Novel cryo-EM image processing methods to address conformational heterogeneity

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

MICHAELA SPIEGEL
aus Wuppertal

Jülich, März 2017

aus dem Institute of Complex Systems - Strukturbiochemie (ICS-6)
am Forschungszentrum Jülich

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Jun. Prof. Dr. Gunnar Schröder
Korreferent: Prof. Dr. Dieter Willbold

Tag der mündlichen Prüfung: 04.07.2017

Zusammenfassung

Die Einzelteilchen Analyse mithilfe von Kryoelektronenmikroskopie ist eine Technik in der Strukturbiologie die gerade in den letzten 5 Jahren immer mehr an Bedeutung gewonnen hat. Hardware, Software und Probenherstellung sind inzwischen soweit, dass immer mehr Strukturen mit hohen Auflösungen besser als 3 Å bestimmt werden können. Hierbei können beliebig große Komplexe untersucht werden.

Mikroskopische Aufnahmen von eingefrorenen einzelnen Proteinen und Proteinkomplexen sind Schnappschüsse von biologischen Maschinen in annähernd nativem Zustand. Je nach Probe ist eine mehr oder wenige breite Palette von struktureller Variabilität in den Daten vorhanden. Dies sind Kompositionsmischungen und Konformationsänderungen. Dies birgt ein enormes Potential biologische Komplexe in unterschiedlichen funktionalen Zuständen zu untersuchen und biologisch zu interpretieren. Hierzu ist es allerdings notwendig die Daten zu klassifizieren was aufgrund des sehr niedrigen Signal-zu-Rausch Verhältnisses ein schwieriges Problem darstellt. Daher ist der Umgang mit Heterogenität in der Kryoelektronenmikroskopie immer noch ein hochaktuelles Thema. In dieser Arbeit wurde eine neue Methode entwickelt, mit der sich Daten in Klassen unterschiedlicher Konformationsänderungen sortieren lassen.

Diese Methode basiert auf einer Hauptbewegungsanalyse. Bei diesem Verfahren wird durch eine statistische Stichprobenwiederholungstechnik (Bootstrapping) die Dichtevarianz der Daten bestimmt und in eine strukturelle Varianz transformiert. Nachdem unterschiedliche Dichtevolumen aus den klassifizierten Aufnahmen rekonstruiert worden sind, lassen sich diese durch ein flexibles Verbiegen entlang der berechneten Hauptbewegungen aufeinander abbilden. Hierdurch ist es möglich über diese verbogenen Dichtevolumen zu mitteln und somit wieder zu einem höher aufgelöstes Dichtevolumen zu gelangen, welches alle Bilder enthält.

Die beiden Methoden zum Sortieren und Verbiegen von Dichtevolumen wurden an einem simulierten Datensatz entwickelt und getestet. Anschliessend wurden sie auf Kryoelektronenmikroskopie Daten angwandt, die das Escherichia coli 70S-Ribosom mit $tRNA^{sec}$ und co-Faktor SelB enthielten. Die verschiedenen Schritte der Datenanalyse wurden dann auf unterschiedliche Weise validiert. Letzendlich konnten für beide Datensätze Dichtevolumen in unterschidelichen konformellen Zuständen berechnet werden.

Anschliessend wurde noch an einer Methode gearbeitet, mit deren Hilfe sich die unterschiedlichen Dichtevolumen wieder auf das mittlere Startvolumen zurück biegen lassen. Die Verbiegung der atomaren Struktur ist durch die Hauptbewegungsanalyse bereits bestimmt worden. Der Eigenvektor, welcher auf die atomare Struktur wirkt, wird nun auf das Gitter des jeweiligen Dichtevolumens angewandt. Nachdem die Dichtevolumen zurück gebogen wurden, können diese gewichtet mit der Fourier-Shell Korrelation aufsummiert werden und so wieder ein Dichtevolumen, welches alle Aufnahmen enthält, berechnet werden. Die Auflösung und Information in dem Dichtevolumen hat sich allerdings verbessert. Diese Methode wurde auf die Dichtevolumen und Eigenvektoren aus den beiden Datensätzen die wir zuvor sortiert haben angewandt, und eine Verbesserung des gemittelten Dichtevolumens im Vergleich zum Ausgangsvolumen konnte gezeigt werden.

Aus Kryo-EM Daten rekonstruierte Dichtevolumen sind oft unscharf und detaillierte Strukturmerkmale sind nicht klar erkennbar. Dies liegt an unterschiedlichen Faktoren des Experimentes wie z.B. mechanische Vibrationen. In dieser Arbeit wurde eine neue Methode zum schärden dieser Dichten entwickelt. Hierbei nutzen wir statistische Informationen die wir über Proteine haben und wenden diese auf Dichtevolumen im Real- und Fourierraum an. Dichtevolumen von Proteinen in unterschiedlichen Auflösungsbereichen lassen sich durch dieses Verfahren optimieren, d.h. die Volumen sind schärfer aufgelöst, Strukturmerkmale sind besser zu erkennen und allgemein ist die Dichteverteilung stärker ausbalanciert.

Summary

Single particle analysis (SPA) using cryo-electron microscopy is a technique in structural biology, which has become increasingly important in the last 5 years. Hardware, software, and sample preparation have improved so much that more and more structures at resolutions better than 3 Å can be resolved. In this case, arbitrarily large complexes can be examined.

Microscopic images of frozen individual proteins and protein complexes are snapshots of biological machines in the near native state. Depending on the sample, a more or less wide range of structural variability is typically present in the data. These are composition mixtures and conformational changes. This has an enormous potential to investigate and interpret biologically complexes in different functional states. For this purpose, however, it is necessary to classify the data, which is a difficult problem because of the very low signal-to-noise ratio. Therefore, dealing with heterogeneity in cryo electron microscopy is still a topical issue. In this thesis a new method was developed with which data can be sorted into classes of different conformations.

This method is based on a principal motion analysis. In this method, the density variance of the data is determined by a statistical random resampling technique (bootstrapping) and transformed into a structural variance. After different density volumes have been reconstructed from the classified images, they can be imaged on top of each other by a flexible bending along the calculated main motions. This makes it possible to average these bent density volumes and thus to achieve a higher resolution density volume which contains all images.

The two methods for sorting and bending density volumes were developed and tested on a simulated data set. Subsequently, they were applied to cryo-electron microscopy data of the E. coli 70S ribosome with $tRNA^{sec}$ and co-factor selB.The different steps of the data analysis were then validated in different ways. Finally, for both data sets several reconstructions in different conformational states could be reconstructed.

Subsequently, a method was developed with the aid of which the different density volumes can be bent back to the mean starting volume. The deflection of the atomic structure has already been determined by the principal motion analysis. The eigenvector which acts on the atomic structure is now applied to the grid of the respective density volume. After the density volumes have been bent back, they can be weighted with the Fourier-Shell correlation and a density volume containing all images can

be calculated. However, the resolution and information in the density volume has improved. This method was applied to the density volumes and eigenvectors from the two data sets we have previously sorted and an optimization of the density volume to the output volume could be shown.

Density maps reconstructed from cryo-EM data are often unsharp and detailed structural features are not clearly recognizable. This is due to different factors of the experiment, e.g. mechanical vibrations. In this work, a new method was developed. We use statistical information about proteins and apply this information to the density volumes in real and Fourier space. Density maps in different resolution ranges can be improved by this method which means the volume gets sharper, structural features are better to recognize and generally the density distribution is more balanced out.

Eidesstattliche Versicherung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist.

Ferner erkläre ich, dass ich nicht anderweitig mit oder ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.

(Michaela Spiegel)

Contents

1 Introduction

1.1 Aim of the Work

The main goal of my doctoral thesis was to develop a method by which the heterogeneity of a cryo-EM data set can be separated into classes of similar conformational states by using the principal motion analysis. The principal motion analysis (PMA) has already been developed in the group, and a new project has emerged from this. The idea to use the PMA for sorting cryo EM images is an approach for solving the still-challenging heterogeneity problem in cryo EM.

This method has the goal to gain a complete picture of the conformational variability and at the same time gain high resolution reconstructions since high-resolution can only be reached if the conformational variability is appropriately accounted for.

In our method images are sorted based on conformational variance which is a new approach and recalls on the common approach of sorting cryo-EM images by the density variance. This yields a number of density maps in slightly different conformations.

Furthermore, one can bring these conformationally different density maps back together by applying the knowledge of principal motions between them. Density maps are elastically deformed to one conformational state and averaged to increase the overall resolution further.

In order to achieve these goals, I had to study software and underlying methods of single particle analysis. I looked at how realistically cryo EM images of a heterogeneous data set could be simulated and applied the newly developed methods to this simulated data.

The reliability of the PMA was tested on a simulated data set and the next sorting steps were developed. After achieving several classes with this sorting method with a high assignment accuracy the method was applied to real cryo-EM data of a ribosome.

After the flexibility and principal motions of this data where studied it became clear that local parts are flexible but these movements do not seem to be correlated. A focused classification with a principal motion sorting was applied and classes in different conformational states from local areas could be reconstructed.

1.2 Thesis Outline and Contributions

The Chapter 2 "Electron Microscopy of Biological Specimen" is intended to provide an introduction into the application of TEMs in Structural Biology. The build up of a transmission electron microscope (TEM) is explained and how the contrast in the cryo-EM images arises. Techniques to increase the contrast in the images are presented, as well as imaging errors which arise due to electromagnetic lens system are explained. Afterward the Fourier slice theorem is explained which is the basis of 3D reconstructions from projection images.

The next Chapter 3 “Single Particle Analysis” (SPA) explains how the samples are produced and how images are detected in the microscope. Next is the image processing part, which is explained a bit more in detail. There are several steps from starting by picking good images of many thousands of micrographs until a 3D density map is reconstructed. Finally, the model bias problem is discussed and it is explained how density volumes can be validated and interpreted and how an x-ray structure can be flexibly fitted into a density map.

After these two chapters about introduction to the cryo-EM field with basic concepts, conventions and methods in Chapter 4 a published work for “Improving the Visualization of Cryo-EM Density Reconstructions” follows. Me and my colleague Amudha Duraisamy developed a method which we called **vi**sualization improvement by **s**tructure factor and **de**nsity histogram correction of cryo-EM **m**aps or short VISDEM. In this project, we have both contributed exactly the same. We have written the scripts together, the manuscript and also the other calculation were all done together.

In Chapter 5 “Sorting Cryo-EM Images by Principal Motions” calculations are shown which were made in order to develop the new sorting method. Here I explain the basic idea of the new method. Followed up by a description how realistic cryo EM images can be simulated. A simple system with images from two different conformations is built and a similar method for sorting cryo-EM images (codimensional density PCA (SparX)) is first tested on this system. Subsequently, the images were sorted according to principal motions first with the atomic structure and then a bead model (mass points connected by springs) is used. Different things like the required number of connections between the beads as well as different techniques in the projection matching are investigated.

The next Chapter 6 “Principal Motions of Molecular Machines from Cryo-EM Data” consists of an almost submitted manuscript. It is about the principal motion analysis (PMA) applied to a simulated test system and GroEL/ES data. The GroEL/ES data was contributed by Dong-Hua Chen, Junjie Zhang and Wah Chiu. The method was developed and the calculations of GroEL/ES data were performed by Gunnar Schröder and Benjamin

Falkner. I simulated a data set on which I applied the PMA and the underlying motions could be determined correctly by this method. I share the first-authorship with Benjamin Falkner.

The next Chapter 7 “Sorting cryo-EM images into classes of similar conformations by principal motion analysis” is also close to being submitted for publication. This is the main project on which I spent most of my time. The method is applied to the same test system as used in the Chapter before and on ribosome cryo-EM data. In the ribosome case it is shown how the method can be applied as a focused classification method and how the individual steps of the method can be validated for a correct proceeding. Additionally images were sorted for the same regions with a focused 3D classification in Relion and the results were compared.

Finally the last project in Chapter 8 “Bending of density maps into different conformational states directed by atomic eigenvectors” is also written in form of a manuscript but it is not yet submitted. It is the extension of the previous chapter, in which the density maps representing different conformational states (which were sorted in the previous chapter) are bent back to be averaged.

The thesis ends in the last Chapter 9 with a conclusion.

2 Electron Microscopy of Biological Specimen

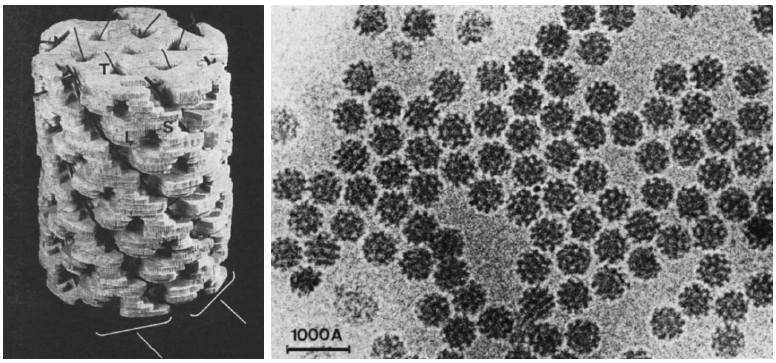
The electron microscope was invented in 1930 by Knoll and Ruska. The achievement of this microscope is that it was possible to resolve structures on a much smaller length scale. It is a useful tool to detect magnified images of surfaces or small objects up to atomic resolution and it has become a mainstream technique in structural biology. There are two possibilities to use it for biological probes depending on the size of the sample.

Subtomogram Averaging One method is to cut out a thin section for example of the inner part of a cell. Then electron micrographs are taken under different tilt angles and are later merged together to get a three-dimensional view of the slice. This is called subtomogram averaging.

Single-Particle Analysis The other method is called single-particle analysis (SPA) and is a method to determine the structure of one single particle, which could be a protein, protein complex or biological machine. The specimen is stabilized by wheter chemical or cryo-fixation and millions of particle images are taken. Threedimensional volumes can be reconstructed out of these particles by applying the Radon Transfromation. During my doctoral thesis I have worked on cryo-EM data of single particles. An overview about all the individual steps from the specimen to the final reconstructed density map is given in Chapter 3.

Early Work The first important work of studying biological particles was done in 1968 by de Rosier and Klug [19]. They solved the first three dimensional density map by a TEM image of the tail of bacteriophage T4. The technique was adapted similiar to techniques used for X-Ray cristallography reconstructions, by using a fundamental relation between detected images and real structure, called radon transformation (Section 2.4). They determined a three dimensional density by only one single electron micrograph taking the helical symmetry of the structure into account. At this time it was not possible to view the density on a computer, so they visualized their calculations in form of a wooden model (Figure 2.1 (a)) to get an idea of the three-dimensional shape.

The first study of biological particles by the “cryo”-EM technique was done in 1984 [2] and a very impresive high contrast micrograph of Semliki Forest



(a) (b)

Figure 2.1: (a) Handmade wood model of Bacteriophage T4 tail, calculated out of one single TEM image in 1968 [19].
(b) Transmission electron micrograph of a frozen sample from Semliki Forest viruses detected in 1984 [2].

viruses is shown in Figure 2.1. These images made it possible to determine the triangulation number $T = 4$ (which describes the symmetry) of this icosahedral virus, which was a controversy before. It is exciting to see that the main ideas of the Single-Particle Analysis were already established so early. For example the freezing of the specimen they introduced is still up to date and this will be explained in more detail in Chapter 3.1. They also already mentioned the requirement of underfocus detection and the oscillatory effect of contrast and the neccessity of varying different defoci and developping a software for putting everything computationally together. All these problems were solved in the following years and the method of Single-Particle Analysis was evolving further.

Resolution Revolution The single particle cryo-EM technique was chosen by the famous journal Nature Methods as the "Method of the year 2015". An overview of the key contributions of the last years of the SPA is given in the historical overview of Eva Nogales [52].

The breakthrough of this method took place in the year 2012 with the "Resolution Revolution" [44]. A sudden resolution jump below the long time thought not able to be overcoming barrier of $4\,\text{Å}$ resolution happened. This happened simultaneously with the development of the direct electron detector. Previously used CCD (charge-coupled device) detectors translated the incoming electrons into photons before detecting them. The new detectors directly detect the electrons and therefore have a much higher detective quantum efficiency (DQE) [51].

New image processing techniques like maximum likelihood techniques [68] and 3D classification procedures, were improving the image processing part further. In the end, it was an unexpected synergy effect of improved

image quality and improved algorithms which led to high resolved density maps [5].

Since the resolution revolution the field is growing faster and faster. The understanding of high/low resolution since I started my PhD in 2013 and 2015 has already changed a lot, as can be seen in Table 2.1.

	low resolution	intermediate	high resolution
2013	< 20Å	~ 8Å	~ 5Å
2015	< 8Å	~ 4-5Å	> 4Å

Table 2.1: What we understand under different resolutions in Single-Particle Analysis?
Gordon Research Conference Three-dimensional Electron Microscopy 2015

2.1 Build up of the Transmission Electron Microscope (TEM)

The basic structure of a transmission electron microscope (TEM) is quite similar to a light microscope as can be seen in Figure 2.2. The photons were replaced by electrons and the optical devices were replaced by electromagnetic devices. The first thing which is needed is an electron source delivering coherent electrons. First TEMs used tungsten wires, nowadays a field emission electron gun (FEG) is used with a tungsten/zirkonium oxide tip. The FEG delivers a more coherent and smaller diameter electron beam.

The electrons are accelerated with a voltage typically between 200-300 keV for cryo-EM samples. With the de Broglie wavelength $\lambda = \frac{h}{p}$ and the relativistic energy equation

$$eU_B + mc^2 = \sqrt{m^2c^4 + p^2c^2}, \tag{2.1}$$

a wavelength of $\lambda_e \approx 2$ pm ($\nu \approx 10^{20}$ Hz) which is in the order of gamma radiation is obtained.

The resolution of a microscope is restricted by the numerical aperture NA of the microscope and described as the distance d_{min} where two points are still distinguishable. This is called Rayleigh criterion:

$$d_{min} = \frac{0.61\lambda}{NA}. \tag{2.2}$$

The numerical aperture, $NA = n \cdot \sin \alpha_{max}$, is defined by the product of the refractive index n with the sinus of the maximum angle of deflection which is still detectable.

