitting or errors in the determination of the model. The final map, called *corrected map*, is therefore given by a combination between the masked and unmasked map. The corrected map obtained at the end of the *RELION postprocess* protocol has a global resolution of 10.5 Å (fig 3.1.c).


(a)

(b)

(c)

Figure 3.1: The gold standard FSC function of unmasked (a), masked (b) map of the corrected electron density map that gives the final global resolution of the model.

The gold standar FSC resolution gives a global estimation of the goodness of the electron density map reconstruction, but is not so relevant in therm of real resolution because unresolved structural heterogeneity or limited representational accuracies may locally blur the maps. The local resolution of the map was evaluated using ResMap (fig. 3.2).

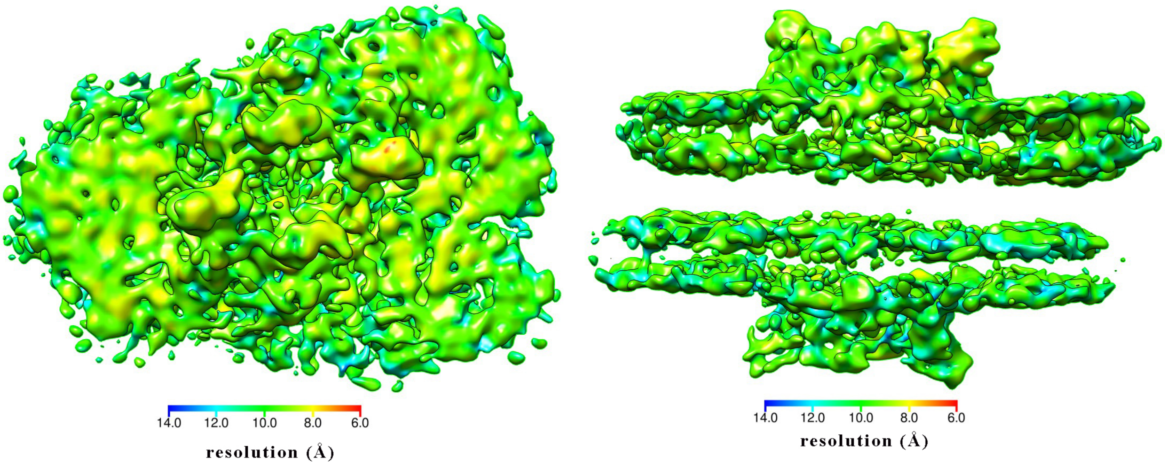


Figure 3.2: side and top view of the local resolution evaluated with ResMap

It is possible to notice that the local resolutions are slightly worse in peripheral areas and especially in the area of the LHCII-M trimer.However local resolution is quite homogeneous and there are not areas where the local resolution differs much from the one given by the FSC curves.

This resolution value depends to the low number of particles used and to the heterogenity of the sample. Morover, the fact that the micrographs used has been recorded with a single frame, instead of multiple ones (movie), makes impossible to perform a correction of the movements of the samples induced by the electrons beam[8, 9].

3.1.2 Final map

Here it is shown the final electron density map of the C2S2M dimer. Dimension were measured with Chimera and are to be considered ± 10.5 Å.

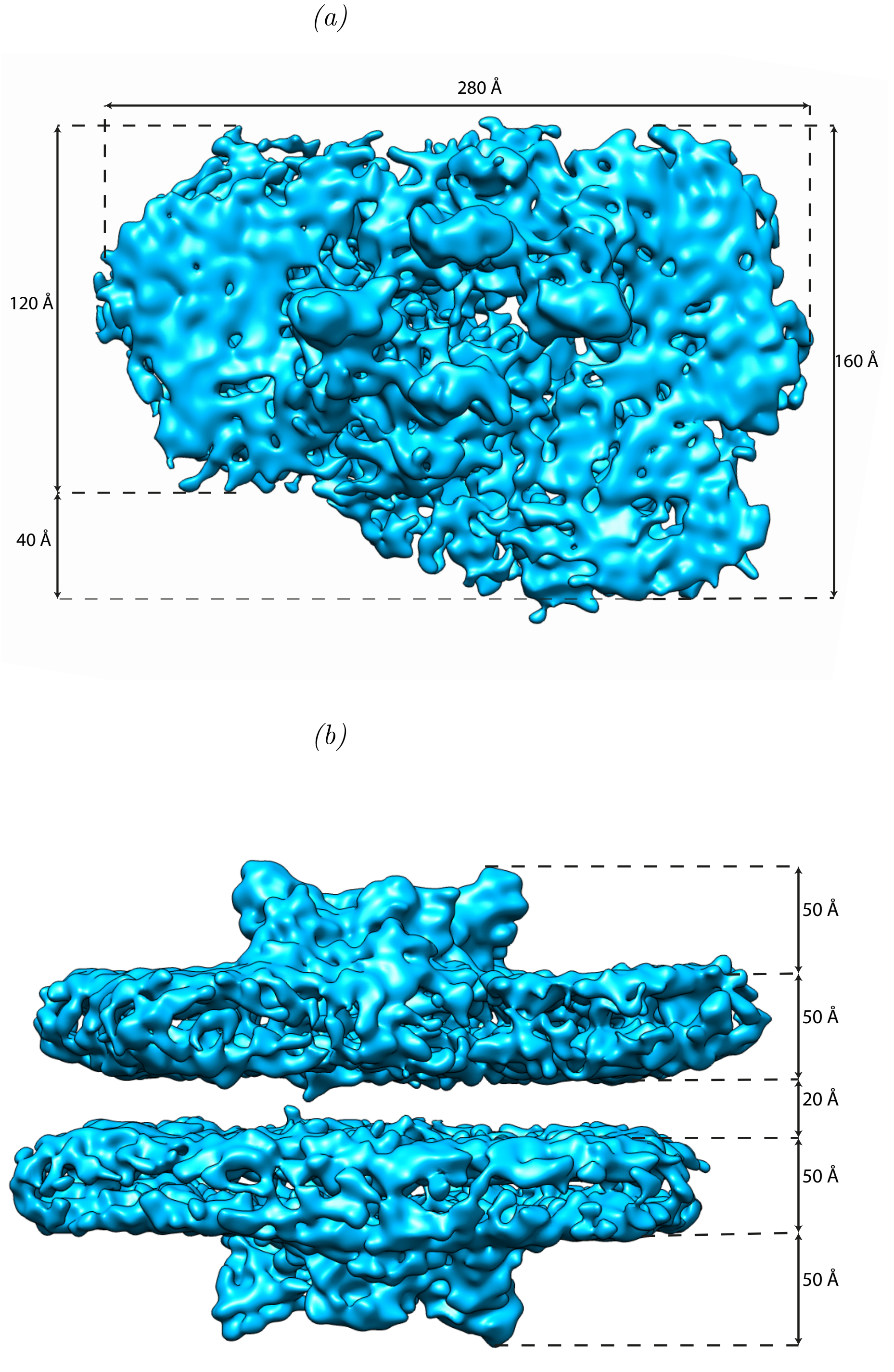


Figure 3.3: Top (a) and side view of the final 10.5 Å electron density map.

The two monomer are not perfectly parallel but are inclined by an angle of 30°as shown in figure 3.4.

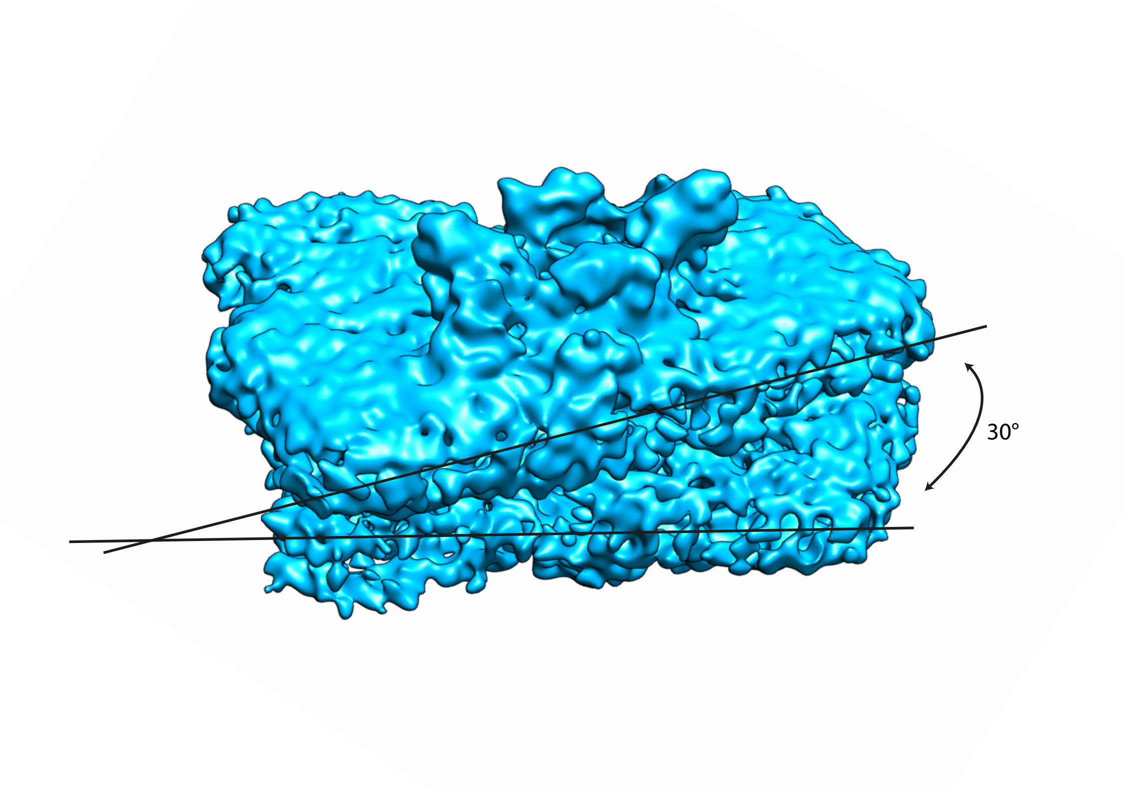


Figure 3.4: The angle between the two monomer

Although no symmetry was imposed during the reconstruction, the complex formed by the two C2S2M monomers has an axial symmetry with the axis of symmetry parallel to the membrane of thylakoids.
Moreover, In agreement with the electron density map resolution it was not only possible to identify the sub units that make up the C2S2M dimer, but also to provide hypotheses on how they are oriented in it, since it was possible to identify the α-helices position (commonly visible for resolution lower than 10 Å).