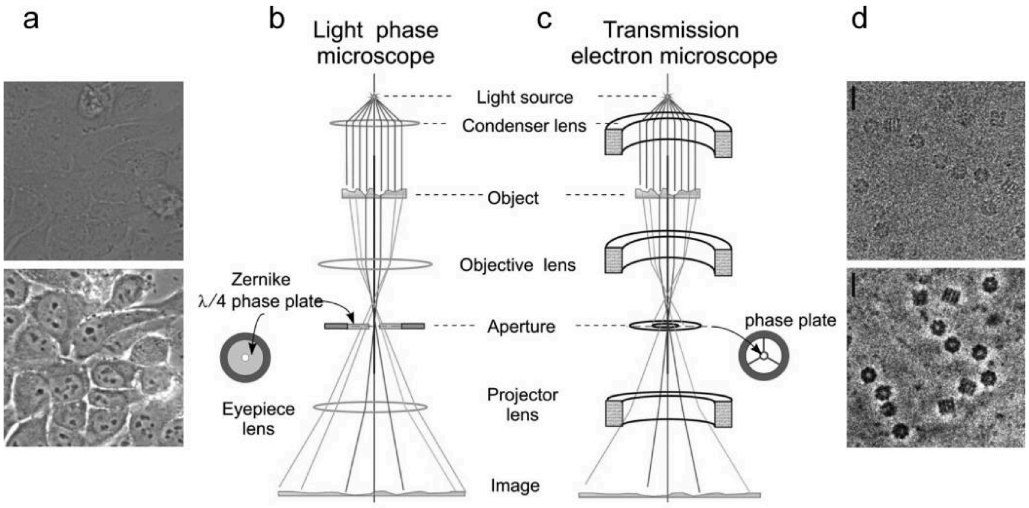


Figure 2.2: The build up of an electron microscope (c) is similar to a normal light microscope (b). The effect of the phase plate, explained in the next Section 2.2 on the image contrast can be seen in light (a) and electron microscopy (d). This picture is taken from [54].

In the electron microscope the resolution is not limited due to the wavelength or numerical aperture but by the sensivity of the biological specimen. In order to stabilize the specimen different techniques can be used like chemical or cryo-fixation (see Chapter 3.1).

The microscope works under ultrahigh vacuum with pressures around 10^{-5}Pa which require a multi-stage pumping system. For cryo-EM the microscope needs a liquid nitrogen-cooled specimen holder and also other parts are cooled down by liquid nitrogen. This is called cryo-shielding it improves the vacuum quality and lowers the specimen drift. The electrons arrive at the specimen after they went through the condensor system which is build by two or three electromagnetic lenses.

After this some modern microscopes have a spherical aberration corrector and an energy filter, which filters out non-elastic scatterers and increases the signal-to-noise ratio.

The last element of the TEM is the detector which can be photographic film, a CCD camera or a direct electron detector. More information on the new direct electron detectors can be found in Section 3.2 about image detection.

2.2 Image Formation and Phase Contrast

Weak Phase Approximation The highly accelerated electrons hit the sensitive specimen and destroy the sample by breaking chemical bonds, but some electrons are reflected by the light nuclei, they undergo an elastic

scattering. These are the electrons which are responsible for the signal in the detected images. They hit the detector later due to their longer pathway which leads to a phase shift in the wave function.

The signal in the image can be described as a two dimensional projection $p(x, y)$ of the three dimensional Coulomb potential of the specimen ($f(x, y, z)$):

$$p(x, y) = \int f(x, y, z)\,dz. \tag{2.3}$$

This equation is correct in the weak phase approximation. Electrons are scattered by light atoms, like typically in biological specimen (H, N, O, C), which lead to a small phase shift that is smaller than the wavelength of the electrons.

The electron beam can be described by a plane wave Φ_0 which undergoes a phase shift due to the Coulomb potential:

$$\Phi_s(x, y) = \Phi_0 e^{i\sigma p(x,y)} \approx \Phi_0(1 + i\sigma p(x, y)) \qquad \text{with} \quad \sigma = \frac{m_e \lambda}{2\pi\hbar^2}. \tag{2.4}$$

Since the phase shift is quite small this leads to a measurable signal of:

$$I(x, y) = |\Phi_s|^2 \approx 1 + (\sigma p(x, y))^2. \tag{2.5}$$

The specimen has to be smaller than the mean free path between two scattering events, which is around $1\mu m$. There is also a minimum thickness of the specimen in order to have enough signal in the image. In terms of particle weight the range lies between 0.5-100 MDa [54].

The phase contrast between scattered and unscattered electrons is very low. For a specimen with thickness of 500 Å and an electron beam with 300 kV, only 5% of the images are elastically scattered, 15% are inelastic scattered and the rest remains unchanged (source: Talk from S.Scheres). The particle signal can be distinguished hardly from noise. But this is truly necessary for further data analysis. There are usually two possibilities to increase the contrast between signal and noise either by phase plates or by detection of the specimen in defocus.

Phase Plate A phase plate is a device which is used for shifting the phase of an incoming electron beam. The easiest one is the Zernike phase plate which is a carbon plate with a hole in the middle. This plate is located behind the specimen directly under the convex lens. The scattered electrons are phase shifted by $-\pi/2$ and the unscattered electrons pass the phase plate unaffected through the hole in the middle. There are several types of phase plates available. Phase plates are still very expensive and not very long-living.

The effect of a $\pi/2$ phase shift on the signal of the image described by Equation 2.5 changes to a much higher contrast in the image with:

$$I(x, y) = |\Phi_{90^\circ}|^2 \approx |\Phi_0(1 - \sigma p(x, y))|^2 \approx 1 - 2\sigma p(x, y). \tag{2.6}$$

The other idea instead of using a phase plate is to induce a stronger contrast in the image by detecting the specimen on purpose under defocus. The combination of spherical aberration and defocus works similar to a phase plate. This is the cheaper and mainly used method and will be explained in more detail in the next section.

2.3 Aberations and Contrast Transfer Theory

Aberations The detected image consists of signal, noise, artifacts and aberrations due to the optical system. Typical image aberrations are astigmatism, chromatic and spherical aberration. Astigmatism originates from asymmetrical lenses, a beam line slightly shifted to the optical axis of the lenses or other asymmetrical set ups, chromatic aberration from fluctuations in the electron wavelengths and spherical aberration from the lens system. Astigmatism can be overcome by an accurate adjustment of the experiment, chromatic aberration by using an energy filter, which filters out "wrong" wavelengths and spherical aberration (C_s) can also corrected by an C_s-corrector [36]. This corrector produces negative spherical aberration which combines with the positive aberration of the objective lens to zero spherical aberration.

But the use of a C_s-corrector is still not common. So digitalized micrographs usually go directly through a numerical correction analysis. In 1948 Scherzer [71] concluded: "*Electron lenses of the usual type cannot be corrected spherically.*" Real lenses are approximated as circular shaped but they have to be parabolic. This approximation leads to focal length of inner/small shifted electrons which is smaller than from more far away shifted ones and the defined focal length (f) is the one from the inner beam.

The spherical aberration of the lenses leads to an aberration effect on the micrographs which can be described by the contrast transfer theory.

Contrast Transfer Theory The contrast transfer function (CTF) describes how the electrons are modulated by the electro magnetic lenses under defocus and spherical aberration in the microscope. It is the Fourier transform of the point spread function which describes how one single point is transformed. To increase the contrast the specimen is detected

on purpose under a small (µm) defocus which is described in the next paragraph in more detail. The CTF function is defined as the following:

$$\mathrm{CTF}(k) = \sqrt{1 - A^2}\sin(\gamma(k)) + A\cos(\gamma(k)) \tag{2.7}$$

$$\gamma(k) = -\frac{\pi}{2} c_s \lambda^3 k^4 + \pi \lambda z k^2. \tag{2.8}$$

The contrast transfer function in frequency space k depends on the wavelength λ of the electrons, the spherical aberration constant c_s of the microscope and the defocus z. The amplitude contrast A is typical between 4 and 7% in cryo-EM but for CTF estimation often set to 10%.

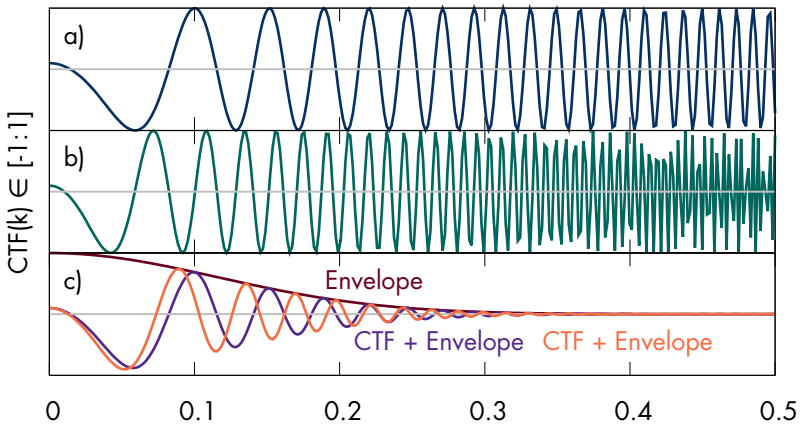


Figure 2.3: Three different contrast transfer functions (CTF), all with a spherical aberration value C_s of $2\,\mathrm{mm}$, defocus z of $-0.8\,\mu\mathrm{m}$ and amplitude contrast A of 10%.
a) CTF with an acceleration voltage of 300kV (λ = 0.019 Å).
b) CTF with an acceleration voltage of 100kV (λ = 0.037 Å).
c) CTF in violet is CTF function of a) multiplied with an envelope function shown in red. The orange curve is the same function under a different defocus(-1µm).

The CTF is shown for some realistic parameters in Figure 2.3. The two upper images show the function with different wavelengths. But there is an additional effect which plays a big role in real images. It is called the B-factor. The B-factor is a constant which describes experimental issues like drift effects or energy spread in the electron beam. Its a sum of different kind of fluctuations. A higher B-factor leads to suppression of high frequency information. This effect can be described with an envelope function (E(k)):

$$E(k) = \exp(-Bk^2/4). \tag{2.9}$$

It is shown in the third image of Figure 2.3, where the CTF from Equation 2.8 is multiplied by the envelope function in Equation 2.9.

Typically used CTF parameters are determined from recently published densities at the Electron Microscopy Database (EMDB)[21]. Densities

which are reconstructed from movie frames were not considered. The used acceleration voltage is typically 200-300 kV, the spherical aberration constants vary between 2-2.7 µm and the used defoci between 0.7-5 µm. Electron doses are between 20-30 $e^-/Å^2$, but are not required for CTF correction.

Defocus As above mentioned specimen are detected under certain defocus to increase the SNR in the image which means the contrast in the micrographs gets higher and we actually are becoming able to distinguish between particle and noise by eye. This is simulated in Figure 2.4. The first image a) shows a projection of a Ribosome density map. In the second image b) noise is added ('pink' noise based on an actual noise curve from a 300 kV microscope `http://blake.bcm.edu/emanwiki/EMAN2`) until the projection is hardly visible and also a CTF function with typical parameters (voltage=300 kV,c_s=2 mm, B-factor=130) and zero defoucd is applied. Even if the image is low-pass filtered (not shown) the particle is difficult to detect. In the third image c) the defocus is changed to 0.8µm. This blurs out the image but at the same time increases the contrast, so that the particle becomes easy to find. A CTF correction with the correct parameters applied on this image is shown in d). Actually a lot of the signal can be preserved. The same is done under an even higher underfocus of 1.5µm in the images e) and f).

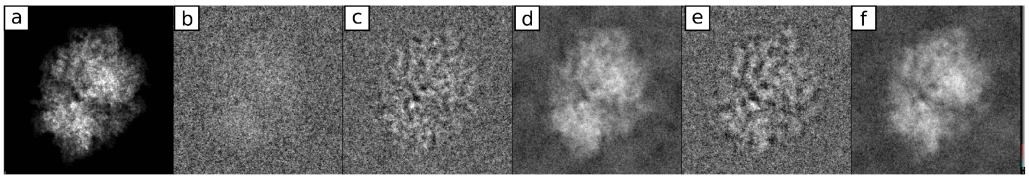


Figure 2.4: Image a) shows a projection image of a Ribosome. The same image plus noise and CTF function with a defocus of zero is shown in b). Image c) shows the same image as b) but with a defocus of 0.8µm, after CTF correction it results in image d). Image e) shows the same image as b) and c) but with a defocus of 1.5µm and image f) is the ctf corrected version of image e).

The influence of defocus and spherical aberration on the contrast of an image was already discussed by Scherzer [71] in 1948. He also calculated the optimal defocus (z) for detecting a single point, which is known as Scherzer focus:

$$z_{\text{Scherzer}} = -\sqrt{\lambda c_s}. \tag{2.10}$$

The whole image A cryo-EM image X_i can be described as projection $\mathbf{P}_\phi$ under unknown orientation ϕ from a Volume V_k of conformation k plus random noise N_i [70],

$$X_i = \text{CTF} \otimes \mathbf{P}_\phi V_k + N_i, \tag{2.11}$$

The problem to recalculate the underlying volumes V_k is inverse, incomplete and ill-posed. In order to invert this formula and determine the volume it is necessary to determine the noise level, CTF parameter and orientation parameter of the detected particles X_i. The whole thing without any model bias. This is complicated due to the low signal-to-noise ratio (SNR), which is defined as $\mathrm{SNR} = var(\mathrm{signal})/var(\mathrm{noise})$ and needs additional information to realize a 3D reconstruction process. All these different steps are explained step by step in Chapter 3.3 in the image processing section.

2.4 Radon Transformation and Fourier Slice Theorem

The EM images of biological specimen are well described by the weak phase approximation as projection images from the specimen. Projections of a three dimensional volume in several different projection directions contain all the information which is needed to reconstruct back the three dimensional volume of the specimen. The relation between the projections and the three dimensional structure is given by the Fourier slice theorem. For better understanding the theorem is first introduced in 2D.

2.4.1 Explanation in two Dimensions

In Figure 2.5 we show an example of a one dimensional projection in pink described by function $p_\theta(r)$ of a two dimensional object in blue described by function $f(x,y)$. The projection is the integral over the object $f(x,y)$, which means $p(r) = \int_{-\infty}^{+\infty} f(x,y)dz$.

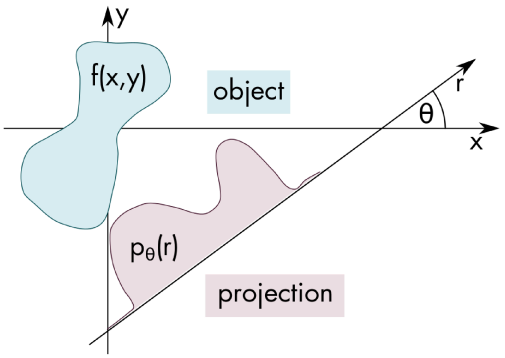


Figure 2.5: The projection (pink) of the object (blue) is proportional to the integral over the volume.

The transformation between the two coordinate systems is given by:

$$\begin{pmatrix} x \\ y \end{pmatrix} = A_\theta \begin{pmatrix} r \\ z \end{pmatrix} \qquad \text{with} \qquad A_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{2.12}$$

The relation between object and projection is called Radon Transformation

Radon Transformation 2D

$$p_\theta(r) = \int_{-\infty}^{+\infty} f\left(A_\theta \begin{pmatrix} r \\ z \end{pmatrix}\right) dz \qquad \text{(Radon Transformation)}$$
$$= \int_{-\infty}^{+\infty} f(r\cos(\theta) - z\sin(\theta), r\sin(\theta) + z\cos(\theta))dz.$$

If there are now a lot of projections $p_\theta(r)$ under different angles θ, it is possible to reconstruct the object by applying the Fourier slice theorem.

Fourier Slice Theorem in 2D Let $p_\theta(r)$ be a projection under angle θ of a two dimensional object $f(x, y)$, with $P_\theta(\rho) = FT\{p_\theta(r)\}$ and $F(u, v) = FT\{f(x, y)\}$, then:

$$P_\theta(\rho) = F\left(\rho\cos(\theta), \rho\sin(\theta)\right). \qquad \text{(FST 2D)}$$

This means to reconstruct the object the Fourier transforms of the projections are combined to two dimensional object and then inverse Fourier transformed.

Proof of Fourier Slice Theorem Let $\theta = 0 (\rightarrow x = r)$ (projection in y-direction, see Figure 2.5)

$$p_\theta(r) = \int_{-\infty}^{+\infty} f(x, y)dy \qquad (2.13)$$
$$P_\theta(\rho) = \int_{-\infty}^{+\infty} p_\theta(r)e^{-i2\pi r\rho}dr \qquad (2.14)$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(r, y)e^{-2\pi(r\rho+y\cdot 0)}drdy = F(\rho, 0) \qquad (2.15)$$

The Fourier transform is rotational invariant. □

2.4.2 Fourier Slice Theorem in 3D

The Fourier slice theorem can also be applied to higher dimensions. For cryo-EM reconstructions it is used in three dimensions and it is shown in Figure 2.6. The EM images of biological specimen are in the very first approximation noisy projection images from the specimen. Many images (100,000 or more) from all different projection directions contain all the information which is needed to reconstruct the three dimensional volume of the detected particle.

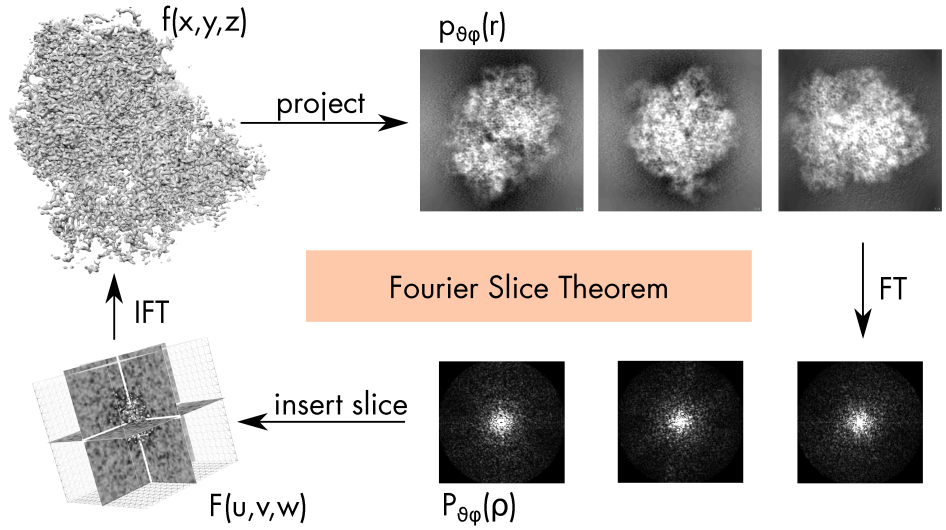


Figure 2.6: **The Fourier Slice Theorem.** This image shows schematically how a three dimensional volume can be calculated from its two dimensional projections.

3 Single Particle Analysis

Single particle analysis (SPA) is a technique to determine the shape or structure of one single particle which could be a large protein, protein complex or biologically machine where other techniques like NMR or X-Ray usually fail. NMR is limited to a size of maximum 400 amino acids and X-Ray is limited by particles which form crystals. Crystallization of proteins often needs a lot of work and luck to find the correct environmental conditions.

Single particle analysis has been applied favorably for determining large virus structures. The high symmetry of viruses makes it relatively easy to achieve a high resolution density map by applying this symmetry during reconstruction and averaging over symmetrical volumes. There exist many viruses with icosahedral symmetry, which means they have 20 equal faces.