3.1.3 Interactions between the two monomer

Reducing the threshold of representation to 0.025, the connecting structures between the two monomers are visible; Their position in the map and their length of approximately 20 Å let us speculate that they are formed by the long loop of proteins CP29 (fig. 3.3 and 3.4).

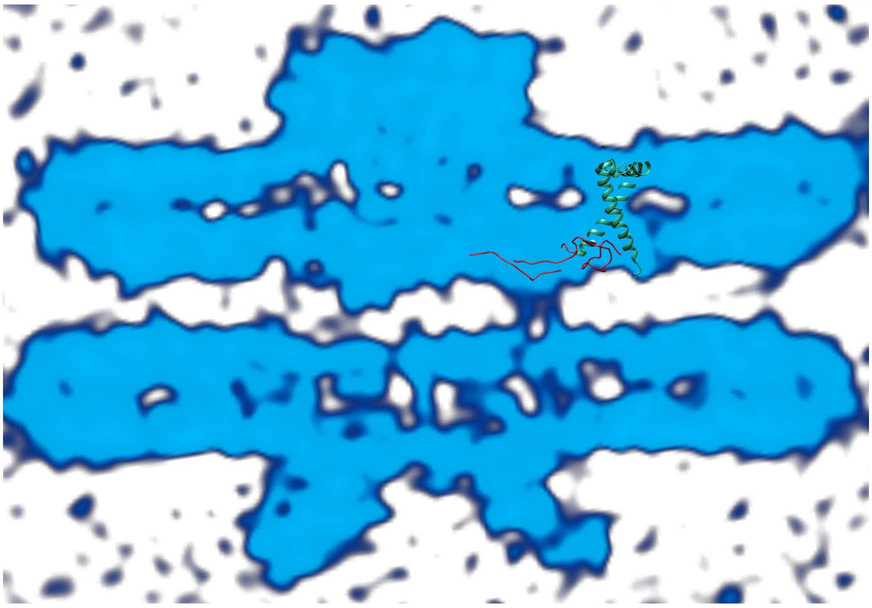


Figure 3.5: Section of the 10.5 Å map at plane 138 along the X axis with a threshold of 0.025. Protein CP29 is represented in green with the loop that is supposed to form the connection between the two monomers marked in red.

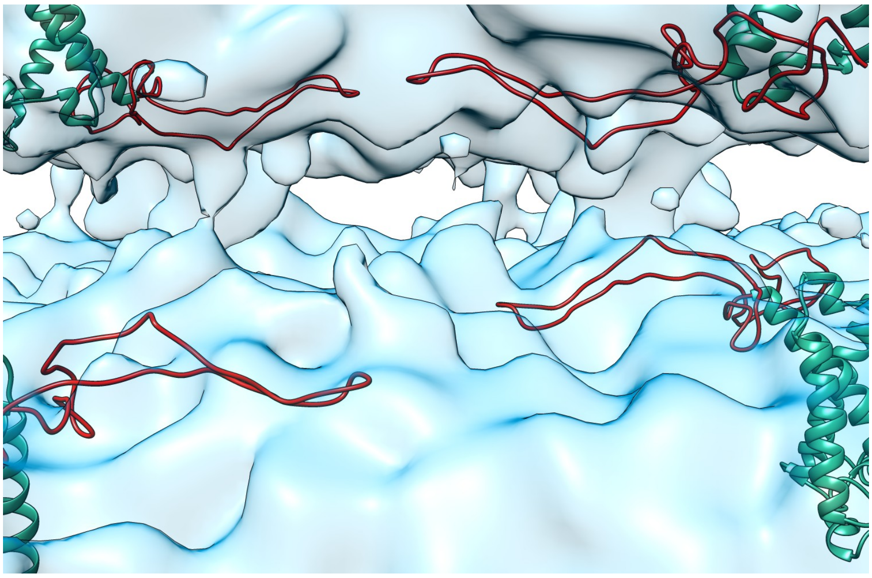


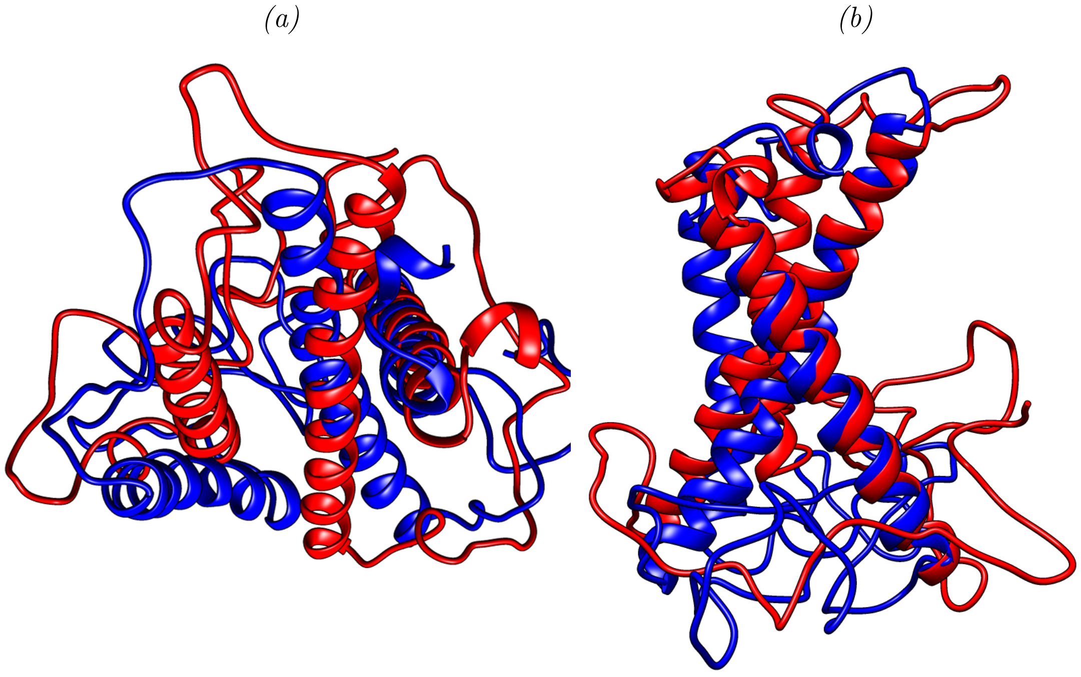
Figure 3.6: 3D view of the connections between the two monomers with a threshold of 0.025, proteins CP29 are represented in green with the loop in red.

3.2 Proteins structure

3.2.1 CP24

The final step of the CP24 refinement end with a R factor(appendix C) of 0.39. This value is higher than that the one of a refined protein that is typically less than 0.2. Moreover it was not possible to evaluate the R-free parameter (i.e. the R factor calculated on some data not included in the refinement) because their percentage in the calculated structure factor was to low to be statistically significant. However the mean R factors of the few structures deposited in the Protein data bank with a similar low resolution cover a range between 0.25 and 0.45.
The table below shows the values obtained at the end of the model refinement:

Resolution	420-12 Å
number of reflections	153449
R factor	0.39
RMS bonds	0.01 Å
RMS angles	1.46 °
Clashscore	30.3
Ramachandran favored	86.27 %
Ramachandran outliers	3.92%
Rotamer outliers	3.16 %
C-beta deviations	0



(c)

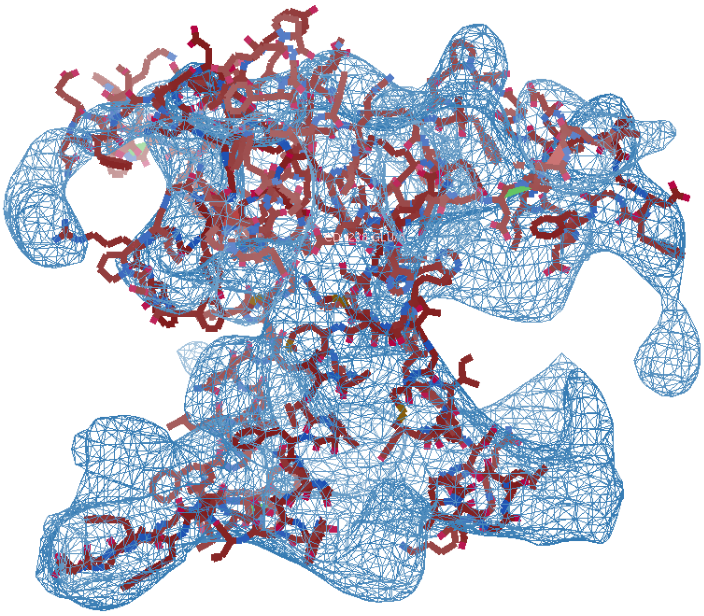


Figure 3.7: Top(a) and lateral(b) cartoon view of the CP24 protein before(blue) and after(red) the refinement.(c) the refined CP24 protein fittet in the extracted map (this figure-has been produced using Coot.

In figure 3.7 it is possible to see how the α-helices have been moved to best fit the map while the loops move much from their initial position. Unfortunately, due to the low resolution of the map the only reliable information is about the position of the α-helices because the resolution is to low to determine reliably that of loops.

3.2.2 PSII-LHCII

The refinement of the PSII-LHCII supercomplex structure, performed in the real space, gives a final cross correlation between the structure and the electron density map of 0.68, which is in agreement with typical CC values of a well refined map with Phenix[63].
The table below shows the values obtained at the end of the model refinement:

Resolution	12 Å
Map CC	0.68
RMS bonds	0.02 Å
RMS angles	1.59 °
Clashscore	12.26
Ramachandran favored	95.76 %
Ramachandran allowed	3.45 %
Ramachandran outliers	0.76 %
Rotamer outliers	5.06 %
C-beta deviations	22

The refined PSII-LHCII structure gives good information about the organization of the subunit in the complex especially as it regards the position of the LHCII-M trimer and of the lhcb6 subunit (protein CP24).