This chapter will give an overview about the whole workflow from sample preparation, image detection, up to image processing and reconstruction of density maps and finally interpretation of them. The focus is more on the computational than on the experimental part.

A good overview about the whole process is given by M. van Heel [87], J.Frank [28] and E. Orlova [54]

3.1 Specimen Preparation

The electron microscope works under high vacuum. This means the particles which are stable in an aqueous solution have to be transferred into a solid state for observation in a way that they do not lose their natural shape. This can be done by chemical fixation which is usually done by negative staining or by freezing the sample and do the measurement in the microscope under cryogenic temperatures. This is called cryo-electron microscopy (cryo-EM). Both methods stabilize the sensitive biological materials and make them more robust against damages.

3.1.1 Sample Preparation of Negative Staining Samples

The samples are fixed on a holey carbon grid (also other materials have been used, e.g. gold grids). A grid is 3mm in diameter. For negative

staining (Figure 3.1 a) an additional support film is applied over the grid. Then a blob of buffer-particle mixture is added on the grid and additionally a 1-2% solution of a heavy metal salt (typically uranyl acetate). Now the specimen has to dry under air.

The heavy metal salt outlines the shape of the particles and acts like a stain later under the electron microscope. The micrographs show a high contrast and the outer shape of the particles is clearly detected. The specimen has been negatively stained.

There is no information about the inner structure in the micrograph. This limits the resolution of the later 3D reconstruction to a resolution around 20Å. Negative staining is often used for a first test and to get an idea of the particle quality and the overall size and shape.

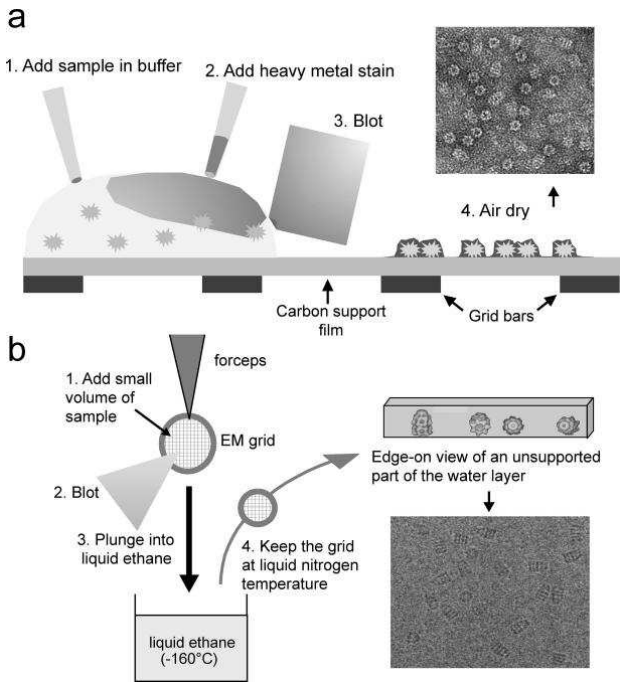


Figure 3.1: a) Sample preparation step for negative staining specimen.
b) Sample preparation steps for cryo-EM samples [54].

3.1.2 Sample Preparation of Cryo-EM Samples

In Figure 3.1 b we can see the steps we need to apply to create a cryo-EM sample. After adding the particle solution on a grid, most of the solution will be blotted dry by a piece of paper such that only a very thin layer remains. The goal is to prepare the sample that just one single particle layer remains.

In the next step the grid is plunged into liquid ethane (90 K). This is usually done by using a gadget where the grid can be attached to a piece which falls into the liquid ethane bath by activating the gadget.

The sample is shock-frozen in such a short time that the water molecules have no time to form out a crystal structure. Instead, they stay in an amorphous state which has a structure similar to liquid water and is called vitreous ice. This operation is also called “plunge-freezing” [17] [39].

A more detailed step by step protocol can be found in [32]. The frozen specimen has to be placed in a liquid nitrogen cooled specimen holder into the microscope. It is important to make sure that the specimen stays in its frozen state the whole time.

3.2 Image Detection

The cooled grids will now be placed into the microscope. Different microscopes can accommodate different numbers of grids (packed in a cartridge) at the same time. The loading of the grids into the microscope, microscope preparation and data collection is explained in more detail in the following protocol [31] on the example of a TECNAI TEM (FEI).

First the grid quality is checked at a low magnification around 100-300 times. If the quality is sufficient the electron gun has to be aligned and an area of interest has to be found. The decision has to be made at which magnification, electron dose and pixel size the images are taken. Then the certain area has to be centered and focused. This needs to be done slowly since the specimen can drift a little bit. After the exposure with electrons the cryo-sample area should be burnt by high electron dose, since after one exposure it got unusable.

The image recording with the TEM is sensitive to mechanical vibrations and especially electrical fields. The most powerful and modern microscopes are usually installed in a single room, and all the operations like lens settings, stage movement position, illumination, tilt focus and so on are all computer controlled from another room outside.

Originally, really low electron doses $20\text{-}30\, e^-/\text{Å}^2$ have been used in order to keep the radiation damage of the sensitive specimen low. The electrons were detected by photographic film and later by digital CCD cameras. The new detector generation of direct electron detectors which are monolithic active pixel sensors (MAPS) have a much higher detective quantum efficiency (DQE) but they are more sensitive and do not allow such high electron doses. In order to get enough signal the movie mode was invented. The same micrograph was detected with a much lower electron dose around $2-5e^-/\text{Å}^2$ but for a longer time and several frames were captured. Later they are corrected for beam induced motion and recombined to one single micrograph.

3.3 Image Processing

The reconstruction of a three dimensional density map by two dimensional cryo-EM images is due to the incompleteness only possible to solve by introducing additional information. It is dangerous because wrong assumptions can lead to wrong density maps. How this problem can be solved and the whole image processing from the detected images to the three dimensional reconstruction works will be explained in this section step by step.

3.3.1 CTF Correction

The first thing to do is to determine the CTF parameter of the micrograph. In Section 2.3 we already explained how the CTF is composed. The power spectrum of the micrograph (amplitude of Fourier transform) shows concentric rings. They are called Thon Rings and stem from the CTF. The highest frequency information available in the micrograph can be determined from the outest Thon ring. The CTF of Equation 2.11 is fitted to this pattern until the correct defocus for each micrograph is determined.

An example of a cryo-EM micrograph and the corresponding power spectrum with visible Thon rings is shown in Figure 3.2

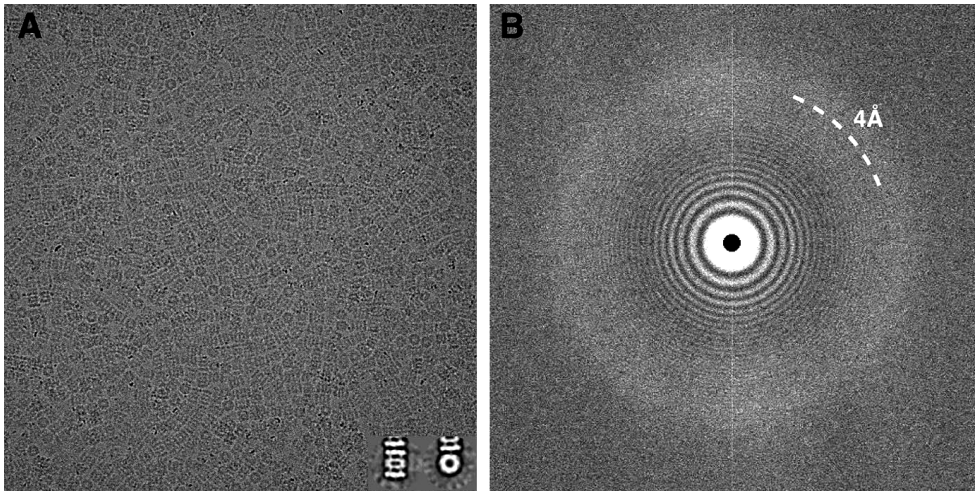


Figure 3.2: This figure is taken from a recently published work by M. Campbell et al [11] and shows a“typical micrograph of ice-embedded T20S proteasome and corresponding power spectrum.”

The most distributed software is CTFFIND from Niko Grigorieffs Lab [62]. It is not possible to totally invert the effect of the CTF even if the correct parameters are known since the function goes through zero. This is one part of the inversion problem in cryo-EM.

There are approximations available to deblur the image and correct for the CTF. First one is the phase flipping and the second one is the Wiener filter, as explained in the following.

Phase Flipping

The phase in the detected micrographs oscillates between negative and positive values described by the CTF. Once the CTF is determined the phase of the micrograph can be multiplied by this function. This is called phase flipping because by multiplying with the "same" oscillating function the negative phases are flipped into positives, the phase flipped images are much sharper.

Wiener Filter

The Wiener filter is a way to restore an image by estimating the real signal in the image as close as possible and distinguish it from noise. It will correct for the CTF and at the same time filter out noisy frequencies.

It is a well known tool in image restoration and the solution of the following problem:

$$g(x,y) = f(x,y) * h(x,y) + n(x,y). \tag{3.1}$$

The image $g(x,y)$ which we want to restore is a convolution of the real image $f(x,y)$ by a point spread function $h(x,y)$ plus random noise $n(x,y)$. A goal is to minimize the error between the real image and the restored image. The Wiener filter is a solution to the mean square error, which is basically the expectation value $E[.]$ of the square root between the difference of the real image $f(x,y)$ and the restored image $\hat{f}(x,y)$.

$$e^2 = E[(f(x,y - \hat{f}(x,y))^2] \tag{3.2}$$

This can be minimized by putting in the definition of the image and derive this expression in order to get the minimized solution.

The solution $\hat{F}(u,v)$ can be found in Fourier space as

$$\hat{F}(u,v) = \frac{H^*(u,v)G(u,v)}{H^2(u,v) + S_{noise}/S_{signal}}, \tag{3.3}$$

with the Fourier transform of $h(x,y) \rightarrow H(u,v)$ and $g(x,y) \rightarrow G(u,v)$. The S_{noise}/S_{signal} is the ratio between the power spectrum of the noise and the signal. The power spectrum is the magnitude of the Fourier transform. This term is frequency dependent and hard to estimate. In many implementations of Wiener filtering this is replaced by a constant

which leads to unsatisfying results. The function $H(u, v)$ is in a cryo-EM image the contrast transfer function (CTF).

3.3.2 Validating Micrographs and Particle Picking

The micrograph needs to show particles in random orientations and particles need to have a certain distance to each other, e.g. do not cluster together. A homogeneous, not too thick ice layer is needed to avoid density gradients on the micrograph. The particles should be able to be distinguished from the background noise.

A micrograph can also be judged by its power spectrum as described below. Ellipse-shaped formed Thon rings indicate an incorrect movie frame alignment or astigmatism errors. Astigmatism can also be corrected during CTF correction but should be avoided. Missing high resolution information in the micrographs leads to reconstructions missing high resolutions. Eventually one has to go back to the sample preparation or to the image detection step.

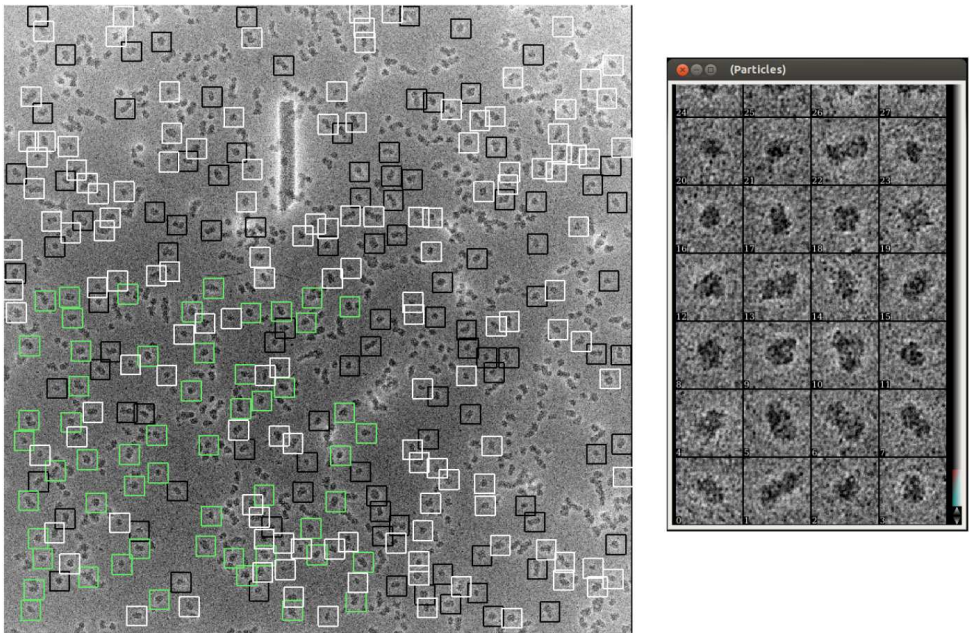


Figure 3.3: Example of a negative stain micrograph and picked particles in EMAN2 (e2boxer.py). The green boxes show particles which are picked by hand and used for automated particle search. Particles found by search are in white boxes and black boxes indicate particles picked by hand. On the right we can see some of the picked particles in higher magnitude.

Only good quality micrographs should be used for particle picking. The higher the contrast the easier it gets to pick “correct” particles. Particles are cut out in a quadratic box which is usually 2/3 of the particle size.

Particle picking can be done manually, automated. Automated particle picking softwares are using pattern recognition algorithms. They are often reference based and need a pattern to search for. This can be manually picked projections, class averages of further picked particles or projections from a reference structure. Only searching for projections especially from a structure determined by a different technique is not a good idea due to model bias, see Section 3.4.1.

In Figure 3.3 a micrograph of negative stained amyloid oligomers is shown. The particles were picked by hand and shown in black boxes. The particles in the green boxes were picked and used by a search algorithm. Found particles are shown in white boxes.

Reference patterns can be searched in the micrograph by calculating the convolution of the pattern with the micrograph or search for it by cross-correlation. There are many more algorithms for particle search available [98].

Manually picking is often not possible due to the huge amount (millions) of particles needed for a high resolution structure. Since the accuracy of pattern recognition in low SNR micrograph still leads to many wrong decisions people will pick a lot of particles by hand than do an automatic search for these particles and later go through all “particles” by eye and delete the wrong and bad ones. This is called semi-automated picking. There are even other techniques to figure out incorrect particles by stochastic scoring functions [92].

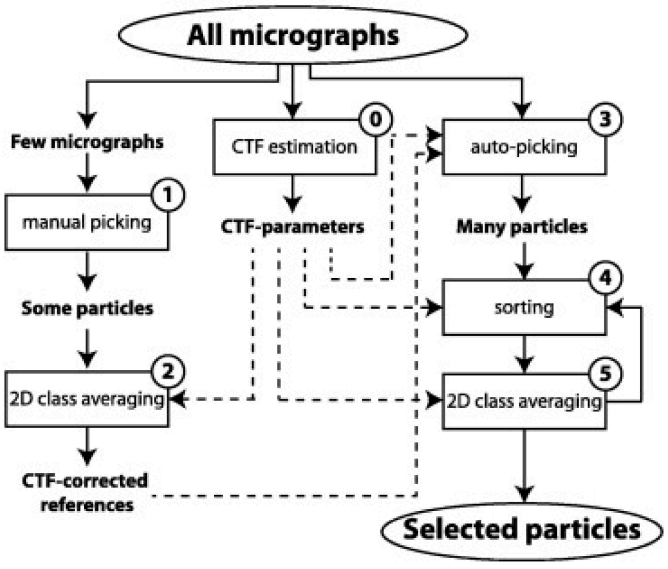


Figure 3.4: Five steps of a semi-automated particle selection proposed for Relion [69].

Although there are still wrong particles inside the data set they can still be sorted out by 2D class averaging as explained in the next step. In Figure 3.4

it is shown how a typical particle selection process is proposed by S.Scheres [69] which is a semi-automated procedure in five steps.

3.3.3 2D Class Averaging

Class averaging is a method to clean the data set of bad particles and see whether the detected image quality is sufficient for a 3D reconstruction. The images are clustered into sets of particles which show the particle in the same orientation. If this is done correctly, an average over many particles should give a much clearer projection image, since the random noise of several images is averaged out and the signal is summed up.

The calculated class averages give an idea about the data quality and if enough images were detected. High frequency information in class averages indicate good chances to achieve a density map containing high frequency information. “Bad” classes with less images and low resolution can contain images of weakly populated conformational states, particles containing artifacts or just wrong particles. Only “good” classes are chosen for further reconstructions. Before the classes can be determined, the images have to be normalized first.

Normalization The picked particles all have different contrast. So they have to be normalized in order to have the same impact in the reconstruction process. They are usually normalized by subtracting the mean density value $\bar{\rho}$ and dividing by the standard deviation σ:

$$\rho_{\text{norm}} = \frac{\rho - \bar{\rho}}{\sigma}. \tag{3.4}$$

Alignment by Maximum Cross-Correlation That means the images must be compared to each other and need to be aligned. To judge if two images show the same particle in the same orientation the images have to be displaced to each other iteratively due to their three degrees of freedom (2 translations and 1 rotation) and checked for similarity.

This can be done by calculating in which orientation the two images have the maximum cross-correlation coefficient (ccc). It is defined for image a and image b with N number of pixel:

$$\text{ccc} = \frac{\sum_{i=1}^{N} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i^2}}. \tag{3.5}$$

One difficulty is that the rotated and shifted image and the second image are not sequential. So the values of one image have to be interpolated on the grid of the other image.

2D Class Averages - Multivariate Statistical Analysis This is a typical procedure of class averaging how it is implemented in programs like Imagic [89] or EMAN2 [80]. The idea of calculating class averages by multivariate statistical analysis was first mentioned by van Heel in 1981 [90].

This method works as followed. First a subset of images is chosen and initial classes are generated. This is explained on `http://blake.bcm.edu/emanwiki/EMAN2`. After the initial guesses of classes are calculated they are sorted by similarity and aligned towards each other. Now the basis images (also called eigenimages) of the aligned class averages are calculated by a principal component analysis. All images are aligned towards a given number of class references and the best matched rotational and translational parameters are used. The projections of all images with a given number of basis images are calculated and the images are clustered to the given number of classes by k-means clustering. This results in new classes and the procedure starts again with sorting the classes, calculating basis images and so on. After some iterations it will converge.

In Figure 3.5(a) a selection of GroEL cryo-EM images are shown. This was a data set which was used in the EMAN2 Tutorial and is available online (`http://blake.bcm.edu/emanwiki/EMAN2`). After only a few iterations one could obtain considerable results. Some of the class averages are shown in Figure 3.5(c) and the eight basis images are shown in Figure 3.5(b).