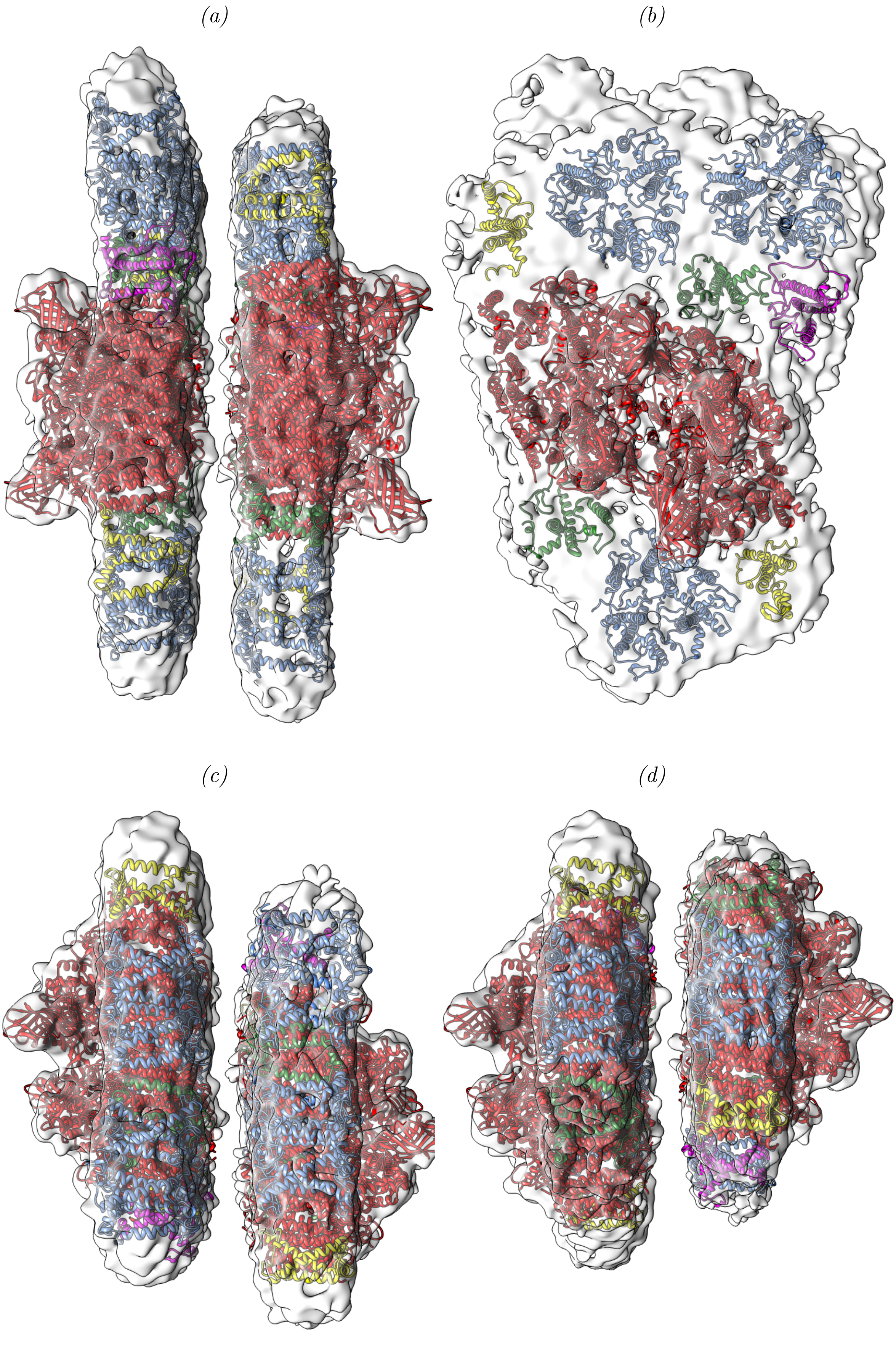


Figure 3.8: (a)Side view of C2S2M dimer supercomplexes. (b), corresponding top view with a rotation of 90°along the membrane plane. (c-d), corresponding front views with a rotation of 90°to the right (c) and to the left (d) along the normal to the membrane plane. PSII (part of 3JCU) is in red, LHCII (2BHW) is in light blue, lhcb5 (part of 3JCU) is in yellow, lhcb4 (part of 3JCU) is in green and lhcb6 is in purple.

3.3 Comparison with literature data

3.3.1 Map comparison

The recent publication of the 3.2 Å structure of spinach PSII-LHCII supercomplex[41] allowed us to compare it with our results.

At first the 3.2 Å resolution map deposited in the EMDataBank (EMD code 6617) was fitted in the post processed electron density map with Chimera. The fit gives a correlation of 0.92 between the two models that indicate a very good agreement between the two PSII-LHCII supercomplexes, taking into account the presence of the trimer LHCII-M in our map(fig 3.9).

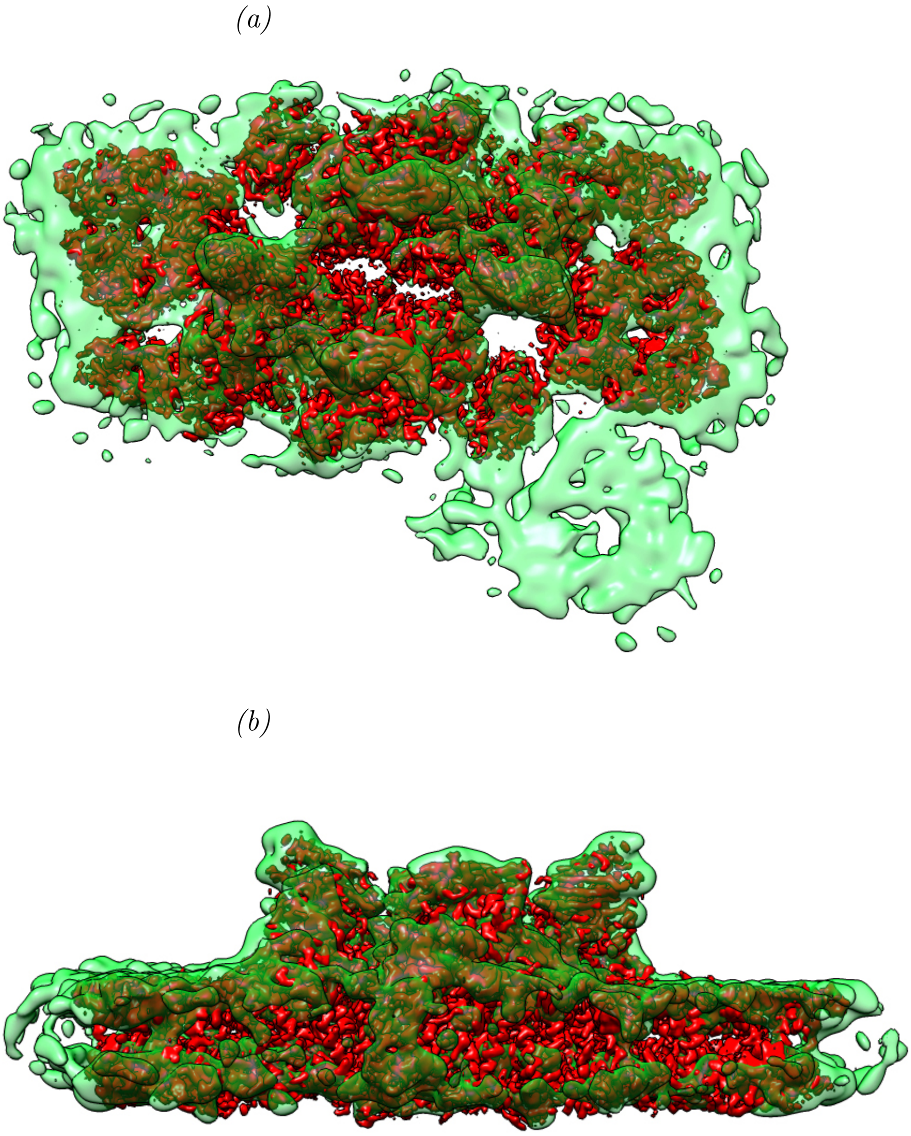


Figure 3.9: Top and side view of the fit of the 3.2 Å map (EMD code 6617) in red, in the 10.5 Åcalculated map in light green

3.3.2 Structure comparison

The comparison between the refined PSII-LHCII structure and the one deposited in the Protein Data Bank (PDB code 3JCU) shows (in addition to the obvious difference given by the presence of the LHCII-M trimer and lhcb6 subunit) a displacement of the peripheral units (fig 3.10).
To evaluate this displacement the distance between the centers of mass of the perimeter subunit of the two model was measured with USCF Chimera. An average shift of 8 Å for the external subunits LHCII-S and Lhcb5 and an average shift of 5 Å for the inner subunit Lhcb4 is measured.

id	subunit	Distance (Å)
1	PSII	0.4
2	LHCII-S	8.3
3	LHCII-S	8.4
4	Lhcb5	7.8
5	Lhcb5	8.3
6	Lhcb4	5.2
7	Lhcb4	4.3

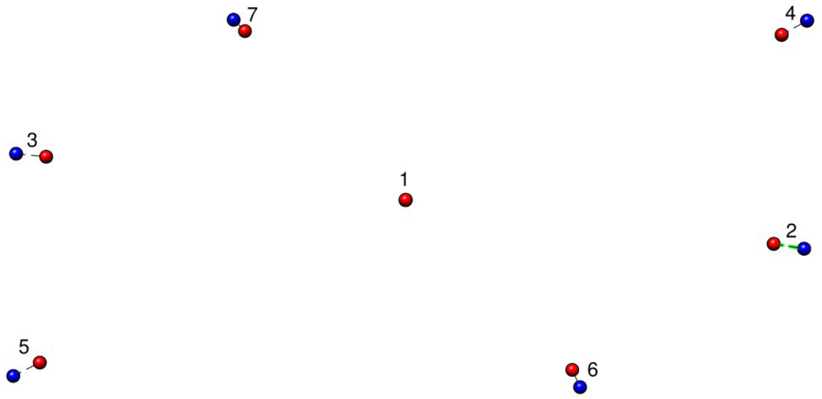


Figure 3.10: Distance between the center of mass of subunits of the deposited (red) and the calculated map(blue).

The outward displacement of the subunits belonging to the calculated map may be due to the different conformational state or to the mobility of these structures.

3.4 Conclusions and future perspectives

Conclusions:

The electron density map obtained at the end of the reconstruction process has a global resolution of 10.5 Å, evaluated with the gold standard Fourier shell correlation method. In parallel, the local resolution is quite homogeneous and there are not areas where the local resolution differs much from the one given by the FSC curves. This resolution value depends on the relative low number of particles used and to the heterogenity of the sample. Moreover the fact that the micrographs used has been recorded with a single frame instead of multiple ones (movie), makes impossible to perform a correction on the movements of the samples induced by the beam. In agreement with the electron density map resolution, it was possible to determine the subunits that compose the C2S2M dimer and allowed us make hypotheses on how they are oriented, since it was possible to identify the α-helices (commonly visible for resolution less than 10 Å) positions in areas of the map with higher resolution.
It was also possible to formulate a hypothesis about the unresolved structure of the lhcb6 subunit (protein CP24) identifying and refining the position of its α-helices and to speculate about the loops organization.
Moreover, fitting the refined structure of CP24 and the resolved structure of the PSII-LHCII subunits in the electron density map it was possible to refine a final structure of the entire PSII-LHCII supercomplex.
The reconstructed maps show also four 20 Å connecting structure between the two monomers. Their position and their dimension let us speculate that they could be formed by the long loop of proteins CP29. However this last statement needs further studies to be confirmed. In any case this thesis shows the ability of cryo-EM single particles techniques to reconstruct the structure of proteins complexes from heterogeneous data.

Future perspectives:

With regard to the structure of the PSII-LHCII dimer, further studies can be conducted in order to identify and characterize the connection structure between the two monomers. To do that, in parallel with fluorescence studies underway at the laboratory of Dr. Pagliano, a molecular dynamics of the complex nestled in the thylakoids membrane is running.
More generally, the techniques learned in this thesis can be applied to other proteins that are difficult to crystallize or resolve with conventional techniques. Alternatively, the cryo-EM allows to obtain electron density maps of multiple protein complex structures in which the individual protein structures obtained by crystallography can be fitted.
Furthermore, the cryo-EM single particles analysis field is in continuous development both for the achievable resolution and for the development of programs able to resolve more and more complicated structures with best efficiency.

Appendices

A | Image formation

The electron microscope aberrations distort the structural information by chaining phases and amplitudes of the recorded electron waves. These distortions can be modeled in Fourier space by the microscope Contrast Transfer Function (CTF).

A.1 Image distortion

Considering spherical aberration, defocus and twofold astigmatism, the total aberration function is[64]

$$\chi(q,\beta) = \frac{2\pi}{\lambda}\left(\frac{1}{4}C_s\lambda^4 q^4 - \frac{1}{2}\Delta f(\beta)\lambda^2 q^2\right) \tag{A.1}$$

where q is the magnitude of the spatial frequency (q_x,q_y), λ is the electrons wavelength, C_s is the spherical aberration coefficient and Δf is the defocus.
From the previous equation it is possible to define the transfer function of the lenses system

$$T(q,\beta) = e^{-i\chi(q,\beta)} \tag{A.2}$$

and the Fourier transform of the electron wave at the back focal plain

$$\tilde{\Psi}(q,\beta) = FT[e^{i\sigma v_z(x,y)}]T(q,\beta) \tag{A.3}$$

where $\sigma = \frac{\lambda m e}{2\pi\hbar^2}$ is the interaction constant and $v_z = \int V(x,y,z)dz$ describes the projection of the potential scattering (**V**) along the z axis i.e. the direction of the incident electrons. Therefore the image intensity is

$$I(x,y) = |\Psi(x,y)|^2 \tag{A.4}$$

were $\Psi(x,y) = FT^{-1}[\tilde{\Psi}(q,\beta)]$

Moreover the temporal and spatial incoherence introduced by the energy spread and the finite source size can be modeled as a damping envelops in the spatial frequency domain. The temporal incoherence can be modeled as

$$K_c = exp\left[-\left(\frac{\pi\lambda q^2 C_c H}{4\sqrt{ln 2}}\right)^2\right] \tag{A.5}$$

where $H = \frac{\Delta E}{E}$ is the ratio between the energy of the incident electron E and its spread andC_c is the chromatic aberration coefficient.
Similarly the spatial incoherence introduced by the finite source size is

$$K_s = exp\left[-\frac{(\pi C_s \lambda^2 q^3 - \pi \Delta f(\beta) q)^2 \beta_i^2}{ln2}\right] \tag{A.6}$$

The global incoherence of the source can be summarized as

$$K(q, \beta) = K_s(q, \alpha) K_c(q). \tag{A.7}$$

lastly the measurement process produces the final Image intensity

$$I(x, y) = [C \cdot N_p(\Phi_e \cdot I_0(x, y))] \otimes PSF(x, y) + I_{rn} + I_{dc} \tag{A.8}$$

where C is the conversion factor of the camera, N_p is the Poisson noise in function of the incident flux of electrons $\Phi_e \cdot I_0(x, y)$ all convoluted ($\otimes$) whit the Point Spread Function (PSF) of the detector that describes the blurring of the image. I_{rn} and I_{dc} are respectively the readout and the integrated dark current.

A.2 Contrast transfer function

In order to estimate the CTF parameters it is assumed that there is a significant overlap of atomic positions in a projection and that the projected sample has noise with a flat frequency spectrum. This is an approximation as every real specimen has limited scattering power. The mean inner potential of the sample introduces a constant phase change of the electron wave which can be neglected in this analysis as it is frequency independent. With these assumptions, the projected potential $v_z(x, y)$ is known and allows us to extract the CTF from the recorded image intensity.
The total intensity for a weak-phase, weak-amplitude object is given by the inverse Fourier transform

$$I_0(x, y) = FT^{-1}[\delta(q) + \sigma \tilde{V_z}(q) CTF(q, \beta)] \tag{A.9}$$

and

$$CTF(q, \beta) = 2A_p(q) K(q, \beta) sin(\chi(q, \beta) - \Phi_a(q)) \tag{A.10}$$

where A_p takes count of the influence of the objective aperture, $\tilde{V} = FT[V_z(x, y)]$ and $\Phi_a(q) = arcsin(W(q))$ with $W(q)$ the amount of amplitude contrast[65].
The CTF parameter is estimated by minimizing the discrepancy between the background subtracted Power Spectrum Densities (PSD) of simulated and measured projections[66] where the PSD is the Fourier transform of the image intensity:

$$PSD(q, \beta) = FT[I(x, y)] \tag{A.11}$$

The PSD of a recorded image shows a pattern called Thon rings[67] (fig 2.1) and their shape is circular if no astigmatism is present, although, with increasing astigmatism the shape gradually transit to elliptical.

The minima of the PSD correspond to zeros of equation A.1 and the PSD zeros (minimal observed contrast) correspond to CTF zeros, that, for the sine therm of CTF, are:

$$\chi(q,\beta) - \Phi_a(q) = k\pi, k \in \mathbb{Z} \tag{A.12}$$

and their location depends on the accelerating voltage, the defocus and the spherical abberation.

B | Particles analysis

B.1 Map recostruction

Almost all existing implementations for cryo-EM structure determination employ the weak-phase object approximation[68], which leads to a linear image formation model in Fourier space:

$$\tilde{X}_{ij} = CTF_ij \sum_{l=1}^{L} \mathbf{P}_{jl}^{\Phi} \tilde{V}_{Kl} + N_{ij} \tag{B.1}$$

where $\tilde{X}_{ij}$ is The j-th component of the 2D Fourier transform $\tilde{\mathbf{X}}_i$ of the i-th experimental image X, CTF_{ij} is the j th component of CTF of the i th image. $\widetilde{V_{Kl}}$ is the l-th component of the 3D Fourier transform $\tilde{\mathbf{V}}_K$ of the K the underlying structure in the data set. $K > 1$ describe a structural heterogeneity in the data[69, 20]. $\mathbf{P}^{\Phi}$ is a J×L matrix where Φ defines the orientation of the 2D Fourier transform with respect to the 3D structure and the operation $\sum_{l=i}^{L} \mathbf{P}_{jl}^{\Phi} \tilde{V}_{Kl}$ extracts 2D slice of the 3D Fourier transform of the K-th structure. N_{ij} is the noise in the complex plane. After the individual particles selection from the micrographs, the experimental observations comprise N images X_i. The estimation of $\tilde{V_K}$ from the images is done by an iterative procedure that requires an initial low-resolution 3D reference structure $\tilde{V}_K^{(0)}$. Every iteration (n) of the refinement process, is divided in two independent process. In the first, the projections of $V_K^{(n)}$ are calculated for different orientations Φ and compared with the experimental images. Then, each image is assigned to its optimal orientation Φ^* group. In the second all images are then combined into a 3D reconstruction that yields the updated model $\tilde{V}^{(n+1)}$. The update formula for V may then be given by:

$$\tilde{V}_l^{(n+1)} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \mathbf{P}_{ij}^{\Phi^{*}T} CTF_{ij} \tilde{X}_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{J} \mathbf{P}_{ij}^{\Phi^{*}T} CTF_{ij}^{2}} \tag{B.2}$$

This refinement is a local optimization procedure that could converge to a local minimum. Consequently, the initial reference model has an important effect on the outcome of the refinement, as strongly influence the refinement result[58]. Moreover the observed data alone are not sufficient to uniquely determine the correct solution. To have a more reliable solution a statistical approach has to be use[54]. This approach uses equation B1.1 formation model but assumes that all noise signals N_{ij} are independent and Gaussian distributed with variance σ_{ij}^2. The variance σ_{ij}^2 is unknown and will be estimated from the data and its variation with resolution allows the description of nonwhite or colored noise.