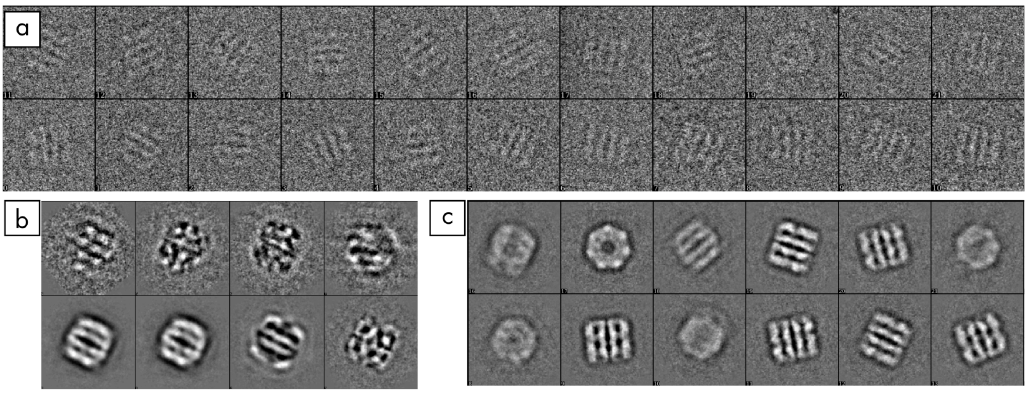


Figure 3.5: Class averages of GroEL Data from EMAN2 Workshop 2014. In (a) a selection of GroEL cryo-EM images are shown. In (b) eight highest eigenimages (principal components) of class averages and in (c) a selection of class averages.

The user has to set up a list of parameters and the most important one is the number of classes. This number is usually set such that each class consists of around 100 images. Many parameter like the number of classes in a subset for alignment, or number of basis images for clustering and so on have to be chosen by the user and optimal values differ for different structures e.g. the number of basis vectors for GroEL classes should be around 3-4 and for ribosome classes between 10-15 in this case due to symmetry. So the user's expertise is necessary to get best possible results

with this method.

2D Class Averages by Maximum Likelihood The class averaging with maximum likelihood approach is another newer approach implemented in software packages Frealign [33] and Relion [67]. I will focus on the implementation in Relion since I worked with this program. The maximum likelihood approach was first derived by Fred Sigworth in 1998 [73]. The whole procedure is an iterative multi reference refinement. So all images are compared to all classes in all possible orientations (rotational and translational). First initial guesses are made by randomly dividing the data into the desired number of classes. This leads to circles with the diameter of the particle. By assigning the images to the classes the resolution of "useful" classes is increasing during each iteration step.

The corresponding classes have small changes which will crystallize out after some iterations. The program will increase the angular sampling accuracy in each iteration step with improvement in class resolution.

In Figure 3.6 an example of class averages of the ribose binding protein in Relion is shown after different iteration steps. The classes of the ribose binding protein look quite useful. But there is also one class shown (lower right one) which did not converge to a useful representation even after 50 iterations. Usually the images of classes like this are then removed before the 3D reconstruction.

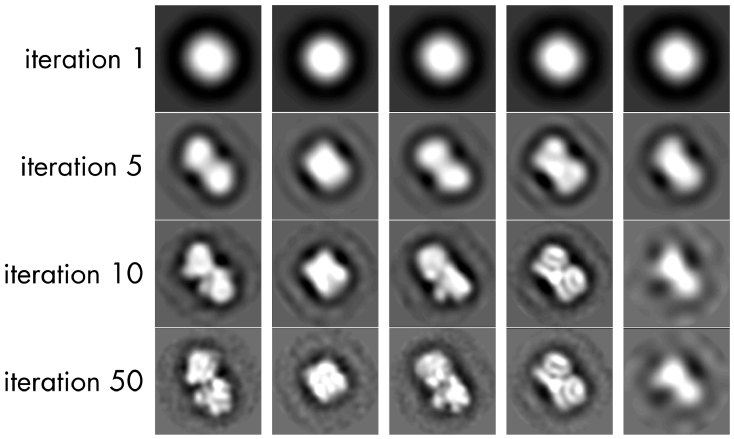


Figure 3.6: Example of five classes calculated by a 2D Classification in Relion after different number of iteration steps.

In the maximum likelihood approach probabilities are calculated and each image is weighted in the class averages by their corresponding probability. This means in each class average different images are weighted differently. Images can also contribute to different classes. A low-pass filter in form of a Wiener filter method is already implemented inside the class averaging calculations. This means the resulting classes are filtered due to the ratio between signal and noise in the 2D averages.

3.3.4 3D Reconstruction

The 3D reconstruction is based on the Radon transformation. A 3D density map can be recalculated by the projection images as described in Chapter 2.4 by Fourier transforming the images, adding them up to a 3D object and then inverse Fourier transforming this object to get the density map. This would be a very easy task if the orientation parameter were known, which means we have to know which image belongs to which projection direction. Since this has to be determined it becomes a more difficult problem.

Each cryo-EM image has five degrees of freedom, the orientation in space can be described in spherical coordinates by two angles. One additional angle describes the in-plane rotation. The orientation is typically described by Euler angles. Since the picked particles are not perfectly centered two additional coordinates have to be determined which are translational shifts in x- and y-direction. One early approach was the common lines approach.

Common Lines Approach The common lines approach first described by van Heel [85] is based on the fact that two 2D projections have at least one common 1D line (dependent on the symmetry also more) in Fourier space. This can be seen in the Section 2.4 about Radon transformation in Figure 2.6. If all these common lines can be correctly determined the Euler parameter are identified. Before this can be done the translational shifts of the images have to be determined. There exist different approaches like e.g. center of mass.

Afterwards a sinogram is calculated which means all lines through the center of the images are written out in a certain step size by rotating the image up to $180°$. Then two sinograms can be scanned for common lines.

This approach was first implemented in the Imagic software package [90], but leads to too inaccurate assignments. It is still applied for calculation of initial models from class averages [49]. The more established approach is the iterative refinement.

Iterative Refinement Approach In the iterative refinement we need to have an initial model first to which we can assign our images. Projections from the model are calculated and compared to the experimental images in order to determine their orientation parameters. This procedure is called “projection matching”. Then the images are reconstructed and the new model is low-pass filtered due to the corresponding resolution and used for a new round of image assignment. In the image assignment, all five degrees of freedom must be determined. The angular sampling accuracy will increase during the refinement. This is very time consuming. So after some iterations the search space will be reduced to an area close to the

determined Euler angles. It is like a local optimization and therefore errors can occur if the initial model is very different from the correct solution. After some iteration the refinement will converge to a density map and the resolution will no longer increase.

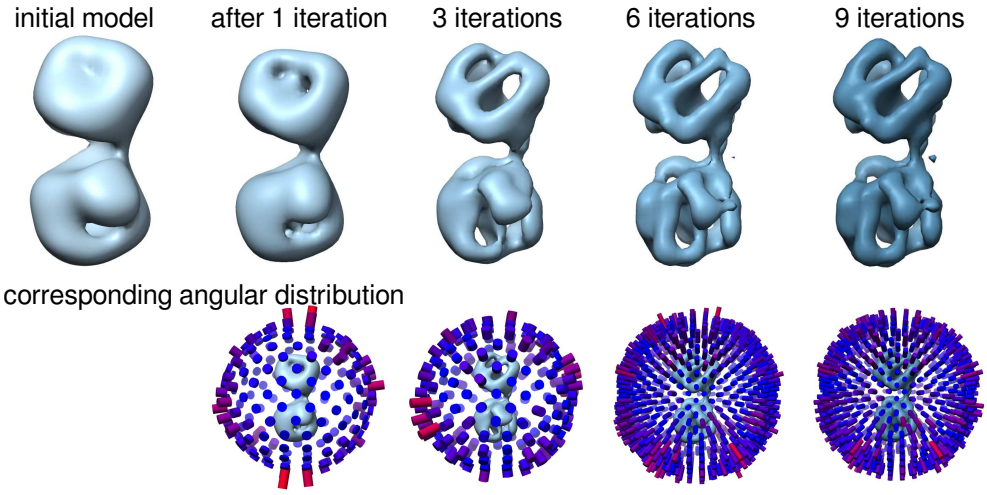


Figure 3.7: This is an example how an iterative refinement works. The blue densities show the reconstructed densities at different iteration steps and below the angular distribution of the images is shown. The angular steps are getting smaller if the resolution of the map gets higher.

An example of how this looks like is given in Figure 3.7. As we can see in this example the initial model is already quite close to the correct shape. So the refinement converged really fast to the optimal solution.

Bayesian Approach for Iterative Refinement An introduction to maximum-likelihood methods (Bayesian approach) is given by Sigworth [74] and Scheres [64].

If we want to solve the problem of finding the correct density reconstruction a solution has to be found such that the probability that our reconstruction is correct dependent on the given cryo-EM data is maximized. If we have a model density maps Θ we know exactly how the experimental data X can be calculated. This relation is described properly by Equation 2.11. So the probability $P(x|\Theta)$ can be calculated easily. But what we are interested in is the inverse problem. We want to know the probability that our reconstructed model is correct for the given cryo-EM images which is $P(\Theta|X)$.

The relation between those two conditional probabilities is given by Bayes Theorem:

$$P(\Theta|X) = \frac{P(x|\Theta)P(\Theta)}{P(X)}. \tag{3.6}$$

The term $\log(P(x|\Theta)P(\Theta))$ is maximized due to computational convenience since $P(X)$ is constant and not changing for different parameters. We call the $P(x|\Theta)$ likelihood and this is what we want to maximize and $P(\Theta)$ is the prior information (a density map from the iteration before). To calculate the likelihood we need to know the orientation parameters (Euler angles and translational shifts) for each image, this information is called Y. The likelihood independent from Y is given over the sum of all possibilities and can be calculated as

$$L(\Theta) = P(X|\Theta) = \int_Y P(X|Y,\Theta)P(Y|\Theta)d\phi = \log(L(\Theta)). \tag{3.7}$$

The probability $P(X_i|\phi,\Theta)$ is given by

$$P(X_i|\phi,\Theta) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^2} exp(\frac{|CTF_{ij}[P_\phi V] - X_{ij}|^2}{-2\sigma_{ij}^2}). \tag{3.8}$$

The number of images is i, the whole probability is the product over all images i. Since all the calculations are done in Fourier space, j is the Fourier component. The frequency dependent noise σ_{ij} is calculated directly from the data and is determined in classes of images from the same micrograph by subtracting projections from images. Especially this probability is extremely computational expensive. Therefore, strategies for domain reduction since a lot of terms contribute nearly zero to the probability.

The probability $P(\Theta)$ is given by

$$P(\Theta) = \prod_l \frac{1}{2\pi\tau_l} exp(\frac{||V_l||^2}{-2\tau_l^2}) \tag{3.9}$$

Therefore the power of the signal τ in the reconstruction is calculated by the gold-standard FSC. Since the signal is smooth in real space, the signal in Fourier space is Gaussian distributed.

The new volume is reconstructed with an integrated Wiener filter and images are weighted due to the determined probabilities.

Initial Model An initial model can be a density map of a further reconstruction or a density map calculated from a structure determined by a different technique. Initial models for data where nothing is known about the structure can be created e.g. in EMAN2. The program creates a random blob and assigns the class averages to this blob. Many different blobs and initial models can be created. The reconstruction for which the class averages fit the projection of the reconstructed initial model best is usually chosen [48]. As mentioned before initial models can also build from class averages by the common lines approach [49].

A good initial model is crucial for the result of the iterative Refinement. It is highly recommended to filter the initial model to a low resolution e.g. for ribosome data to 60Å to avoid any model bias. Wrong features in the start structure procreate and gain more and more "signal" in the course of the refinement.

3.4 Validation and Interpretation

Since we are dealing with very much noisy data in the single-particle cryo-EM, it is of utmost importance to pay attention to some things in order not to make any mistakes when processing the data and interpreting the reconstructed density maps. This section therefore presents techniques for validating the image processing, determining the resolution of density maps and interpretating density reconstructions by fitting models into density maps.

3.4.1 Model Bias - Einstein from Noise

First we want to present an example of a SPA work from the year 2013 which was strongly discussed and criticized in the cryo-EM community. It is about the determined structure of HIV-1 envelope glycoprotein trimer which was determined to 6Å resolution [50]. The density map of this complex from a previous work (11Å) was used to pick particles from the micrographs. For this purpose, projections of this density map were calculated and used as references. This work and the reconstructed data could not convince other scientists.

There were several responses from experts in the field. One was van Heel [86] who demonstrated how it is possible to get Einstein from noise class averages, by searching this pattern in random noise micrographs as it is shown in Figure 3.8. In A) he shows the four references he is searching for and in B) the noisy "micrograph". If you have enough particles, you can reproduce all references in class averaging as it is shown in D).

Subramaniam had the same opinion and asks the author to make his original micrographs publicly accessible [77] because the "electron micrographs do not provide convincing evidence for the presence of molecular images of HIV-1 Env trimers." "It is the combination of small (or invisible) particles low (or zero) contrast, and the use of automatic particle (or non particle) picking procedures that is most dangerous." was the reply by R.Henderson[37]. To overcome this problem he suggest to record images "at large defocus with relatively high-dose exposures, such as $80, 100$ or even up to $140\, e^{-}/\text{Å}^2$ ", so that particles become visible by eye.

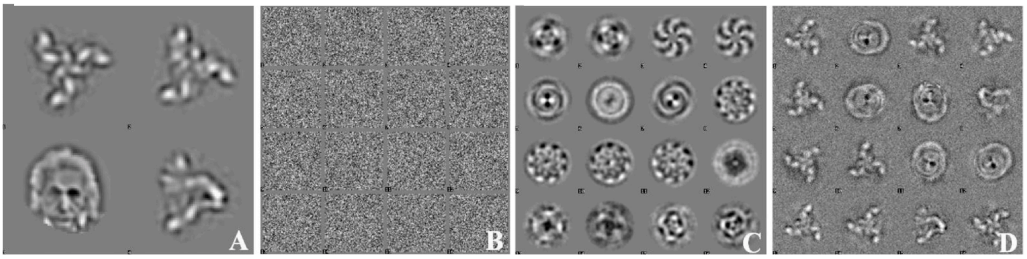


Figure 3.8: A) shows four images used for searching particles in pure noise. B) shows the picked particles. The eigenimages of picked particles are shown in C) and classes in D). All four references can be reconstructed back by total noise[86]

3.4.2 Resolution Criteria

For studying biological complexes with Cryo-EM the electrons are typically accelerated up to 100 - 300 keV which leads to wavelengths up to 0.19Å (300 kV). In contrast to light microscopy the resolution is not limited by wavelength, one limiting factor are the aberrations of the lenses and the sensitivity of the specimen. High energy electrons destroy the specimen so only low electron doses can be used. This ends up in images with a very low signal-to-noise ratio (SNR).

In cryo-EM the resolution of a density map is defined by the frequency at which the density values of two reconstructions each consisting of half of the images are correlated in Fourier space up to a certain value of 0.143 [88] [25]. Important is that the data set has to be splitted before the angular refinement which is the “gold-standard” criteria [70]. The assumption which is made is that the signal in the two reconstructions is correlated whereas the noise is uncorrelated. The correlation is measured by the Fourier shell correlation (FSC), defined by

$$\mathrm{FSC}(r) = \frac{\sum\limits_{r_i \in r} F_1(r_i) F_2(r_i)^*}{\sqrt{\sum\limits_{r_i \in r} |F_1(r_i)|^2 \sum\limits_{r_i \in r} |F_2(r_i)|^2}}. \tag{3.10}$$

This function calculates the frequency dependent similarity of two objects (in this case density maps). The two objects are Fourier transformed ($F_{1/2}(r_i)$) and then the normalized scalar product of resolution shells is calculated. This ends up in a function which looks typically like in Figure 3.9.

The FSC (blue curve) shows the correlation of both maps for different shells in Fourier space. For low frequencies both maps have a correlation of 1, which means they are identical. If we go to higher frequencies the correlation decreases more and more until it falls ultimately to zero. High similarity between the two maps can arise from “real” correlated signal

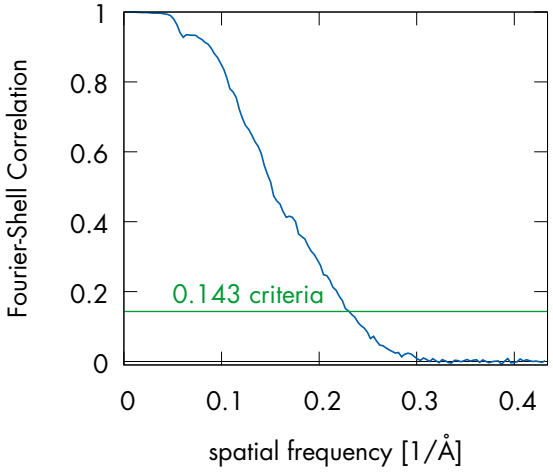


Figure 3.9: Example for a typical Fourier Shell Correlation.

or from systematically image processing errors applied to both data sets which is also a critical point in this resolution definition. Spikes in the FSC can originate if certain frequencies were suppressed due to a low range of different defoci. So we would read out the value at which the FSC crosses the 0.143 threshold and take the reciprocal to get the resolution. In this example in Figure 3.9 the resolution is around $4.2\,\text{Å}$.

The 0.143 criterium When one agreed on the FSC calculation for resolution determination, the resolution was first read off at a correlation value of 0.5, but this definition had changed to 0.143. This value was proposed by Rosenthal and Henderson in 2003 [63]. They show that the FSC of a density map with a perfect reference FSC_{ref} and the FSC of the whole map FSC_{full} can be expressed by the normal FSC between the half maps as:

$$\text{FSC}_{\text{ref}} = \frac{\sum |F_1||F_{ref}| \cos\Delta\phi}{\sqrt{\sum |F_1|^2 \sum |F_{ref}|^2}} = \sqrt{\text{FSC}_{\text{full}}} = \frac{2\text{FSC}}{1+\text{FSC}}. \tag{3.11}$$

The suggestion was to chose a FSC_{ref} value of 0.5 as resolution criteria, which corresponds to a normal FSC value of 0.143. The argumentation is that this also corresponds to a $\Delta\phi$ of $60°$ by Equation 3.11 which is equivalent to the phase error definition (figure of merit) in X-ray crystallography and describes when a map is regarded as interpretable. This means that the different resolution definitions in X-ray crystallography and cryo-EM should be more comparable by this definition.

Local Resolution

Often the resolution of the density map is higher resolved in the center than on the outer parts, one subunit is much better resolved than the

other. These effects can be measured by calculating the local resolution of a density map with ResMap [43]. ResMap estimates the noise by the region surrounding the particle or two half maps. The programs calculates the resolution for each voxel. The local resolution can be estimated for maps between 4 and 40 Å resolution.