Considering the ensemble of possible solution, the reconstruction has the aims to find the model Θ that has the highest probability of being the corrected one in the light of both the observed data X and the prior information Y. This is a posterior distribution that can be factorizes as:

$$P(\Theta|X,Y) \propto P(X|\Theta,Y)P(\Theta|Y) \tag{B.3}$$

where the first term, called *likelihood*, expresses the probability of observing the data given the model, and the second, called *prior*, quantifies the probability to obtain the model given the prior information.
The assumption of independence in the noise allows the probability of observing an image given its orientation and the model to be calculated as a multiplication of Gaussians over all its Fourier components[70], so that:

$$P(\tilde{X}_i|\Phi,\Theta,Y) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^2} exp\left(-\frac{|\tilde{X}_{ij} - CTF_{ij}\sum_{l=1}^{L}\mathbf{P}_{jl}^{\Phi}\tilde{V}_l|^2}{2\sigma_{ij}^2}\right) \tag{B.4}$$

where it is assumed not to have model heterogeneity($K = 1$). Unlike random canonical tilt, in single particles analysis the orientations Φ of the images are not known. They are treated as hidden variables and are integrated out. The corresponding marginal likelihood function of observing the entire data set X is then given by:

$$P(X|\Theta,Y) = \prod_{i=1}^{N} \int_{\Phi} P(\tilde{X}_i|\Phi,\Theta,Y)P(\Phi|\Theta,Y)d\Phi \tag{B.5}$$

where $P(\Phi|\Theta,Y)$ gives information about the distribution of the orientations. To Calculate the *prion* it must be assumed that the all Fourier components $\tilde{V}_l$ are, like the noise, independent and Gaussian distributed with zero mean and an unknown variance τ_l^2. So the *prior* could be It can be expressed as the product of independent Gaussian distributions:

$$P(\Theta|Y) = \prod_{l=1}^{L} \frac{1}{2\pi\tau_l^2} exp\left(-\frac{|\tilde{V}_l|^2}{2\tau_l^2}\right) \tag{B.6}$$

Equation B.4 and B.6 can be combined in equation B.3 which leads to a more accurate update algorithm that aims to find the best values for $\tilde{V}_l$, τ_l^2 and σ_{ij}^2:

$$\tilde{V}_l^{(n+1)} = \frac{\sum_{i=1}^{N}\int_{\Phi}\mathbf{G}_{i\Phi}^{(n)}\sum_{j=1}^{J}\mathbf{P}^{\Phi_{ij}T}\frac{CTF_{ij}\tilde{X}_{ij}}{\sigma_{ij}^{2(n)}}d\Phi}{\sum_{i=1}^{N}\int_{\Phi}\mathbf{G}_{i\Phi}^{(n)}\sum_{j=1}^{J}\mathbf{P}^{\Phi_{ij}T}\frac{CTF_{ij}^2}{\sigma_{ij}^{2(n)}}d\Phi + \frac{1}{\tau_l^{2(n)}}} \tag{B.7}$$

$$\sigma_{ij}^{2(n+1)} = \frac{1}{2}\int_{\Phi}\mathbf{G}_{i\Phi}^{(n)}|\tilde{X}_{ij} - CTF_{ij}\sum_{l=1}^{L}\mathbf{P}_{jl}^{\Phi}\tilde{V}_l^{(n)}|^2 d\Phi \tag{B.8}$$

$$\tau_l^{2(n+1)} = \frac{1}{2}|\tilde{V}_l^{(n+1)}|^2 \tag{B.9}$$

$\mathbf{G}_{i\Phi}^{(n)}$ is the posterior probability of Φ for the i-th image, given the model at iteration n, that could be expressed as

$$\mathbf{G}_{i\Phi}^{(n)} = \frac{P(\tilde{X}_i|\Phi, \Theta^{(n)}, Y)P(\Phi|\Theta^{(n)}, Y)}{\int_{\Phi'} P(\tilde{X}_i|\Phi', \Theta^{(n)}, Y)P(\Phi'|\Theta^{(n)}, Y)d\Phi'} \tag{B.10}$$

Instead of assigning an optima orientation Φ_i^* to each image a probability integral is calculated over all possible orientations. Moreover, the update formula (eq. B.7) is derived by an optimization of the posterior distribution and does not involve any arbitrary decisions. Both the power of the noise and the power of the signal are learned from the data in an iterative manner through eq. B.8 and B.9 so the update formula will yield an estimate of $\tilde{V}$ that is both CTF corrected and low-pass filtered, and in which uneven distributions of the orientations of the experimental images are taken into account.

The derivation of equation B.7-9 depends on the assumption of independence between Fourier components of the signal and unfortunately the signal of a macromolecular complex has a limited support in real space due to its low SNR. This leads to an underestimation of the power in the signal that produce a oversmoothing of the reconstruction. This problem was heuristically solved multiplying all estimated value of τ_l^2 by a constant T, in attempt to account for the correlations between Fourier components.

B.2 Resolution estimation

The introduction of the Fourier Shell Correlation (FSC) function [71], defined as:

$$FSC(\nu, \Delta\nu) = \frac{\sum_{[\nu,\Delta\nu]} F_1(\nu)F_2^*(\nu)}{\left[\sum_{[\nu,\Delta\nu]} |F_1(\nu)|^2 \sum_{[\nu,\Delta\nu]} |F_2(\nu)|^2\right]^{\frac{1}{2}}} \tag{B.11}$$

where typically ν dimension is [Å^{-1}], provided a single particle counterpart to SNR estimation. In this procedure the data is randomly divided into two independent half sets, which are reconstructed independently and, at the end of each cycle are compered for consistency. To do so the cross correlation coefficient (CC), between 3D maps reconstructed from the two half data sets is plotted in resolution shells and indicates where the observations are in good agreement (CC=1.0) or where the data are essentially uncorrelated and consist only of noise (CC=0.0).

The importance of having two independent sets reconstructed individually It is closely linked to the low SNR of the Cryo-EM images. Other technique, such as differential phase residual (DPR) or spectral signal to noise ratio (SSNR), work quite well for structure calculated from images of negative stained particles, where SNR is high, but problems appeared when 3D structures were calculated from cryoEM images of ice-embedded single particles, caused by the lower contrast and higher noise, resulting in the need to combine many more single particle images[72]. Therefore, if there are components in the noise image that agreed with those of reference, this could lead to an inaccurate overfitted map[73].

Moreover it was shown[74] that a FSC resolution of 0.143, that relates the two half

sets in such an unbiased refinement, is equivalent to an FSC of 0.5 between the structure calculated from all the data. The term *gold-standard* FSC denotes the calculation of the FSC between two completely independent halves of a data set that had been separated at the start of a 3D structure analysis and compared only after the refinement had been completed[75].
Operationally at the end of every reconstruction interactions an FSC curve between the two independent map is calculated and this curve is converted into an estimate for the resolution dependent SNR:

$$SNR(\nu) = \frac{FSC(\nu)}{1 - FSC(\nu)} \tag{B.12}$$

which is then used to estimate the power spectrum of the signal:

$$\tau^2(n)(\nu) = \frac{SNR(\nu)}{\frac{1}{N_\nu} \sum_{l \in \nu}^{N_\nu} \sum_{i=1}^{N} \int_\Phi \mathbf{G}_i^{\Phi(n)} \sum_{j=1}^{J} \mathbf{P}_{lj}^{\Phi T} \frac{CTF_{ij}^2}{\sigma_{ij}^2} d\Phi} \tag{B.13}$$

where $l \in \nu$ indicates that the l-th 3D Fourier component lies in the resolution shell ν and N_ν is the total number of Fourier components that lie within that resolution shell. The two subsets are joined to calculate a single reconstruction only when the refinement converge. This final reconstruction will have a higher SNR than the two reconstructions from the independent halves of the data, but in order to prevent overfitting it may no longer be used in refinement. To estimate the final resolution of the final electron density map the FSC is modified to:

$$FSC^{'}(\nu) = \sqrt{\frac{2FSC(\nu)}{1 + FSC(\nu)}} \tag{B.14}$$

Consequently, the frequency where the gold-standard FSC is equal to 0.143 indicates the estimated resolution of the map[25, 74].

C | Structure reconstruction

C.1 Structure factors

The structure factors are a mathematical description of how a material scatters incident radiation that can be generically describe as

$$\mathbf{F}_{hkl} = \sum_{j=1}^{N} f_j exp(2\pi i \mathbf{r}_j \tilde{\mathbf{r}}_{hkl}) \tag{C.1}$$

where $\tilde{\mathbf{r}}_{hkl}$ is a generic lattice vector of the reciprocal space with vectorial index hkl, $\mathbf{r}_j$ is a generic direct lattice vector associated with the position of the $j-$th atom in the crystal cell, and N is the number of atoms in the unit cell. f_j is the scattering function of the $j-$th atom, defined *atomic scattering function.* f_j represents the Fourier transform of the electron density of the single atom, ρ_j. If a spherical symmetry is assumed for ρ_j and if the atom is in the origin of the system, the f_j can be also written as:

$$f_j(\tilde{\mathbf{r}}) = \int_0^\infty U_j \frac{sin(2\pi \mathbf{r}_j \tilde{\mathbf{r}}_{hkl})}{2\pi \mathbf{r}_j \tilde{\mathbf{r}}_{hkl}} d\mathbf{r} \tag{C.2}$$

where $Uj(\mathbf{r}) = 4\pi\mathbf{r}2\rho_j(\mathbf{r})$ is the radial distribution function.
The structure factor can be calculated as the Fourier transform of the electron density map $\rho(\mathbf{r})$ can be calculated as

$$\mathbf{F}(\tilde{\mathbf{r}}) = \int \rho(\mathbf{r}) e^{2\pi i \mathbf{r}\tilde{\mathbf{r}}} d\mathbf{r} \tag{C.3}$$

The electron density is a continuum in the whole tridimensional space and its maxima represent the atomic positions.

C.2 Structure refinement

The calculated structure factor $\mathbf{F}_{map}$ (eq. C.3) can be compared with the calculated structure factor of the atomic model $\mathbf{F}_{mod}$, The goodness of the model can be evaluated by the crystallographic R factor defined as:

$$R = \frac{\sum_{hkl} |\mathbf{F}_{map}(hkl) - K\mathbf{F}_{mod}(hkl)|}{\sum_{hkl} |\mathbf{F}_{map}(hkl)|} \tag{C.4}$$

where K is a scale factor. The sum is extended over the reflections in a resolution range decided by the operator.
In order to improve the fitting of the molecular model in the electron density, the least-square methods is often used. It tries to minimize the difference between observed and calculated data.

$$S = \sum_i w_i (F_{i(map)} - F_{i(mod)})^2 \tag{C.5}$$

Unfortunately this method implies a large number of variables, giving rise often to an under determined system.

rigid body refinement

To overcome this problem and provide a significant decrease of the observed data, especially for low resolution maps, the rigid-body refinement (constrained least square method) could be use. Rigid body refine treats a selected group as a rigid entity, the torsional angles are fixed and the group is only allowed to translate and rotate.