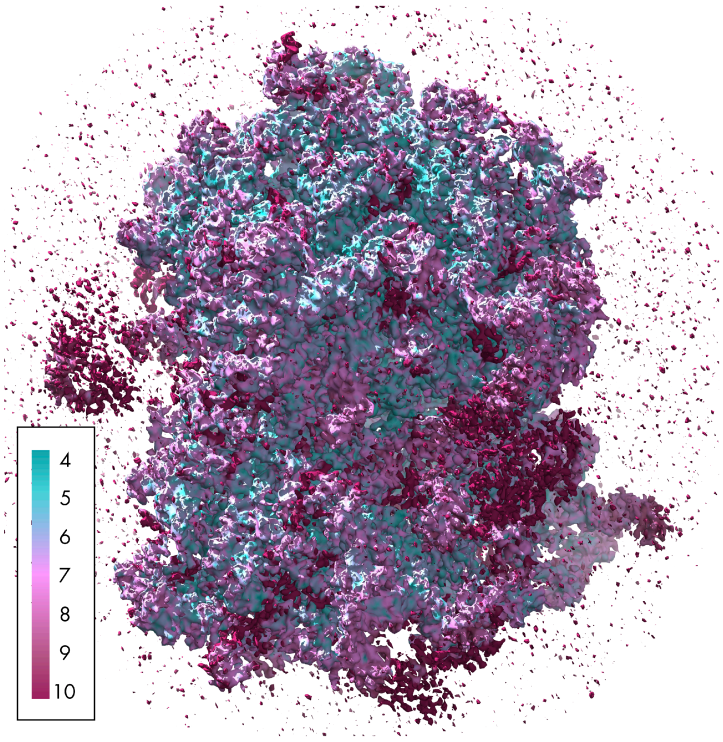


Figure 3.10: Local resolution of ribosome density map calculated with Resmap. (On this ribosome reconstruction we applied our new sorting approach on principal motions in Chapter 7).

An example of a local resolution calculation is given in Figure 3.10. The local resolution was calculated on the raw, unsharpened density reconstruction. But is shown on the sharpened map for a clearer view. Most parts of the density are at a resolution of 4 Å. Here we changed the threshold such that also regions with a much lower density became visible. Often central regions are much better defined due to higher angular accuracy and to lower flexibility. The local resolution is a good measure to check regions of interest for resolution and to gain an idea where areas that are for whatever reason less resolved.

3.4.3 Postprocessing

Due to the suppression of high frequency information as described by the B-factor in Section 2.2 the raw reconstructed density map does not show the full information which it contains. But still the particle density should be clearly visible over the noise. A way to recover the effect of the

B-factor, if the B-factor of Equation 2.9 is determined, is to multiply the Fourier transform by this function with a positive sign in the exponent. The B-factor can be determined by a Guinier plot as proposed by [63].

If an atomic model for the density map is known e.g. X-ray crystallography the structure factor of a density model calculated from it can be applied to the reconstructed map. This will sharpen the density map.

Since I also worked on a postprocessing method to sharpen and enhance the overall representation of density maps, I will refer for more information and for our new method for postprocessing (VISDEM) to Chapter 4.

3.4.4 Interpretation by Modeling

It is not only important to see whether the FSC curve shows a high resolution but also to check if the density map looks reasonable. If we assume that we did everything correct and we end up at $15\,\text{Å}$ we should be able to distinguish between different domains of a complex. At this relatively low resolution it is possible to rigidly fit X-ray structures into the different domains to get an idea about the conformational state and the overall arrangement. This can be done e.g. in Chimera [30].

Fitting X-ray structures is a useful tool to interpret density maps, since in the end the goal is to learn something about the structure. A flexible fitting makes sense for density maps lower than $15\,\text{Å}$. At a density resolution lower than $9\,\text{Å}$ secondary structure elements become visible. Alpha-helices are shown as rods and beta-sheets as more planar objects. Features in secondary structure elements get more pronounced around 5-$6\,\text{Å}$. Pitches are arising from helices. Around $4.8\,\text{Å}$ beta-strands starting to separate. If an X-ray structure is available for the particle/parts of the complex this structure can be fitted into the density map. This is done flexibly since the conformation of the X-ray structure and the one which is presented in the density map are usually not identical. There are many different programs available which use different approaches.

One is MDFF which fits the structure by running an MD simulation plus an additional force which pulls the atoms into the density map [82]. Another one is iMODFIT which calculates normal modes and uses them to find a fit for the map [47].

MDFF has the disadvantage that a MD simulation is really computationally expensive and the fitting can only be done in very small time steps. It could be that the structure will be stuck in a local minimum or will move into the wrong direction first. The iMODFIT uses normal modes and since the normal modes are not sufficient to describe the correct conformational state represented by the density map will not work. Another option is the flexible fitting in DireX which will be explained in the next subsection since it was used in all my calculations.

If the resolution of the density map is higher than $4\,\text{Å}$ it is possible to build a model de novo. This can be done by programs like COOT [22] or Phenix [3] which were developed in the first place for solving X-Ray structures.

3.4.5 Flexible Fitting with DireX

We are working in our group with a program for flexible fitting which was developed by G. F. Schröder [72].

An atomic model or also just a backbone trace can be fitted into the density map. Therefore the atoms have to move to high electron density regions. The program calculates a density map from the atomic model. For each atom a stochastically density gradient $\bar{g}_i$ is calculated on the difference map ρ_{diff} between target and atomic model density map,

$$\bar{g}_i = \nu_{s_c} \frac{1}{12} \sum_{j=1}^{12} \rho_{\text{diff}} \frac{\bar{r}_j - \bar{x}_i}{\bar{r}_j - \bar{x}_i}. \tag{3.12}$$

The gradient is calculated by 12 random vectors $\bar{r}_j$ around the atom position $\bar{x}_i$.

It is important that the stereochemistry of the structure is preserved during the refinement. DireX is based on an algorithm used in a program called Concoord [34]. Concoord randomly shifts the atoms of a protein into arbitrary directions, after this it corrects for pairwise interatomic distance restraints. So, two atoms in a protein have usually a certain distance to each other due to experimental observations. If the distance between two atoms due to the distortions got higher or lower than in a certain allowed interval both atoms are shifted for the same amount towards/outwards each other. The corrections are done iteratively. For a small protein this converges to a structure in a different conformation than the start structure after 100-300 iterations.

In DireX three forces act on the atoms: the electron density map, the Concoord restraints and the DEN restraints which are the key point of this method. DEN stands for deformable elastic network. Between atoms in a certain distance are randomly distance restraints refined. These atoms are connected by a harmonic potential. The effect of this can be explained on a small example shown in Figure 3.11. The density map is shown in a green mesh and the structure which we want to fit in orange. While fitting the structure the secondary structure information (helix) should not be destroyed.

The idea is to shift the atoms towards the density map, keep the stereochemistry and also secondary structure elements. DEN restraints are shown as black springs in Figure 3.11.

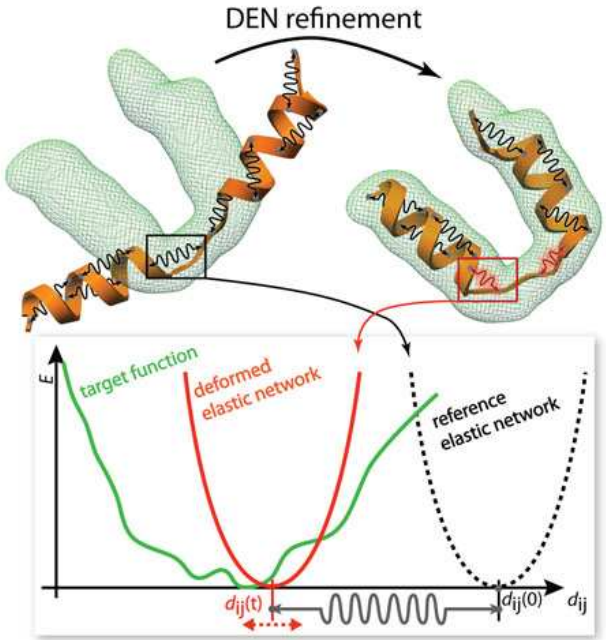


Figure 3.11: The upper picture shows an alpha helix in orange and a density of a kinked helix in a green mesh. The helix is flexible fitted by randomly distributed springs (in black) into the green mesh. In the lower picture we show the energy potential of the spring in black and the target position which is given by the density map in green. The spring adjusts during the fitting to the red curve which is much closer to the target. (Figure from `schroderlab.org`)

During the refinement when the helix in Figure 3.11 adapts to the density map the equilibrium distance d^0_{ij} of the springs in the kinked regions is updated after each generated structure. In the Figure 3.11 the black equilibrium distance adapted during the refinement to the red equilibrium distance. This happened in several steps and is calculated iteratively and can be described as follows

$$d^0_{ij}(n+1) = (1-\kappa)d^0_{ij}(n) + \kappa[\gamma d_{ij}(n) + (1-\gamma)d_{ij}(0)]. \tag{3.13}$$

The strength of the DEN restraints (γ-factor) can be set up from 0 not flexible at all to 1 totally flexible. The κ is scaled down during the refinement and gives some kind of weight on the DEN restraints but should not be changed by the user. All parameters and different weightings can be changed but are kind of adjusted for regular cases.

Cross-Validation An over-fitting which means fitting into noise can lead to misinterpretation of the density map. In DireX a cross-validation procedure is integrated which prevents such an overfitting [24]. The cross-validation uses a work set to which the method is applied and then calculates whether it also improved the untouched test set. In our case the two sets are

resolution shells of the density map. The work set shell contains the resolution of the density map up to infinity. The atomic model is flexibly fitted into this low-pass filtered density. The programs calculates after each iteration step the correlation between the model density and the target input density map. After some iterations this correlation usually converges to some value where the atomic model is fitted completely into the density map. This correlation value is called C_{work}. Simultaneously the program calculates the correlation to the unused resolution shell which is called C_{free}. The interval for the free set is usually chosen to a resolution slightly lower than the resolution of the map up to the resolution.

Optimally the density information used for the fitting is enough for the refinement but not too much that the model is not over fitted and the interval of the C_{free} value is chosen that the signal is still enough for validating and not too much noise. This can be tested and observing a decreasing C_{free} value usually indicates over-fitting, which can be overcome by adjusting the strength of the DEN restraints or decreasing the gamma value.

4 Improving the Visualization of Cryo-EM Density Reconstructions

This chapter is about a method for sharpening of cryo-EM density maps. The work was inspired by the IEEE - Signal processing cup challenge in 2014. My colleague Amudha Kumari Duraisamy and I started to work on this challenge data together. The data consists of density maps in which the signal should be increased and the noise decreased which was measured by Fourier-shell correlation of the given density maps to a high resolution density map calculated from the correct underlying structure.

Why do the density maps have to be sharpened? A reconstructed density map always contains errors due to misaligned noisy images. High frequency information is suppressed by the envelope of the CTF, which is further explained in Chapter 2.3. Usually the density map will be sharpened which means to improve the visualisation and increase values of high frequencies in Fourier Space.

Very often different parts in the density map are not equally good resolved. This means some regions have much more signal then others, this makes it difficult to interpret the entire structure. It happens for example often that the small subunit in a ribosome is worse resolved than the big subunit. This occurs because the images are mostly aligned on the thicker part with the higher signal. With our method we can sharpen the whole map and also balance out the density value level difference which improves the visualisation, but additionally compared to other methods we also improve the resolution of the map.

We call our method VISDEM as abbreviation for **v**isualization **i**mprovement by **s**tructure factor and **de**nsity histogram correction of cryo-EM **m**aps. We published our results in Journal of Structural Biology [76]. We both are the first authors and contributed equally to this work.

The program we developed (dx_visdem) comes with DireX (Low Resolution Structure Refinement) and can be downloaded from `https://simtk.org/home/direx`.

4.1 Abstract

Cryo-electron microscopy yields 3D density maps of macromolecules from single-particle images, tomograms, or 2D crystals. An optimal visualisation of the density map is important for its proper interpretation. We have developed a method to improve the visualisation of density maps by using general statistical information about proteins for the sharpening process. In particular, the packing density of atoms is highly similar between different proteins, which allows for building a pseudo-atomic model to approximate the true mass distribution. From this model the radial structure factor and density value histogram are estimated and applied as constraints to the 3D reconstruction in reciprocal- and real-space, respectively. Interestingly, similar improvements are obtained when using the correct radial structure factor and density value histogram from a crystal structure. Thus, the estimated pseudo-atomic model yields a sufficiently accurate mass distribution to optimally sharpen a density map.

4.2 Introduction

Cryo-electron microscopy (cryo-EM) is a powerful technique to determine the structure of large macromolecules. In cryo-EM, a three-dimensional density map is reconstructed from a series of single-particle 2D images, tomograms or sub-tomogram averages, helical reconstructions or 2D crystals. An optimal visualisation of the reconstructed density map is important for its interpretation, and, if the resolution of the density map is high enough, for atomic model building. However, density reconstructions often suffer from distortions or artifacts which arise from both electron optics as well as the density reconstruction process.

For example the contrast transfer function modulates the amplitudes of the structure factor, and in particular the envelope function of the transfer function dampens higher-resolution features in the density map. To improve the visualisation of the reconstruction, typically B-factor sharpening [20, 25] is used to amplify high-resolution features in the density map. Since B-factor sharpening only affects the radial structure factor amplitudes, i. e., all structure factors within the same resolution shell are scaled by the same factor, this method improves the visualisation but by definition not the resolution. A more advanced regularized deconvolution technique has recently been developed to address this problem [35].

Another density artifact that originates from the reconstruction process is a blurring of the particle periphery due to uncertainties in the angular assignment of particle image orientations. This leads to a density that is increasingly reduced and smeared out the farther away it is located from the center of the particle. Recently, a spherical deconvolution algorithm has been developed to reduce this blurring [42]. But as for B-factor sharpening,

the resolution as determined by the Fourier Shell Correlation (FSC) is not affected by this deconvolution approach. A similar correction can be done also on 2D class-averages [55].

Density modification procedures such as for example solvent flattening [93] (or solvent flipping [1]) are often used to improve phases of electron density maps obtained from X-ray diffraction experiments [16]. For solvent flattening a mask needs to be determined, which defines the region that is covered by the molecule (or by the solvent). In addition the expected histogram of density values is often used as a constraint on the density map [96]. At high-resolution, where the density shows atomicity of the molecular structure, this histogram is rather shape-independent and depends mainly on the resolution (Fourier cut-off) of the density map. Therefore, in X-ray crystallography typically a generic resolution-dependent density histogram is applied.

Here we present a technique for improving the visualisation of cryo-EM density maps by applying two constraints: 1) in Fourier space on the radial structure factor, and 2) in real-space on the density value histogram. The estimates of the radial structure factor and the density histogram are obtained from an approximate model of the mass distribution within the reconstructed particle density. This approximate model is built using only the density reconstruction and statistical information on proteins calculated from the Protein Data Bank (PDB); no further knowledge about the protein structure is required. Our approach is therefore free of any model bias.

We tested our approach, called VISDEM (**V**isualisation **I**mprovement by **S**tructure factor and **D**ensity histogram correction of cryo-**EM** maps), on three different experimental density maps of fatty acid synthase, GroEL, and the transient receptor potential channel TRPV1 which have been determined at different resolutions of 18 Å, 8.9 Å, and 3.3 Å, respectively. For all of these cases either a docked X-ray structure or (for the highest resolution case) an atomic model built from the EM density was available for evaluating our results.

4.3 Results

3D density reconstruction procedures typically do not take into account the fact that one knows that the particle is, for example, a protein. However, this knowledge can provide additional information which yields restraints on the density map. The main idea of our method is to use information on the mass distribution of average protein structures to improve the visualization of cryo-EM density maps of proteins. In particular we are using knowledge of the average atomic density and average composition of elements in proteins. This provides constraints on both the radial structure factor (in reciprocal space) and the distribution of density values (in real space). For this, we need to create an approximate (pseudo-atomic)

model of the mass distribution, which we refer to as the bead model. This bead model is supposed to be a very rough approximation to the true protein structure. The beads represent only the non-hydrogen atoms; adding the weakly scattering hydrogen atoms does not change the results in a noticeable way. The shape of the particle is given by the density map and the beads are placed randomly into the density map, i. e., the beads are placed in regions where the density is above a certain threshold. The question is only which density threshold defines the correct protein boundary or volume?

As the atomic density does not vary much within a protein and is very similar in all proteins [83], the number of beads to be placed is given by the volume. However, the volume of the particle is defined by the density threshold. At high resolution (3–5 Å) the volume does not depend strongly on the density threshold and is therefore relatively well defined. With such an estimate of the volume the number of atoms can also be estimated using the known average atomic density. At lower resolution (worse than 5 Å) the volume depends more strongly on the density threshold. In that case the molecular volume cannot easily be determined and one needs to assume that the number of atoms is known.

As will be shown below, the exact placement of the beads in the volume is not critical for determining the radial structure factor and the density value histogram. But since the structure factor and the distribution of density values are not independent, one cannot set one without changing the other. We therefore modify the density by iteratively applying the constraints in reciprocal- and real-space, such that eventually a density is obtained which has a radial structure factor and distribution of density values close to the expected ones. The VISDEM method consists of the following steps:

1. Estimate volume of the particle with known or estimated number of atoms.
2. Generate a pseudo-atomic model (or "bead model").
3. Apply radial structure factor from bead model.
4. Apply density value histogram from bead model.
5. Repeat steps 3 and 4 one more time.

These steps are explained in more detail in the following.

4.3.1 Estimating the Volume

It is important to use a number of beads that is close to the actual number of atoms in the structure. Often the protein sequence is known which directly yields the number of atoms. The beads are then placed randomly at positions where the density is above a given threshold. This threshold

determines the volume covered by the bead model and therefore the average bead density. To decide which density threshold is best we compare the obtained bead distribution with the average atom density of protein structures in the Protein Data Bank (PDB). To do this, the cumulative radial distribution function (cRDF) of all atoms is calculated. The cRDF of atoms in protein structures is very similar at small distances for all proteins and then converges to the total number of atoms for larger distances. The average cRDF value for all non hydrogen atoms at a distance of 5 Å is 16.9 ± 0.85 calculated for 140 protein structures randomly chosen from the PDB. The best threshold value is the one for which the value of the cRDF function at 5 Å is closest to 16.9.

4.3.2 Estimating the Number of Atoms

Oftentimes an EM reconstruction does not show equally strong density for all atoms. Due to their high flexibility some loop regions or even entire protein domains might not be visible. We therefore propose a method to estimate the number of atoms that are visible in an EM reconstruction. At fixed density map threshold, we generate models with different numbers of beads. These bead models are then refined with the program DireX [72] into the EM density. The number of beads that yields the best cRDF is then chosen as the best estimate. This approach works only reliably for resolutions better than about 5 Å, since at lower resolution the boundaries of the protein are not well defined and the refined bead model will be blurred far outside the protein surface, which yields a cRDF curve that cannot be compared with the average cRDF obtained from PDB structures. The blurring of the refined bead model will then also lead to a wrong structure factor and would not improve the map sharpening.

The structure factor and the density value histogram strongly depend on the types of the scattering atom. We therefore randomly assign atom types to the beads in the pseudo-atomic model such that its composition is the same as the average composition observed for proteins in the PDB, i. e., 62.2% carbon, 17.2% nitrogen, 20.1% oxygen, and 0.5% sulfur. The obtained bead model is then an approximation of the real protein structure in terms of shape, mass distribution and elemental composition. If the sequence of the amino acids in the protein is known, one could directly use the correct composition of elements, but in practice this does not make a noticeable difference.

4.3.3 Matching Radial Structure Factor and Density Histogram

In the third step of the VISDEM protocol the radial structure factor of the EM reconstruction is matched to that of the bead model. First, a density map at high-resolution (typically two times the grid spacing) is created from

the bead model with the program DireX, which uses theoretical electron scattering factors approximated by a sum of five Gaussian functions [60]. Then the radial structure factor is computed from the bead density map and applied to the original EM reconstruction. For these operations we use the program *e2proc3d.py* as part of the EMAN2 software [80].

In the fourth step, the real-space density value histogram of the structure-factor corrected EM reconstruction is matched to that of the bead model. Since the density value histogram is very different for different resolutions it is important to filter the bead density map to the final resolution at which the EM reconstruction will be examined. To match the density histogram of one map A to another map B, the density values in both maps are first sorted. Then the density values of map A are assigned in the same order to the corresponding grid points of map B. The new density map B' contains the same density values as map A (and therefore has the identical density histogram), but they are assigned to possibly different grid points. This procedure is implemented in the program *dx_matchhist* which is part of the DireX package (http://www.simtk.org/home/direx).

Steps 3 and 4 of the VISDEM protocol are repeated one more time, since the structure factor has to be corrected after applying the density value histogram. We did not observe consistent improvement when performing more than this one iteration.

It should be noted that no explicit mask is applied; the density map is affected only by the radial structure factor and density histogram constraints estimated from the bead model. While the bead model itself is created by defining a density threshold, which divides the density map into particle and solvent regions, the density values in the solvent region are not set to zero (as would be done in masking). One could in principle also apply a mask (e. g. using the density threshold) and set all density values outside the mask to zero, which would further shift all FSC curves to higher resolutions, but here we wanted to demonstrate only the effect of the VISDEM sharpening procedure without the additional effect of masking.

In the case that the atomic structure is already known, e. g., by X-ray crystallography, this known structure can be used to estimate both the radial structure factor and the density value histogram to be used in steps 2 and 3 of the VISDEM protocol. We refer to this as the ideal-VISDEM sharpened map, which we show in the application examples below for comparison to the regular bead model based VISDEM protocol.

4.4 Application Examples at Different Resolutions

The method was applied to three published test cases of different resolutions taken from the EMDataBank (EMDB): The density map of Mycobacterium tuberculosis fatty acid synthase multienzyme complex (EMDB-

2538) with a resolution of 18 Å [15], the chaperonin GroEL/ES density map (EMDB-2325) with a resolution of 8.9 Å [13] and the transient receptor potential channel V1 (TRPV1) (EMDB-5778) with a resolution of 3.3 Å [46] were chosen as examples. Since there is no general rule on how EMDB deposited maps are processed, sometimes the maps are not optimally filtered or sharpened. Comparing the VISDEM sharpened maps with the original EMDB maps may thus exaggerate the improvement. We therefore show optimally B-factor sharpened EMDB density maps instead for comparison with the VISDEM sharpened maps. The B-factor sharpening was done by simply multiplying the Fourier components with $\exp(-\mathrm{B}\,\mathbf{s^2})$ using a negative value for B; no further weighting was applied.

In the first example, the density map of fatty acid synthase (FAS) was used to test VISDEM. For FAS, the cRDF was calculated from a bead model generated from the FAS density map at different threshold values. The number of beads was set to 125,670, which corresponds to the expected number of non-hydrogen atoms. The optimal density threshold was then found by comparing the cRDF of the bead model to an average cRDF obtained from a random selection of PDB structures (cf. Fig. 4.1(b)). The cRDF curves for different density threshold values were calculated and the optimal value of 3.0 yielded a cRDF curve (red) closest to 16.9 at 5 Å. This optimal density threshold corresponds to a volume of 2760 nm^3.

With the known molecular weight of 1.98 MDa and assuming an average partial specific volume of 0.714 ml/g [83] the expected volume is 2350 nm^3.

The FSC was calculated to evaluate the similarity between the atomic PDB model and the differently sharpened density maps. For this a 3 Å density map was computed from the atomic PDB model. Figure 4.1(c) shows the improvement of the FSC curve of the VISDEM sharpened map (blue curve) compared to the FSC curve obtained for the original EMDB deposited density map. Interestingly, the FSC curve for the VISDEM sharpened map is almost identical to the ideal-VISDEM sharpened map, where the radial structure factor and density value histogram were computed from the PDB model instead of from the bead model. The VISDEM sharpened map (cf. Fig. 4.1(a), blue) shows more pronounced features on the periphery of the structure whereas the density of the B-factor sharpened map (orange) falls off more quickly towards the outside. The same trend can also be seen when comparing the sharpened maps with a density map computed from the PDB structure filtered to 18 Å (cf. Fig. 4.1(a), green).

As a second example GroEL is shown in Fig. 4.2. To calculate the bead model a threshold value of 0.9 was determined which yields a value closest to 16.9 at 5 Å in the cRDF. We used this value for further calculations even though a slightly better result could be achieved with a different threshold value of 1.1. As in the example for FAS, FSC curves were calculated for

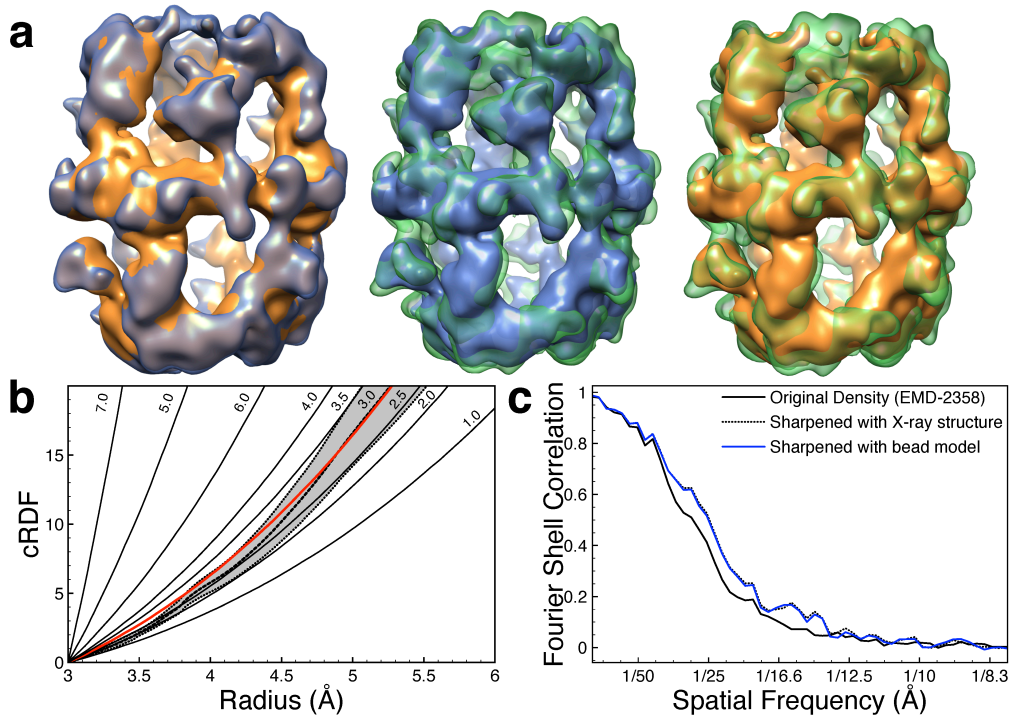


Figure 4.1: VISDEM was tested on the density map of the fatty acid synthase (FAS) protein from the EMDataBank (EMDB-2358) with a resolution of 20 Å. (a) The density map calculated from the X-ray structure (transparent green) is compared to the B-factor sharpened EMDB density map (orange) and the VISDEM sharpened map (blue). Density thresholds are chosen to enclose the same volume in all three maps. (b) The cumulative radial distribution function (cRDF) is plotted against the radius. The cRDF curves (black lines) of the bead models were calculated for different density threshold values. A density threshold of 3.0 (red cRDF curve) yields the best agreement of the bead model cRDF curve with the cRDF curve averaged over 140 randomly chosen PDB structures (black dashed line) at a radius of 5 Å. The $\pm 1\sigma$ cRDF curves are shown as dotted lines. (c) The Fourier shell correlation (FSC) was calculated between the deposited PDB structure and the original EMDB density map (black line), the VISDEM sharpened map using the bead model (blue line), and the ideal-VISDEM sharpened map (dotted line). The improvement in the FSC is the same for the bead model as for the X-ray structure sharpened VISDEM maps.

GroEL between a density calculated from the X-ray structure (PDB ID 3zpz) and the original EMDB map (black), the VISDEM-sharpened map (blue) and the ideal-VISDEM sharpened map (black dotted curve) which are shown in Fig. 4.2(c). Again, the FSC curves indicate better agreement of the VISDEM sharpened maps with the X-ray structure. The differently sharpened density maps of GroEL/ES are shown in Fig. 4.2(a). The outer regions in the VISDEM sharpened density map (blue) have overall stronger density and are on a more similar scale as the central regions of the particle

when compared to the B-factor sharpened map (orange). Also visually, the VISDEM sharpened map shows more similarity to the density map computed from the X-ray structure (green).

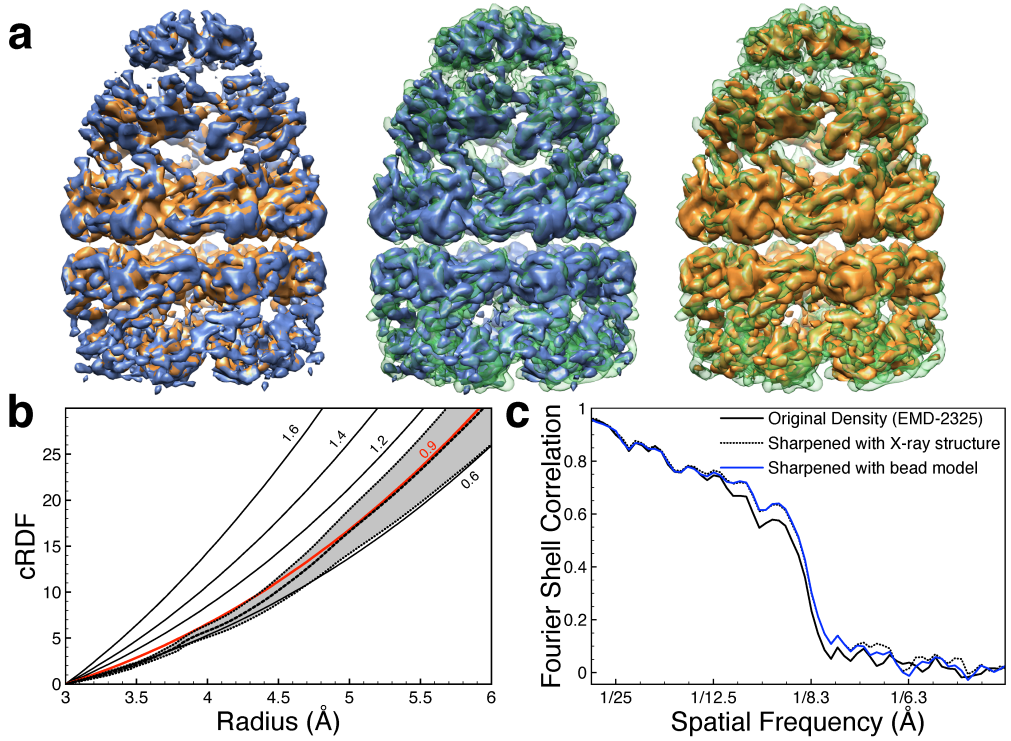


Figure 4.2: Results of the VISDEM sharpening protocol are shown for a GroEL/ES density map (EMDB-2325) with a resolution of 8.9 Å. (a) Density maps of GroEL created for X-ray structure (transparent green), sharpened by VISDEM (blue) and B-factor only (orange). (b) Cumulative radial distribution functions of the bead models (black curves) are shown for different threshold values. The optimal threshold of 0.9 yields a curve (red) closest to the PDB average (black dashed line). (c) FSC curve calculated between X-ray model map, filtered to 7 Å and original map (black), VISDEM bead model sharpened map (blue) and VISDEM X-ray sharpened map (black dotted).

In the third example, the results for TRPV1 are shown in Fig. 4.3. The outer domains of the tetramer show reduced density which suggests that not all atoms are equally well visible in the density map. Since the resolution of 3.3 Å is very high, the number of (visible) atoms can be estimated directly from the density map. For this, bead models with different numbers of beads were created which were then refined into the density without any restraints between the beads using the program DireX. At this resolution the different refined bead models have a similar clearly defined boundary and thus enclose a similar volume, the only difference is the number of the beads. The best cRDF curve was obtained with 12,500 beads, which yields a volume of 220 nm^3. Whereas the volume estimated from the molecular weight of 245 kDa is 290 nm^3, which suggests that 25% of the atoms are

not clearly visible in the density. These atoms likely belong to flexible domains that are not well resolved in the density.

Interestingly, the FSC curve for the bead model sharpened map indicates better agreement with the PDB model than the PDB model (ideal-VISDEM) sharpened map. The reason is likely the fact that the PDB model includes the ankyrin repeat domains, which are not completely visible in the density, i. e., the model sticks out of the density. The density histogram and radial structure factor computed from this PDB structure therefore do not match well the (smaller) particle shown by the density. Our bead model reflects better the visible density and yields a more appropriate estimate for the density histogram and structure factor.

The black FSC curve in Fig. 4.3 (c) which we calculated between the original density map and the fitted atomic model differs from the comparison in the original publication [46] as we did not apply a mask for the calculation of the FSC.

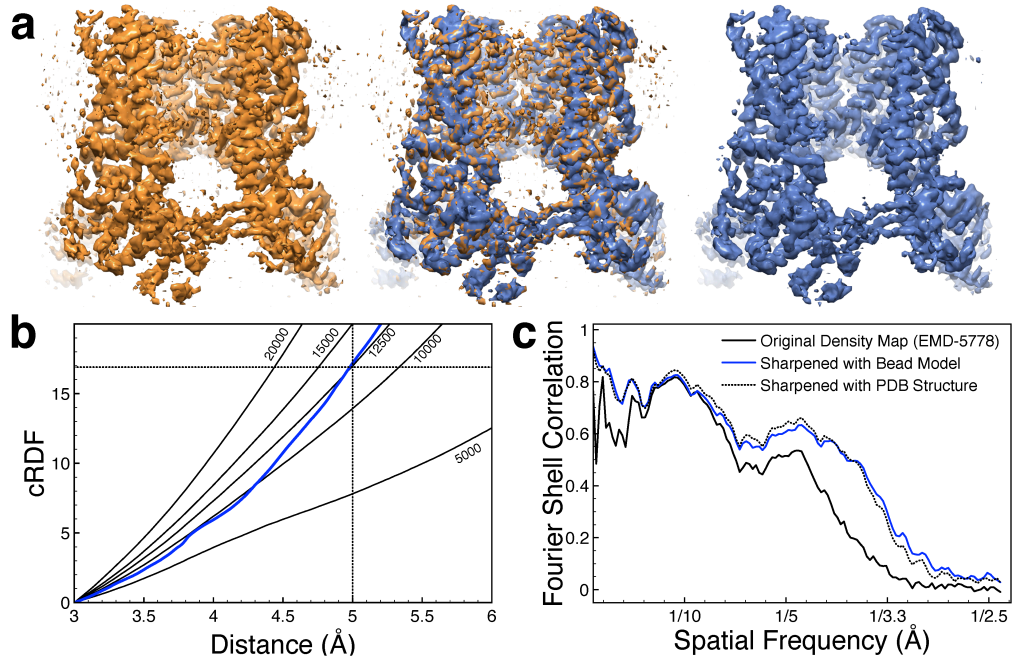


Figure 4.3: Showing results of the VISDEM sharpening protocol for the transient receptor potential channel V1 (TRPV1) (EMDB-5778) with a resolution of 3.3 Å. (a) The density contour level was chosen to yield the same contour for both maps in the central pore region. (b) Cumulative radial distribution function (cRDF) of bead models created with different numbers of beads and then refined into the density map (black curves). The bead model with 12,500 atoms has the same cRDF value as the PDB average (blue) at a radius of 5 Å. (c) FSC curve for TRPV1 .

B-factor sharpening emphasizes high-resolution features such as side-chains, but at the same time it leads to more noise all over, which results in

a smaller correlation with the PDB model. Our sharpening shows similar details as the best B-factor sharpened map, but with much less additional noise. It also puts peripheral regions on a more similar scale as interior regions (cf. Fig. 4.4) this leads to better balance of the density over the entire structure.

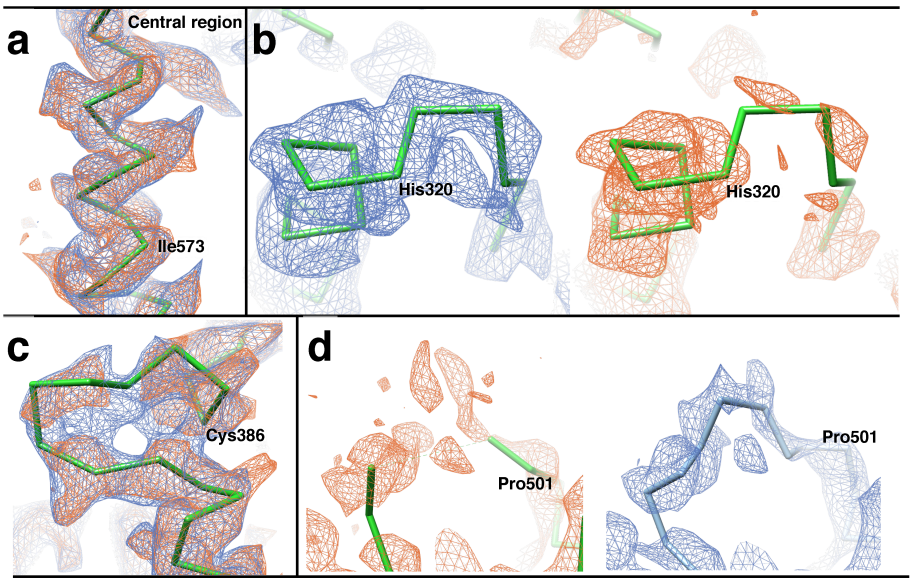


Figure 4.4: Showing details of the density improvement upon VISDEM sharpening for the TRPV1 channel. (a) The threshold was chosen to yield a similar surface for the B-factor sharpened map (orange) and the VISDEM sharpened map (blue) at the central region of the protein. The deposited PDB model is superimposed (green). For this threshold, density regions are shown in (b-d) at regions farther away from the center of the particle, showing stronger and better connected density for the VISDEM sharpened map. (d) The backbone in the loop around Pro501 could not be traced with the original density map, while the VISDEM sharpened map suggests a Ca-trace (blue trace) due to the better connected density, even if the assignment of the amino acid sequence in this loop is still ambiguous.