Simulate annealing

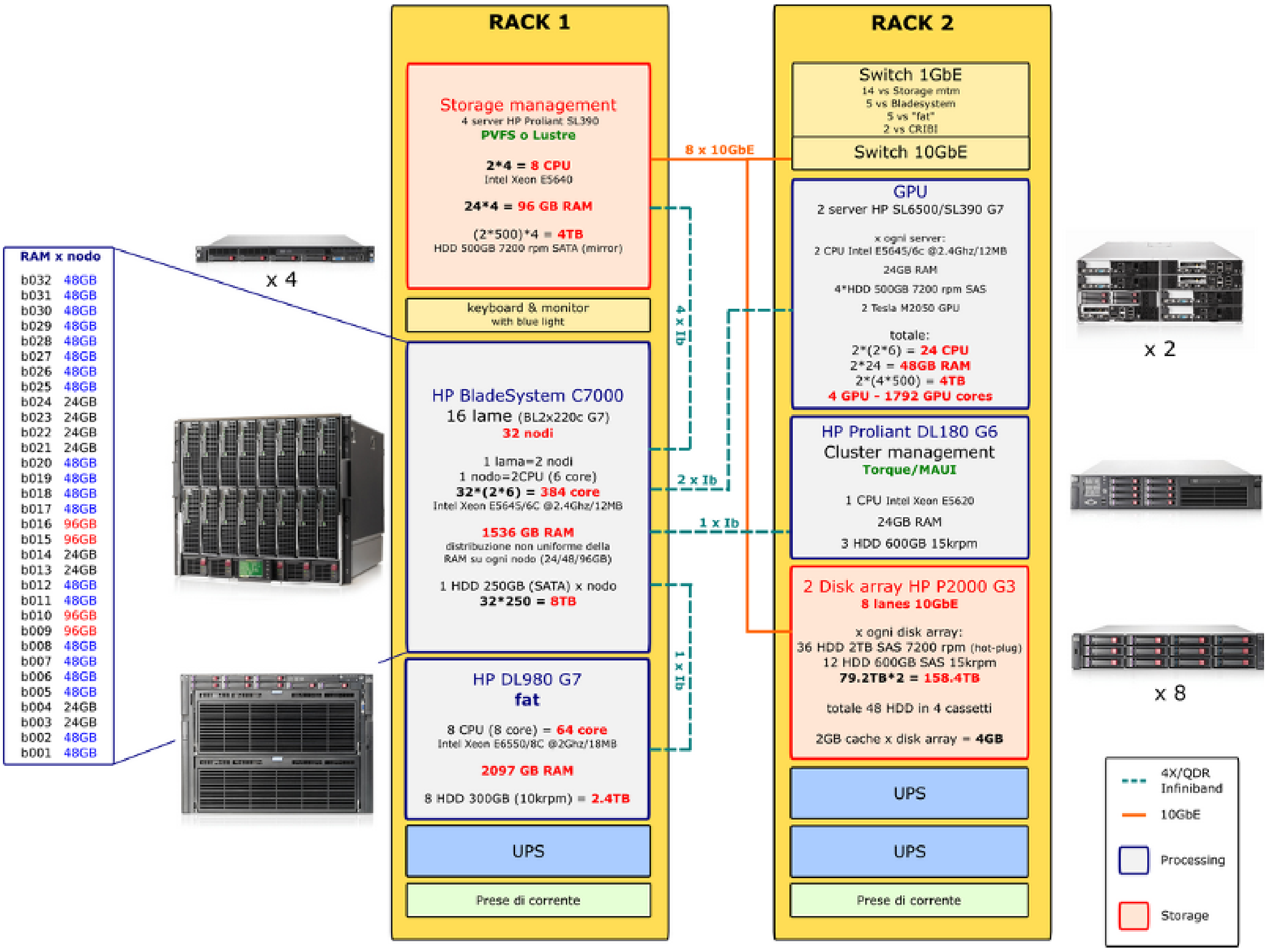
Molecular dynamics can also be used to refine the structure. The method consists in integrating the Newton classical motion equation:

$$m_i \frac{d^2 x_i}{dt^2} = -\nabla E_{pot} \tag{C.6}$$

each atom is associated with an initial velocity, usually related to the Maxwellian distribution of the selected temperature and moves following a trajectory that starts from its position at time $t_0 = 0$ and moves towards a minimum of energy. The potential energy for refinement of a structure can be evaluated from the valence angles, dihedral torsion, bonds and non bonded interactions.
The analysis performed at temperature higher than the ambient one is defined *simulated annealing* and induces the trajectory to move towards the absolute minimum of energy.

D | CRIBI HPC cluster

The main limit in cryo-EM data analysis is the massive computational cost of the script used[54]. Each protocol of the analysis workflow could take from hours to days to be completed and produces a massive RAM usage. Data analysis of this thesis has begun on a eight intel ® Core ™ i7-4790K CPUs @ 4.00 GHz desktop computer with 32 Gb of RAM, but soon we became aware of the need to use a more powerful system. Scipion was then installed on the HPC cluster of *Centro di ricerca interdipartimentale per le biotecnologie innovative* (CRIBI) (fig D.1).



RAM x nodo	
b032	48GB
b031	48GB
b030	48GB
b029	48GB
b028	48GB
b027	48GB
b026	48GB
b025	48GB
b024	24GB
b023	24GB
b022	24GB
b021	24GB
b020	48GB
b019	48GB
b018	48GB
b017	48GB
b016	96GB
b015	96GB
b014	24GB
b013	24GB
b012	48GB
b011	48GB
b010	96GB
b009	96GB
b008	48GB
b007	48GB
b006	48GB
b005	48GB
b004	24GB
b003	24GB
b002	48GB
b001	48GB

Figure D.1: CRIBI cluster scheme

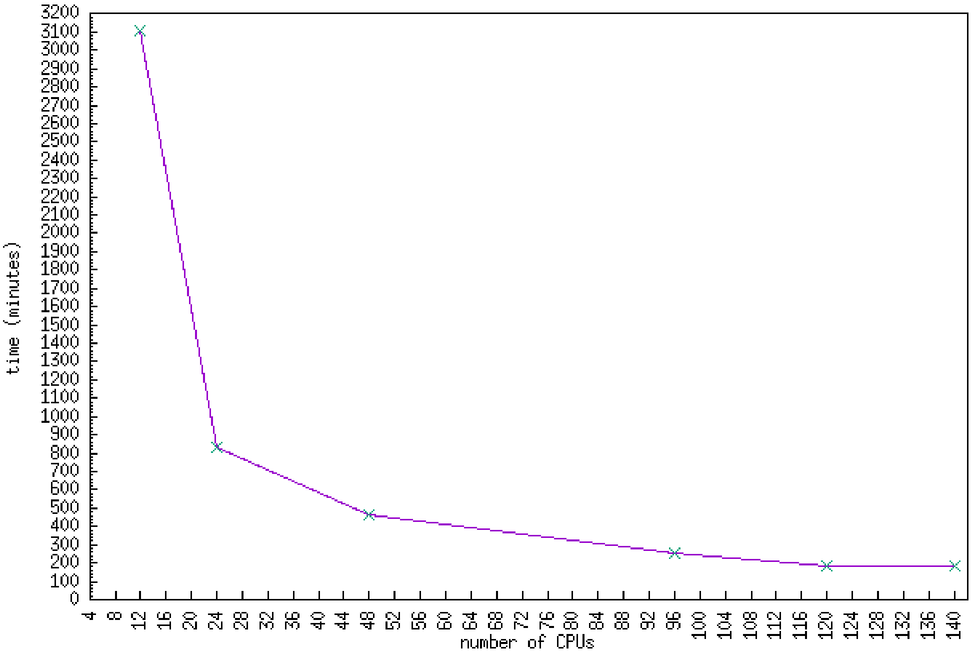
D.1 Infrastructure

The CRIBI HPC cluster is composed by a set of 32 servers (each node has 2x6 cores and 24, 48 or 96 Gb RAM, together called *blade*) and one big servers (called *fat*) with 64 cores and 2 Tb RAM.
There is also a storage system (lustre) of about 150Tb spaceshared between all the machines.
Finally there is a computer, called the *masternode* that acts as a gateway and it is used to launch each process with TORQUE resource manager[76]. Therefore the Scipion *host.conf* script has been modified in order to submit the protocols to the CRIBI queue system. Unfortunately the large number of CPUs required for each protocol does not allow to choose the desired amount of RAM so this parameter was not set and changes according to the availability of the cluster.

D.2 Performance analysis

To optimize the number of processors needed to run the reconstruction protocols the analysis of the Scipion performance on the Cluster was carried out.
like in a similar analysis[77] 10000 particles were reanalyzed with an increasing CPU's number for 2D and 3D classification. In accord to the number of particles the 2D classification was performed with 50 classes (K=50) and the 3D classification with 2 classes (K=2). In order to make the analysis shorter was analyzed the time required for the first three steps.

(a)



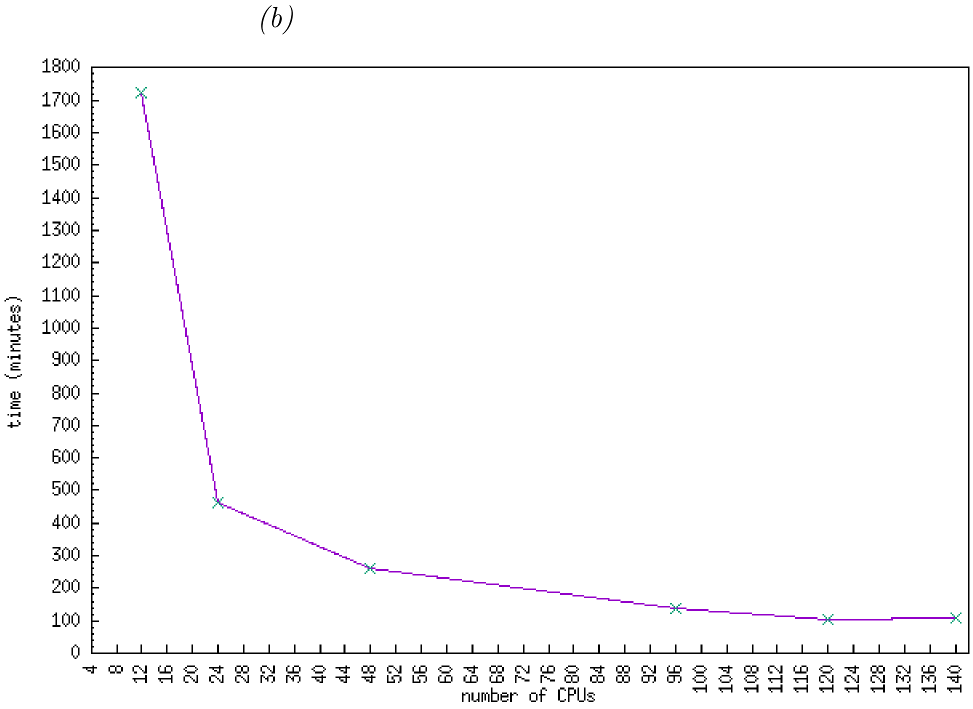


Figure D.2: Time required in minutes in function of the number of CPU used by the 2D (a) and 3D (b) RELION classification protocol.