However, one should be aware of the fact that lower density regions are suppressed, as can be seen by the noisy density around the membrane bound region which likely comes from amphipol molecules which were used in the experiment to stabilize the membrane protein and therefore has physical meaning. The corresponding density is too weak to be filled with beads and therefore does not contribute to the structure factor estimate. The visualisation of such low-density features in the map, thus, does not benefit from the VISDEM sharpening.

To further evaluate our results the cross-correlation coefficient was calculated between the differently sharpened maps and a density calculated from the corresponding PDB structures, which served as the best possible answer. We assume that a better EM density is more similar to this best available (PDB) atomic model and we use this similarity to assess the quality of the sharpening. The different models that were compared to the PDB structure were the original density map as deposited to the EMDB, an optimally B-factor sharpened density map, the VISDEM and ideal-VISDEM sharpened maps.

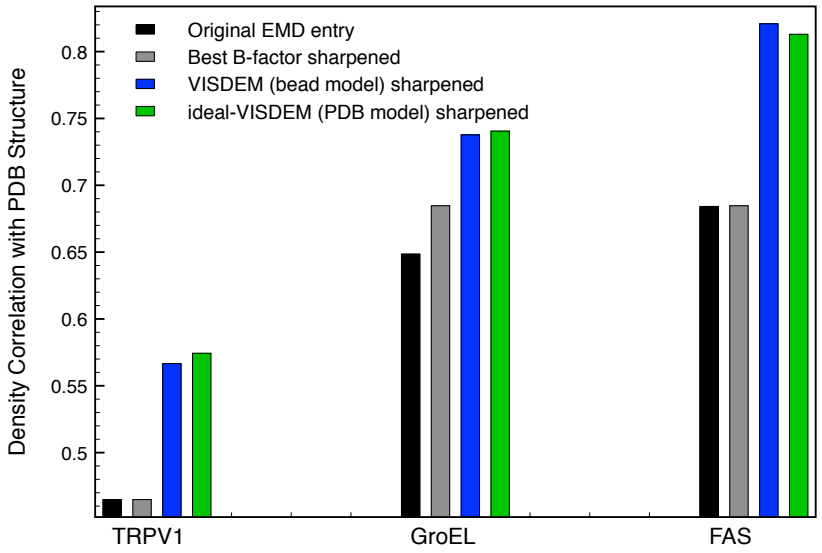


Figure 4.5: The graph shows cross-correlation coefficients for all three test cases which are always calculated to a filtered map of the corresponding PDB structures. The correlation with the original EMDB density map is shown in black, the B-factor sharpened map (-200 Å^2) in gray, the VISDEM sharpened map (bead model) in blue and the ideal-VISDEM sharpened map (PDB structure) in green.

All density maps were filtered to a resolution of 18 Å, 8.9 Å, and 3.3 Å for the FAS, GroEL, and TRPV cases, respectively, before calculating the cross-correlation coefficient. The obtained cross-correlation coefficients are compared in Fig. 4.5. From the graph it is clear that the VISDEM sharpened maps have the highest correlation coefficients when compared to both the original map from EMDB and the best B-factor sharpened map. The VISDEM sharpened maps for the bead model and the PDB model are highly similar, which shows that approximate bead model carries sufficient information to optimally sharpen the density maps and that knowledge of the "real" structure would not even improve the density further.

4.5 Discussion

We have developed the VISDEM method to sharpen cryo-EM density maps by using constraints on the radial structure factor as well as on the density value histogram. Application of this method to three examples shows clear improvement in the density map (Figs. 1-3) when compared to B-factor sharpened maps. Further, the VISDEM sharpened maps are significantly more similar to the known atomic models than the B-factor sharpened maps. This similarity was quantified by computing FSC curves and density cross-correlation coefficients between the sharpened maps and maps computed from the atomic models.

The regular VISDEM method uses a simple bead model to approximate the mass distribution of the particle, whereas the ideal-VISDEM method uses a known atomic model. Interestingly, both approaches yield very similar results, which means that the simple bead model provides sufficient information on the radial structure factor and density histogram, without the risk of any model bias.

The resolution in cryo-EM is defined by the FSC between two 3D reconstructions that were obtained from two independent half-sets of the image data. While B-factor sharpening can significantly improve the visualization of density maps it does not affect FSC curves. The VISDEM method instead does affect FSC curves and the resolution estimate in fact improves when applying the VISDEM procedure to both half maps. However, even when constructing the bead models independently for both half maps, it is not clear how the procedure of placing the beads imposes similar information on both half maps and therefore introduces correlations, which could artificially improve the FSC. We therefore suggest not to use VISDEM sharpened maps for estimating the resolution.

We have shown that the volume (optimal threshold) of a low-resolution density map can be found, if we assume that the number of visible atoms is known. And we have also shown that the number of visible atoms can be estimated from the density, if the resolution of the density map is better than 5 Å.

Density maps of macromolecules containing mixtures of DNA, RNA, and protein cannot as easily be improved by VISDEM, since nucleic acids produce stronger density than amino acids. The position of the nucleic acids influences the structure factor and density histogram and therefore the placement of nucleic acids in the map needs to be at least approximately known. The same holds true in principle for every deviation from the average atom distribution in proteins, such as for example metalloproteins.

The VISDEM sharpening procedure improves density maps of proteins over a large range of resolutions. Such a sharpened map is a better target map for atomic model building and refinement, as some of the artifacts from imaging and density reconstruction (both in real and reciprocal space)

are removed and the sharpened map therefore agrees better to density expected from an atomic model without adding a model bias.

5 Sorting Cryo-EM Images by Principal Motions

5.1 Introduction

Cryo-EM gives us the opportunity not only to determine the structure of a particle in a very specific conformational state such as X-ray crystallography but a whole range of conformational states. A cryo-EM recording always reflects a whole distribution of conformational states in which the particles in the aqueous sample actually were before they were shock-frozen. There exist methods in sample preparation to stabilize the particle, especially complexes of particles by cross-linking.However, some rest flexibility will be there. The flexibility of the whole complex or just local parts limits the achievable resolution. Very flexible parts are often not even resolved in X-ray structures. Many people have already been working on this heterogeneity problem and designed methods to estimate conformational variance contained in cryo-EM recordings and methods to split the images into classes which represent different conformational states. This is necessary to achieve high resolution density maps. Typically all the particles which do not really fit to the conformational are discarded during the refinement. The flexibility of biological assemblies or proteins often helps us to understand their function better and in more detail.

In this chapter a new approach on how to sort cryo-EM images is presented. First we want to present the method and the idea itself. Then we show how a heterogeneous cryo-EM data set can be generated. To have a system with an exactly known solution is very useful for developing the method.

5.2 Strategy

In our method we sort the images by principal motions which we estimate from the cryo-EM data. Afterwards we build models along the principal motions which are used for sorting the images into classes. How this works will be explained on schematic Figure 5.1. The images were made from calculations on a simulated test data set. How this data has been simulated and which results our method provides will be explained in more detail in the next sections.

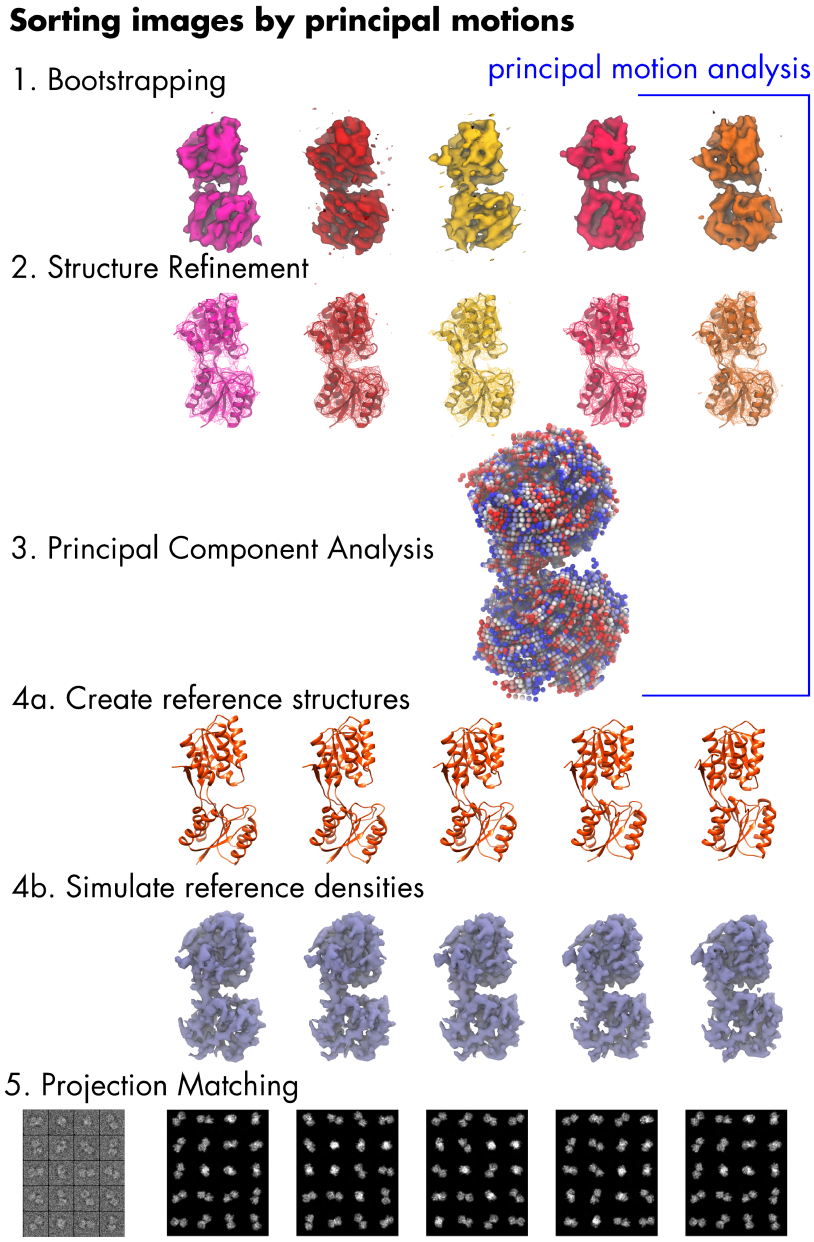


Figure 5.1: Schematic Overview of our new sorting method.

In order to sort the images by the principal motions we first have to estimate them. We call this "principal motion analysis". This method was developed by Benjamin Falkner during his PhD [23]. It is described by the first three steps of the method.

1. Bootstrapping The first step is the bootstrapping. The bootstrapping approach was introduced to the cryo-EM community by Penczek in 2002 [56]. It is a statistical resampling technique that relies on random sampling with replacement. In cryo-EM, a certain number of images are drawn randomly from the entire data set. Many image stacks are created by this

procedure. From each of the image stacks, density maps are reconstructed. Each density map is composed of a different set of images. In a heterogeneous data set with images from different conformations, each bootstrap density map contains the same number of images, but with a different weighting of images from different conformations. The variance between the slightly different density maps needs to be interpreted in terms of conformational variance.

2. Refinement The density variance needs to be described on a structural level. Therefore in the second step a structure which can be an atomic model or bead model is flexibly fitted against the bootstrap density maps. Usually the model structure factor is applied on the density maps first and the refinement is monitored by the cross-validation parameters. Since the variance between the bootstrap maps are often very small a very precise fitting has to be made.

3. Principal Component Analysis Having now an ensemble of fitted structures, the principal component analysis (PCA) is applied to the atom positions. The same approach is typically used to determine the main motion of molecules from molecular dynamics simulations. In principle, the PCA is applied by setting up a covariance matrix for the observables and diagonalize it. The eigenvectors specify the principal components. In our case, we call the principal components principal motions. We assume that the greatest variance between the structures reflects the real underlying conformational variance. Benjamin Falkner offers in his dissertation [23] a solution for calculating a PCA on very large structures with a large number of atoms.

4. Calculation of References The eigenvalues of the PCA provide an estimate of the variance of the structures. The real variance of the system is unknown. Only the direction is determined by the principal motions. The average structure is linearly shifted along positive and negative direction to generate reference atomic models From these reference models reference density maps are computed.

5. Projection Matching In the last step the images are compared to projections of the reference density maps. This is done by calculating projections in the same orientation from the density map which have been determined for the images. Each image is compared by a cross-correlation to a projection from each density map. The images are assigned to classes by the highest correlation values.

Important Point This was a brief overview of the individual points of our method. It is important to emphasize that this is not a supervised classification in the sense that we simply predict reference densities. All reference densities are defined by principal motions. These principal motions were computed directly from the cryo-EM data.

5.3 Simulating a Cryo-EM Test Data Set

For testing different types of image sorting procedures a simulated heterogeneous EM data test set is created. The flexible ribose-binding protein is used with 271 residues (2002 atoms + hydrogen). Even though the protein is too small to be visible in cryo-EM in reality the results obtained with this test system are easily comparable to larger systems. The small protein is used only to reduce computational cost. Its main motion is a swinging from an "open" to a closed conformation. The dominant conformation under normal condition is opened by $43°$. Three different conformational states from this protein can be found in the protein data base (PDB), see Figure 5.2.

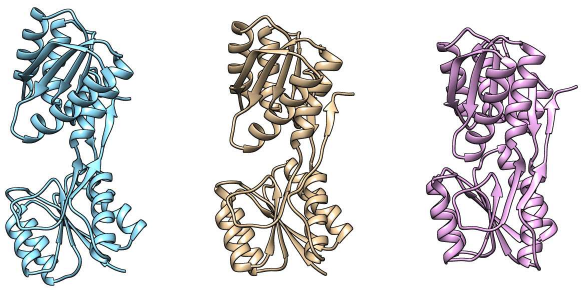


Figure 5.2: Three different conformations of the ribose-binding protein (PDB entries: 1ba2, 1urp and 2dri). The most likely conformation is the one in the middle.

5.3.1 Simulation of Projections

The workflow to realistically simulated cryo-EM images from an atomic model using equation 2.11 is shown in Figure 5.3. First, noise was added to the projections, not only after applying the CTF but also before on the clean projections. This random noise simulates instrumental noise and noise of the ice layer which is also affected by the CTF. The Fourier transformed image is multiplied by the CTF with a B-factor and then in the end additional Gaussian noise is added. The pixels in the images are binned afterwards to adapt the pixel size to the information available in the image. A too high sampling can lead to artifacts and higher computational cost and is not necessary.

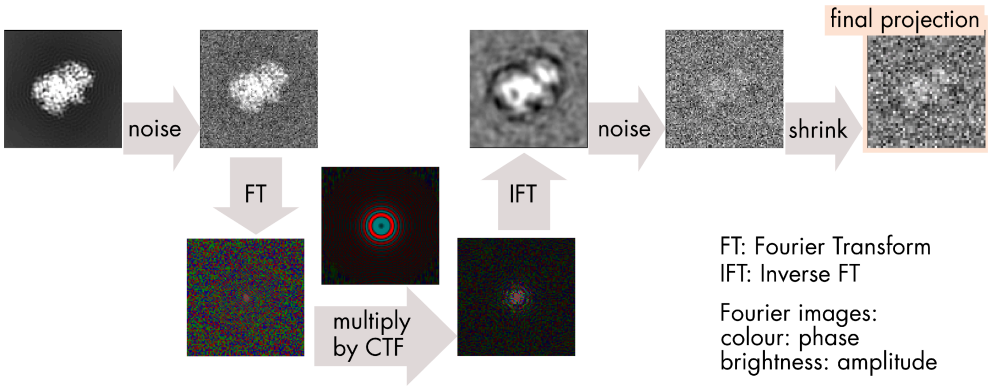


Figure 5.3: Example how cryo-EM images are simulated. Images were made in EMAN2.

5.3.2 Creating the Projections

For creating the test data set only two conformations were chosen, the middle conformation (1urp) and the closed one (2dri) as shown in Figure 5.2. The RMSD between these two structures is 4.24 Å. For both conformations were 12,000 projections from random orientations calculated. This means the whole dataset contains 24,000 images. The projections from one conformation contain respectively 5 different defoci values in their CTFs. This means

$$\mathrm{N}_{\mathrm{proj}} = 2,400 \text{ projections} \cdot 5 \text{ defoci} \cdot 2 \text{ models} = 24,000 \text{ projections}. \quad (5.1)$$

The used CTF parameter for the calculations can be seen in Table 5.1.

		set	defocus	B-factor
apix	1.5	1	0.8 µm	250
voltage	300kV	2	1.0 µm	160
cs	2mm	3	1.2 µm	290
ac	10%	4	1.4 µm	190
		5	1.6 µm	220

Table 5.1: **The parameters used for calculating the CTFs.**
Left are the fixed parameters and right are the variable parameters.

The set number in Table 5.1 indicates to which “micrograph” the images belong and under which defocus they were detected. Typical values were used for creating a realistic system.

5.3.3 Angular Refinement and Final Map

Although the correct orientation parameters for each image are known, this information will not be used here in order to make the system more realistic. Therefore new orientation parameters are determined. This is done in a few steps. First CTF parameters are determined and images are corrected. As an example how this looks images before and after Wiener filtering and Phase flipping are shown in Figure 5.4.

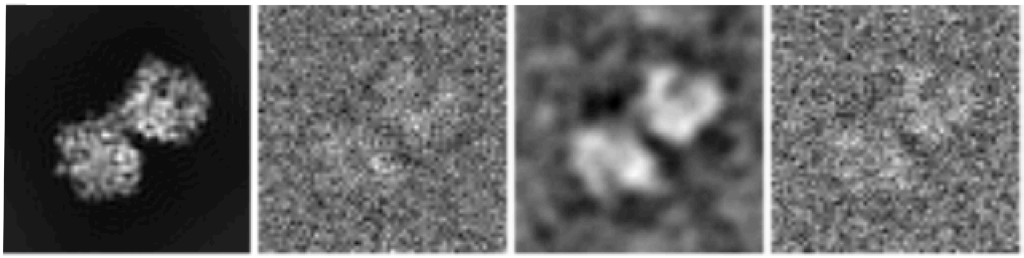


Figure 5.4: CTF correction of images in EMAN2. The first picture shows the direct projections, the next picture shows the final simulated image with noise and CTF, the third image shows the Wiener filtered image and the last one shows the phase flipped image.

The phase flipped images are used for calculating 2D classes. After this, initial models are calculated by refining the class averages to random blobs. The resulting initial models are all quite similar, so the first model is used for further steps. The used initial model and corresponding initial random blob are shown in Figure 5.5(a). All these steps were done using programs of the EMAN2 software package. The initial model fits quite well to the class averages which can be obtained by comparing projections of the model to the classes as shown in Figure 5.5(b).