On the CRBI cluster the computational time does not significantly decrees when more than 120 CPUs are used. As well the difference of execution time with the same number of CPUs between the 2D and 3D classification is strictly related to the number of classes used in classification (K).

The lower number of CPUs a job required the lower is its waiting time in the cluster queue before the execution. Therefore, each protocol was submitted to the cluster requiring 120 or less CPUs and the classes number was set to be the smaller possible according to the data.

E | Abbreviation

Å	Ångstrom
C	Core complex
Ca	Calcium
CC	Cross Correlation
CCD	Charge-Coupled Device
Chl	Chlorophyll
CMOS	Complementary Metal-Oxide Semiconductor
CPU	Central Processing Unit
CRIBI	Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative
Cryo-EM	Cryo Electron Microscopy
CTF	Contrast Transfer Function
DDD	Direct Electron Detector
DDM	Dodecyl D Maltoside
DISAT	Dipartimento Scienza Applicata e Tecnologia
DPR	Differential phase residual
DQE	Detection Quantum Efficiency
EMD	Electron microscopy Data bank
FSC	Fourier Shell Correlation
FT	Fourier Transform
HPC	High Performance Computing
LHCII	Light Harvesting Complex second
LMM	Low Molecular Mass subunit
M	Moderatly bound LHCII trimer
MES	Morpholino Ethanesulfonic Acid
$MgCl_2$	Magnesium Chloride
Mn	Manganese
MTF	Modulation Transfer Function
NaCl	Sodium Chloride
NCS	Non Crystallographic Symmetry
NPQ	Nonphotochemical Quenching
NPS	Noise Power Spectrum
O	Oxygen
OEC	Oxigen Evolving Complex
PDB	Protein Data Bank
Pheo	pheophytin molecule
PSII	Photosystem second
Psb	polypeptides subunits

PSD	Power Spectrum Density
PSF	Point Spread Function
Q	plastoquinone molecule
RC	Reaction Center
REM	Reflection Electron Microscope
S	Strongly bound LHCII trimer
SET	Scanning Electron Microscope
SPA	Single Particles Analysis
SNR	Signal Noise Ratio
SSNR	spectral signal to noise ratio
TEM	Transmission Electron Microscopy
RAM	Random Access Memory

Bibliography

[1] E Ruska. The development of the electron microscope and electron microscopy. Nobel lecture, December 1989.

[2] DB Williams and CB Carter. *The Transmission Electron Microscope*. Springer US, Boston, MA, 1996.

[3] RS Ruskin, Z Yu, and N Grigorieff. Quantitative characterization of electron detectors for transmission electron microscopy. *Journal of structural biology*, 184(3):385–393, 2013.

[4] G McMullan, AR Faruqi, D Clare, and R Henderson. Comparison of optimal performance at 300kev of three direct electron detectors for use in low dose electron microscopy. *Ultramicroscopy*, 147:156–163, 2014.

[5] AR Faruqi and S Subramaniam. Ccd detectors in high-resolution biological electron microscopy. *Quarterly reviews of biophysics*, 33(01):1–27, 2000.

[6] R. Meyer and A. Kirkland. Direct electron detector, November 15 2007. US Patent App. 11/628,184.

[7] M Kuijper, G van Hoften, B Janssen, R Geurink, S De Carlo, M Vos, G van Duinen, B van Haeringen, and M Storms. Fei's direct electron detector developments: Embarking on a revolution in cryo-tem. *Jurnal of Structural Biology*, 192(2):179–87, Nov 2015.

[8] SHW Scheres. Beam-induced motion correction for sub-megadalton cryo-em particles. *Elife*, 3:e03665, 2014.

[9] X Li, P Mooney, S Zheng, CR Booth, MB Braunfeld, S Gubbens, DA Agard, and Y Cheng. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em. *Nature Methods*, 10(6):584–90, Jun 2013.

[10] U Luecken, HG Van, F Schuurmans, and JA De. Method of using a direct electron detector for a tem, November 2 2011. EP Patent App. EP20,100,161,243.

[11] H Shigematsu and FJ Sigworth. Noise models and cryo-em drift correction with a direct-electron camera. *Ultramicroscopy*, 131:61–69, 2013.

[12] XC Bai, G McMullan, and SHW Scheres. How cryo-em is revolutionizing structural biology. *Trends Biochem Sci*, 40(1):49–57, Jan 2015.

[13] E Nogales. The development of cryo-em into a mainstream structural biology technique. *Nature methods*, 13(1):24–27, 2016.

[14] W Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[15] L Wang and FJ Sigworth. Cryo-em and single particles. *Physiology*, 21(1):13–18, 2006.

[16] V Cabra and M Samsó. Do's and don'ts of cryo-electron microscopy: a primer on sample preparation and high quality data collection for macromolecular 3d reconstruction. *J Vis Exp*, (95):52311, 2015.

[17] RF Thompson, M Walker, CA Siebert, SP Muench, and NA Ranson. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods*, 100:3–15, 2016.

[18] FEI$^{\text{TM}}$. *Titant Krios: Visualizing life at the molecular level*, Apr 2015.

[19] Y Cheng, N Grigorieff, P A Penczek, and T Walz. A primer to single-particle cryo-electron microscopy. *Cell*, 161(3):438–449, 2015.

[20] P Cossio and G Hummer. Bayesian analysis of individual electron microscopy images: towards structures of dynamic and heterogeneous biomolecular assemblies. *Journal of Structural Biology*, 184(3):427–37, Dec 2013.

[21] JM Carazo, COS Sorzano, J Oton, R Marabini, and J Vargas. Three-dimensional reconstruction methods in single particle analysis from transmission electron microscopy data. *Archives of biochemistry and biophysics*, 581:39–48, 2015.

[22] M Van Heel and M Schatz. Fourier shell correlation threshold criteria. *Journal of structural biology*, 151(3):250–262, 2005.

[23] PA Penczek. Resolution measures in molecular electron microscopy. *Methods Enzymol*, 482:73–100, 2010.

[24] H Y Liao and J Frank. Definition and estimation of resolution in single-particle reconstructions. *Structure*, 18(7):768–75, Jul 2010.

[25] SHW Scheres and S Chen. Prevention of overfitting in cryo-em structure determination. *Nature Methods*, 9(9):853–4, Sep 2012.

[26] R Henderson, A Sali, M L Baker, B Carragher, B Devkota, KH Downing, EH Egelman, Z Feng, J Frank, N Grigorieff, W Jiang, SJ Ludtke, O Medalia, P A Penczek, PB Rosenthal, MG Rossmann, MF Schmid, GF Schröder, AC Steven, DL Stokes, JD Westbrook, W Wriggers, H Yang, J Young, HM Berman, W Chiu, GJ Kleywegt, and CL Lawson. Outcome of the first electron microscopy validation task force meeting. *Structure*, 20(2):205–14, Feb 2012.

[27] A Kucukelbir, FJ Sigworth, and HD Tagare. Quantifying the local resolution of cryo-em density maps. *Nature methods*, 11(1):63–65, 2014.

[28] J Barber. Photosystem ii: the engine of life. *Quarterly reviews of biophysics*, 36(01):71–89, 2003.

[29] LA Staehelin and GWM van der Staay. Structure, composition, functional organization and dynamic properties of thylakoid membranes. In *Oxygenic photosynthesis: The light reactions*, pages 11–30. Springer, 1996.

[30] J Barber. Engine of life and big bang of evolution: a personal perspective. *Photosynth Res*, 80(1-3):137–55, 2004.

[31] P Albanese, J Nield, JAM Tabares, A Chiodoni, M Manfredi, F Gosetti, E Marengo, G Saracco, J Barber, and C Pagliano. Isolation of novel psii-lhcii megacomplexes from pea plants characterized by a combination of proteomics and electron microscopy. *Photosynth Res*, Jan 2016.

[32] AZ Kiss, AV Ruban, and P Horton. The psbs protein controls the organization of the photosystem ii antenna in higher plant thylakoid membranes. *Journal of Biological Chemistry*, 283(7):3972–3978, 2008.

[33] Y Umena, K Kawakami, JR Shen, and N Kamiya. Crystal structure of oxygen-evolving photosystem ii at a resolution of 1.9 å. *Nature*, 473(7345):55–60, 2011.

[34] B Loll. Ge &amp; science preis für dr. bernhard loll.

[35] G Jackowski, K Kacprzak, and S Jansson. Identification of lhcb1/lhcb2/lhcb3 heterotrimers of the main light-harvesting chlorophyll a/b–protein complex of photosystem ii (lhc ii). *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1504(2):340–345, 2001.

[36] J Standfuss, ACT van Scheltinga, M Lamborghini, and W Kühlbrandt. Mechanisms of photoprotection and nonphotochemical quenching in pea light-harvesting complex at 2.5 å resolution. *The EMBO Journal*, 24(5):919–928, 2005.

[37] T Wan, M Li, X Zhao, J Zhang, Z Liu, and W Chang. Crystal structure of a multilayer packed major light-harvesting complex: implications for grana stacking in higher plants. *Molecular plant*, 7(5):916–919, 2014.

[38] C Pagliano, J Nield, F Marsano, T Pape, S Barera, G Saracco, and J Barber. Proteomic characterization and three-dimensional electron microscopy study of psii-lhcii supercomplexes from higher plants. *Biochimica et Biophysica Acta*, 1837(9):1454–62, Sep 2014.

[39] X Pan, Z Liu, M Li, and W Chang. Architecture and function of plant light-harvesting complexes ii. *Current opinion in structural biology*, 23(4):515–525, 2013.

[40] X Pan, M Li, T Wan, L Wang, C Jia, Z Hou, X Zhao, J Zhang, and W Chang. Structural insights into energy regulation of light-harvesting complex cp29 from spinach. *Nature structural &amp; molecular biology*, 18(3):309–315, 2011.

[41] X Wei, X Su, P Cao, X Liu, W Chang, M Li, X Zhang, and Z Liu. Structure of spinach photosystem ii-lhcii supercomplex at 3.2 å resolution. *Nature*, 534(7605):69–74, Jun 2016.

[42] EJ Boekema, B Hankamer, D Bald, J Kruip, J Nield, A F Boonstra, J Barber, and M Rögner. Supramolecular structure of the photosystem ii complex from green plants and cyanobacteria. *Proceedings of the National Academy of Sciences*, 92(1):175–179, 1995.

[43] EJ Boekema, H van Roon, F Calkoen, R Bassi, and JP Dekker. Multiple types of association of photosystem ii and its light-harvesting antenna in partially solubilized photosystem ii membranes. *Biochemistry*, 38(8):2233–2239, 1999.

[44] J Nield and J Barber. Refinement of the structural model for the photosystem ii supercomplex of higher plants. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1757(5):353–361, 2006.

[45] GS Singhal, G Renger, SK Sopory, KD Irrgang, et al. *Concepts in photobiology: photosynthesis and photomorphogenesis*. Springer Science &amp; Business Media, 2012.

[46] J Barber. Photosystem ii: The water-splitting enzyme of photosynthesis. In *Cold Spring Harbor symposia on quantitative biology*, volume 77, pages 295–307. Cold Spring Harbor Laboratory Press, 2012.

[47] C Pagliano, S Barera, F Chimirri, G Saracco, and J Barber. Comparison of the α and β isomeric forms of the detergent n-dodecyl-d-maltoside for solubilizing photosynthetic complexes from pea thylakoid membranes. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1817(8):1506–1515, 2012.

[48] S Barera, C Pagliano, T Pape, G Saracco, and J Barber. Characterization of psii-lhcii supercomplexes isolated from pea thylakoid membrane by one-step treatment with α- and β-dodecyl-d-maltoside. *Philos Trans R Soc Lond B Biol Sci*, 367(1608):3389–99, Dec 2012.

[49] S Järvi, M Suorsa, V Paakkarinen, and EM Aro. Optimized native gel systems for separation of thylakoid protein complexes: novel super-and mega-complexes. *Biochemical Journal*, 439(2):207–214, 2011.

[50] FEI$^{\mathrm{TM}}$. *Falcon II: 16 megapixel TEM electron detector with back thinned sensor technology*, 2013.

[51] JM de la Rosa-Trevín, A Quintana, L Del Cano, A Zaldívar, I Foche, J Gutiérrez, J Gómez-Blanco, J Burguet-Castell, J Cuenca-Alba, V Abrishami, J Vargas, J Otón, G Sharov, JL Vilas, J Navas, P Conesa, M Kazemi, R Marabini, COS Sorzano, and JM Carazo. Scipion: A software framework toward integration, reproducibility and validation in 3d electron microscopy. *Journal of Structural Biology*, 195(1):93–9, Jul 2016.

[52] JM de la Rosa-Trevín, J Otón, R Marabini, A Zaldívar, J Vargas, M Carazo, and COS Sorzano. Xmipp 3.0: an improved software suite for image processing in electron microscopy. *Journal of Structural Biology*, 184(2):321–8, Nov 2013.

[53] G Tang, L Peng, PR Baldwin, DS Mann, W Jiang, I Rees, and SJ Ludtke. Eman2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46, 2007.

[54] SHW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519–30, Dec 2012.

[55] JA Velaquez-Muriel, COS Sorzano, JJ Fernąndez, and JM Carazo. A method for estimating the ctf in electron microscopy based on arma models and parameter adjustment. *Ultramicroscopy*, 96:17–35, 2003.

[56] COS Sorzano, R Marabini, J Velázquez-Muriel, JR Bilbao-Castro, SHW Scheres, JM Carazo, and A Pascual-Montano. Xmipp: a new generation of an open-source image processing package for electron microscopy. *Journal of structural biology*, 148(2):194–204, 2004.

[57] SHW Scheres. Classification of structural heterogeneity by maximum-likelihood methods. *Methods Enzymol*, 482:295–320, 2010.

[58] SHW Scheres. A bayesian view on cryo-em structure determination. *J Mol Biol*, 415(2):406–18, Jan 2012.

[59] SHW Scheres, R Núñez-Ramírez, COS Sorzano, JM Carazo, and R Marabini. Image processing for electron microscopy single-particle analysis using xmipp. *Nature protocols*, 3(6):977–990, 2008.

[60] EF Pettersen, TD Goddard, CC Huang, GS Couch, DM Greenblatt, EC Meng, and TE Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[61] K Arnold, L Bordoli, J Kopp, and T Schwede. The swiss-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, 2006.

[62] PD Adams, PV Afonine, G Bunkóczi, VB Chen, IW Davis, N Echols, JJ Headd, LW Hung, G J Kapral, RW Grosse-Kunstleve, et al. Phenix: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):213–221, 2010.

[63] PV Afonine, JJ Headd, TC Terwilliger, and PD Adams. Phenix tools for validated refinement of atomic models into maps (low-resolution, cryo-em, x-ray or neutron). *Computational Crystallography Newsletter*, Volume 4, part 2:43–44, 2013.

[64] H Kohl and L Reimer. *Transmission Electron Microscopy: Physics of Image Formation.* Springer-Verlag New York, 2008.

[65] M Vulović, E Franken, RBG Ravelli, LJ van Vliet, and B Rieger. Precise and unbiased estimation of astigmatism and defocus in transmission electron microscopy. *Ultramicroscopy*, 116:115–134, 2012.

[66] JJ Fernández, JR Sanjurjo, and JM Carazo. A spectral estimation approach to contrast transfer function detection in electron microscopy. *Ultramicroscopy*, 68(4):267–295, 1997.

[67] F Thon. Phase contrast electron microscopy. *Electron Microscopy in Material Sciences*, pages 571–625, 1971.

[68] EJ Kirkland. Linear image approximations. In *Advanced computing in electron microscopy*, pages 19–39. Springer, 1998.

[69] SHW Scheres, H Gao, M Valle, GT Herman, PPB Eggermont, J Frank, and JM Carazo. Disentangling conformational states of macromolecules in 3d-em through likelihood optimization. *Nature methods*, 4(1):27–29, 2007.

[70] SHW Scheres, R Núñez-Ramírez, Y Gómez-Llorente, C San Martín, PPB Eggermont, and JM Carazo. Modeling experimental image formation for likelihood-based classification of electron microscopy data. *Structure*, 15(10):1167–1177, 2007.

[71] G Harauz and M van Heel. Exact filters for general geometry three dimensional reconstruction. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conf*, volume 73, pages 146–156, 1986.

[72] N Grigorieff. Resolution measurement in structures derived from single particles. *Acta Crystallographica Section D: Biological Crystallography*, 56(10):1270–1277, 2000.

[73] R Henderson. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences*, 110(45):18037–18041, 2013.

[74] PB Rosenthal and R Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J Mol Biol*, 333(4):721–45, Oct 2003.

[75] S Chen, G McMullan, AR Faruqi, GN Murshudov, JM Short, SHW Scheres, and R Henderson. High-resolution noise substitution to measure overfitting and validate resolution in 3d structure determination by single particle electron cryomicroscopy. *Ultramicroscopy*, 135:24–35, 2013.

[76] A Telatin. *Using the CRIBI HPC cluster: a very short tutorial.* CRIBI Biotechnology Center Genomics and Bioinformatics Unit, Dec 2011.

[77] MA Cianfrocco and AE Leschziner. Low cost, high performance processing of single particle cryo-electron microscopy data in the cloud. *eLife*, 4:e06664, 2015.

Ringraziamenti

Ho sempre considerato la vita come una scalata in montagna: a volte gli appigli si trovano facilmente e la via da seguire si disegna agile, altre, invece, basta solo un passaggio sbagliato per trovarsi lungo un percorso difficile e pericoloso. Peggio ancora, può accadere che un temporale costringa a rimanere inchiodati nello stesso punto per ore, senza poter ne salire e scendere. Arrivato ad una sosta lungo il mio percorso, mi sembra doveroso fermarmi un attimo per ringraziare tutti i compagni di cordata che mi hanno accompagnato e tutte le guide che mi hanno indicato la via da seguire.

In primo luogo vorrei ringraziare il Professor Giuseppe Zanotti per la sua grande disponibilità e per avermi dato la possibilità, i mezzi e i giusti consigli per scoprire la crio microscopia elettronica. Ringrazio anche il Professor Flavio Seno per la disponibilità e l'interesse dimostrato per il progetto. Ringrazio la Dott.ssa Cristina Pagliano e il Dott. Pascal Albanese per il supporto datomi nelle fasi iniziali di questo lavoro e per il confronto sui risultati ottenuti. Ringrazio il Dott. Andrea Telatin e Diego Bacchin per i consigli su come meglio utilizzare il cluster di calcolo del CRIBI. Ringrazio Paola per il sostegno durante gli alti e bassi di questa tesi e per aver ascoltato tutte le mie elucubrazioni. Insieme a lei ringrazio anche Tito, Francesca, Domenico, Mattia, Laura e Lorenzo per avermi accolto tra loro e, ognuno a modo suo, aver contribuito al completamento di questo lavoro. Un ringraziamento particolare va a Valentina per il costante scambio di opinioni e per i mille consigli. Infine ringrazio di cuore tutti coloro che, tra compagni di corso e professori, mi hanno aiutato a superare i molti momenti difficili, vi devo molto perchè senza di voi non sarei potuto arrivare a questo punto.

Un ringraziamento speciale va a quello sparuto gruppo di persone che mi ha s(u-o)pportato in questi anni, grazie per i cineforum, per le chiacchiere davanti a una birra, per le serate sotto le stelle e per non avermi ancora mandato a quel paese.

Un pensiero particolare va a Vittoria, grazie per aver condiviso al mio fianco quasi un terzo delle nostre vite, grazie per essere cresciuta con me tra tante gioie e qualche dolore e soprattutto grazie per aver deciso di passare il resto della tua vita con me. Ovunque saremo, se ci sarai tu al mio finaco, sarò a casa.

Per ultimo un grandissimo ringraziamento va ai miei genitori, a mia madre che mi ha sempre lasciato fare le mie scelte in completa autonomia e a mio padre che per primo mi ha parlato con il magnifico linguaggio della fisica, grazie per tutto quello che mi hai insegnato e per la passione che mi hai trasmesso per questo mondo. Di sicuro, se sono qui ora lo devo maggiormente a te. Però mi rimarrà sempre il grande rammarico di non esserci arrivato prima che tu te ne andassi.

Ora però è il momento di ripartire perché la carne è molto vicina all'osso, ma l'osso non è ancora stato stritolato.