To finally get the orientation parameters the images are refined to the initial model by Relion. in cryo-EM. This leads finally to a reconstructed map with a resolution of 5.1Å(see reconstructed density map in Figure 5.5(c) and corresponding FSC curve in Figure 5.5(d)).

The phase flipped particles reconstructed in EMAN2 with the correct angular information lead to a resolution of 5.3Å. The resolution is lower because it is not possible to do an accurate Wiener filter with the EMAN2 software like it is done inside Relion. The “new” determined orientation parameters in Relion are used for further calculations. Different images are weighted differently and can contribute in more than one direction to the reconstructed map with different weights. For further calculation only the maximum contribution for each image is used. The resulting density map with each image contributing once leads to a little bit smaller FSC of 5.5Å.

To make all the information in the maps visible the final map can be sharpened by different methods. The post-processing in Relion masks and

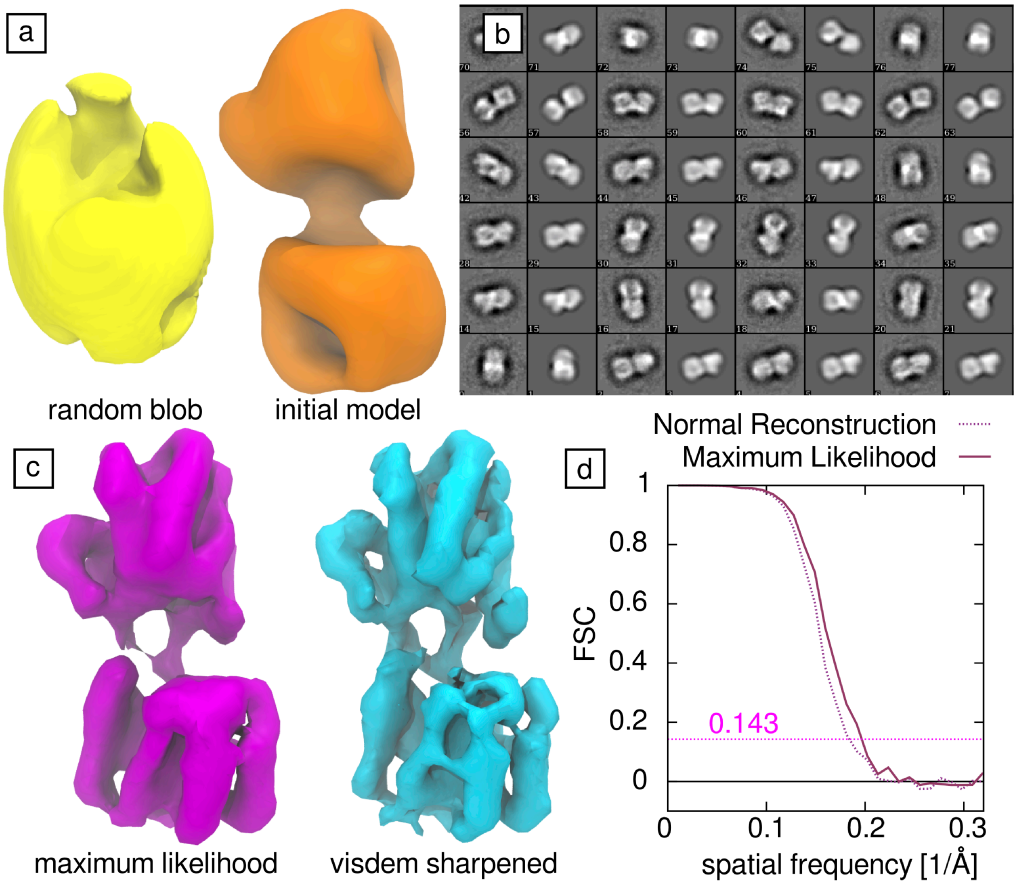


Figure 5.5: Initial Model, class averages and final reconstructed density map for 1urp test data. (a) The random blob and from this calculated initial model are shown. In (b) class averages and projections of the initial model are shown. (c) The final refined density map is shown in purple and the visdem sharpened map in blue. In (d) the FSC curves for the reconstructed density map with and without maximum likelihood approach.

B-factor sharpens the map which lead to a final resolution of 4.9Å. For comparison the final map is also sharpened by VISDEM sharpening which leads to similar results. The VISDEM sharpened map is also shown in Figure 5.5.

The refinement was done in order to create some realistic uncertainty on the orientation parameter. Since the correct ones are known the average error was calculated and is shown in Table 5.2.

	conformation 1		conformation 2	
parameter	**average error**	**standard deviation**	**average error**	**standard deviation**
space angle [°]	3.01	0.041	7.095	0.096
rotational angle [°]	3.46	0.078	7.89	0.18
translational shift [Å]	0.036	0.0006	0.147	0.0015

Table 5.2: Angular and translational accuracy on determined orientation parameters.

5.4 Bootstrapping

The bootstrapping technique was applied to the image stack to get an idea about the variance in the system. This means stacks consisting of 100 density maps are created and each map is reconstructed by the same number of images N_i. Many different ensembles are created for different values of N_i. If this number increases the resolution of each map also gets higher, but at the same time they are becoming more and more equal and the variance gets more difficult to detect. In Figure 5.6 the average resolution of the different bootstrap ensembles is plotted. It falls down until it converges slowly to the resolution of the map with all images to 5.5Å. The function falls like $47.2x^{-0.5} + 5.5$. Additionally three bootstrap maps reconstructed by different number of images are shown.

If the aim is to reveal smaller motions, a higher resolution is needed. This means the optimal number of images can be very different and At the same time, it must still be possible to distinguish the bootstrapped maps from the mean map so that a refinement is possible at all. depends on the data and on the size of the conformational change of the molecule.

5.5 Sorting with Density PCA à la Penczek

Our approach for sorting cryo-EM images is influenced by Penczeks approach [59]. He calculates the principal composition of a bootstrap ensemble directly on the density maps and determines the variance on density regions.

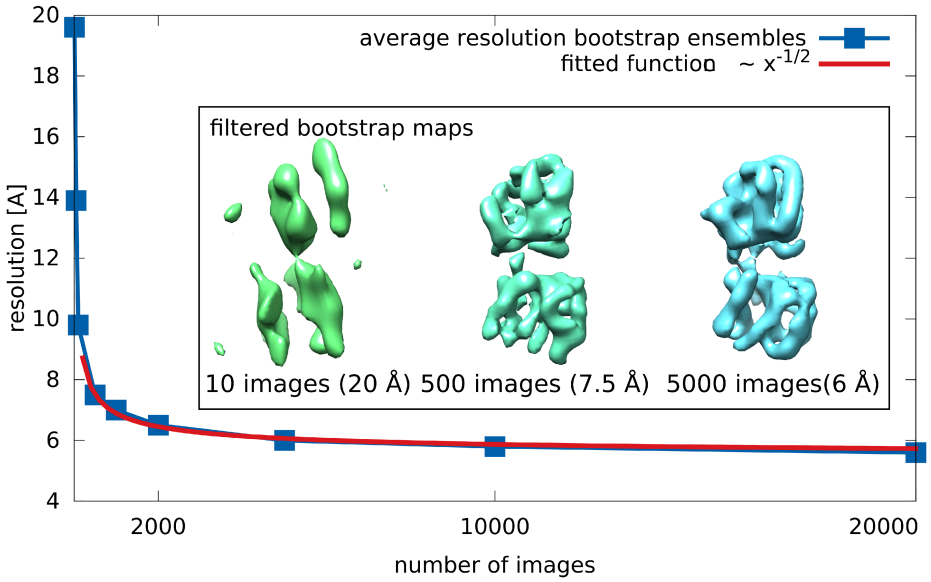


Figure 5.6: Average resolution of bootstrap ensembles and three example density maps containing different number of images.

We applied his density PCA on our simulated data. All the calculations are done in SparX. In Figure 5.7 examples of the different steps are shown, which will bes explained step by step now. The first image shows three different bootstrap maps which look quite noisy but show a lot of information by low-pass filtering. A tanh filter is used and the parameters of the filter were adapted to the FSC curves of the bootstrap ensemble. In this example each map contains just 50 images.

The second picture shows the average map around which a mask was created. The principal component analysis is applied on the bootstrap maps. In picture three the variance map is shown. But not really much can be seen. The eigenvectors have the dimension of three dimensional volumes and therefore named eigenvolumes as shown in the next picture. To get an impression of how the underlying density maps look, the eigenvolumes are added to the averaged density map. The corresponding eigenvalues contain the variance of the bootstrap ensemble and give an idea about the factor by which the normalized eigenvolumes have to be multiplied to be on the right scale for describing the different conformations.

5.5.1 Results of Sorting

The images are sorted into classes by calculating their factorial coordinates, which is practically the multiplication of the image with the projection of the eigenvolume. The factorial coordinates along the first three eigenvolumes were calculated. Afterwards the images were clustered in factorial space by k-means clustering. The whole procedure was calculated on different bootstrap ensemble in which the density maps contained different

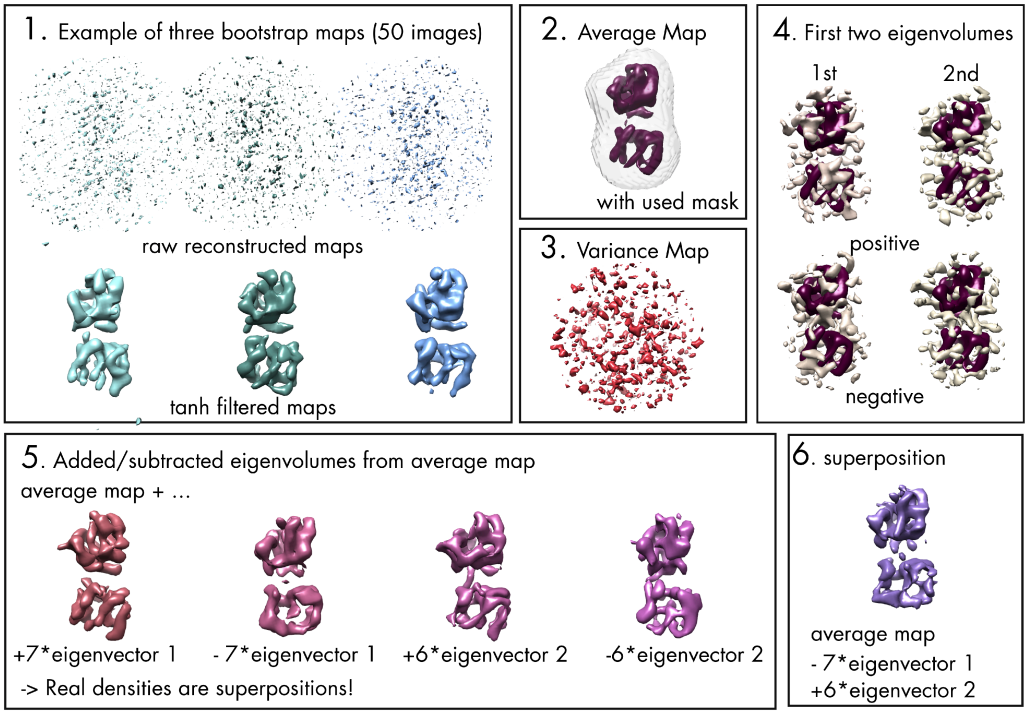


Figure 5.7: Example density maps for the different steps of density PCA (Penczek) on a bootstrap ensemble of our simulated test system.

number of images. The bootstrap ensemble where only 50 images were

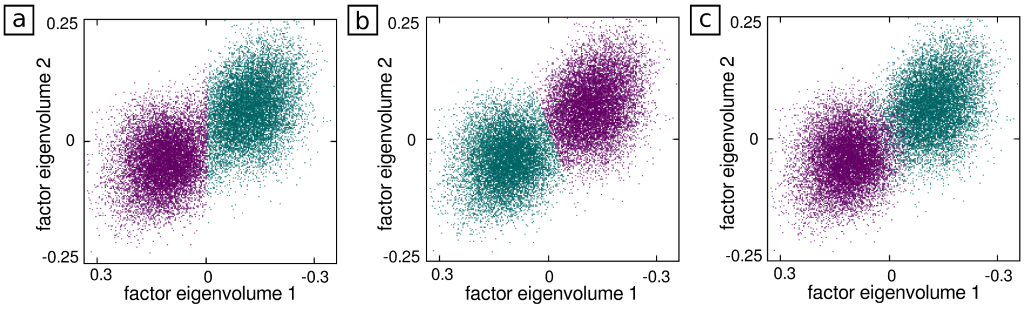


Figure 5.8: (a) Clustering images into two classes by using the first factorial coordinate.
(b) Clustering images into two classes by using the first two factorial coordinates.
(c) Real distribution of images in factorial space.

used per map worked best. In Figure 5.8 this example is shown. The first plot shows the factorial coordinates of the images as points and how they are split into two classes by taking the first eigenvolume into account. In the next plot the clustering into two classes is shown by taken also the second eigenvolume into account. The last plot shows the exact solution. This worked very well and 97% of the images could be sorted back correctly.

5.6 Sorting of Images by Principal Motion Analysis

5.6.1 Determining Principal Motions

To calculate principal motions a structure has to be fitted to all different bootstrap maps. For the flexible fitting the original pdb file of the open conformation (1urp) is used as first test case and the default parametrization of DireX. Only the resolution range and cross-validation range of the refinement is matched to the respective ensemble.

For analyzing the refinement the cross-validation parameter (C_{free} in DireX) was used. It shows that in the chosen resolution range for cross-validation the c_free parameter stayed alternating around the start value or also went slightly done. An increasing C_{free} value indicates an overfitting.

After the refinement the structures are aligned to each other, in a way that they are all shifted back to the average density. This leads to an ensemble of slightly different structures. The PCA was applied to these structures and the first eigenvector was compared to the vector between the two starting structures. This was done for all different ensembles. It looks like the eigenvector is better for smaller number of images per map. The highest scalar product was calculated for the ensembles containing 50 and 100 images per bootstrap map with 0.97. This means we have determined exactly the correct movement of the system by the principal motion analysis. Five frames along this eigenvector can be seen in Figure 5.9. The opening and closing motion is clearly visible.

Figure 5.9: Five atomic structures shifted along the first eigenvector determined by principal motion analysis.

5.6.2 Projection Matching Results

After the eigenvectors of the dominant motions are known the average structure is shifted into both directions along the first eigenvector (the best first eigenvector is used) by factor 190. This factor was used to come close to the starting structures and is usually not known but was calculated in this step in order to have a good starting point to test the projection

matching. For both structures density maps with a resolution of 5Å were calculated.

The images are now sorted into two states. For each image the orientation parameter are known and a projection in exactly the same orientation is calculated from all reference densities (in this case two). The correlation between the image and the projections were calculated.

Also the raw images were used, the Wiener filtered images and the phase flipped images. It turned out that the raw images could not be well assigned, the Wiener filtered pictures a bit better and the phase flipped images best. Therefore, three different comparators were tested: ccc, dot product and frc (Fourier Ring Correlation). The FRC did not work at all and the ccc was slightly better than the dot product.

Afterwards different modifications on the images were tested like masking, filtering or structure factor matching. This is shown in Figure 5.10.

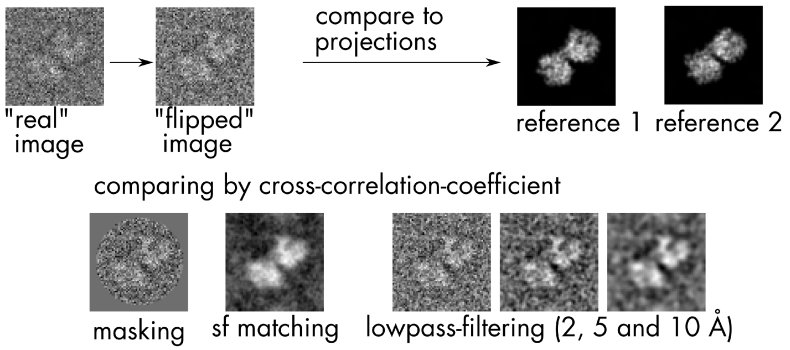


Figure 5.10: Example how the projection matching looks like the cross-correlation coefficient is calculated between the phase flipped image and the projections therefore the image is masked before, the structure factor of the projections is applied and the images are low-pass filtered to different resolutions.

The phase flipped simulated image looks quite noisy. It is not really possible decide manually which of the two conformations it belongs. However, these images can be represented by a simple cross-correlation with a probability of 90.7% as shown in Figure 5.11. A mask around the image does not increase the assignment.

Applying the structure factor significantly worsens the assignment. Low-pass filtered images lead to a slightly improved assignment. The best result with 92% was achieved by filtering the images to 10 Å resolution.

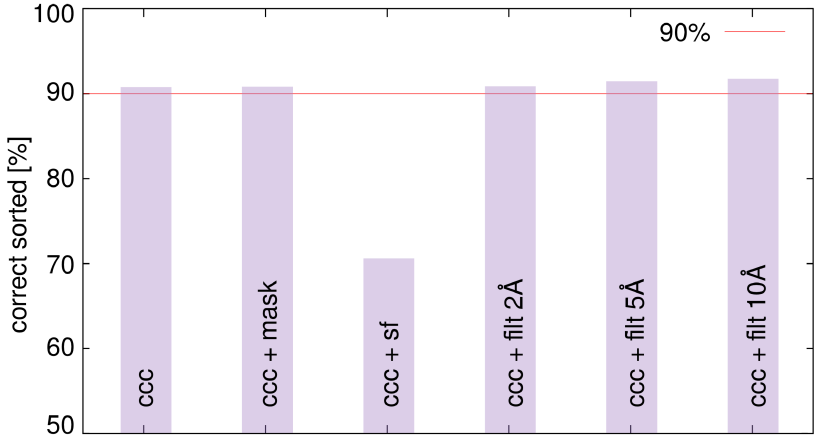


Figure 5.11: Showing the number of images which were assigned correct to the two starting conformations.

5.7 Sorting by Principal Motions with Bead Model

5.7.1 Bead Model Generation

A bead model is built inside the average density map. The beads are generated on the grid points of the density map if the density value of this point is above a certain threshold and if a bead has at least a given number of neighbor beads in a defined radius around him. Three created bead models are shown in Figure 5.12. The sampling rate was changed from every grid point, to every second and every third. In this example the beads were only placed on the grid if they have at least 13 neighbors in a distance of 13Å. The density values of the map are saved in the pdb file of the bead model as occupancy values.

Figure 5.12: Three bead models for the ribose binding protein. A sampling of every grid point (1.5 Å) , every second grid point (3 Å) and every third grid point (4.5 Å) was chosen.

To ensure that the bead model stays intact during the refinement, additional distance restraints need to be defined. These so-called DEN restraints are

placed between neighboring beads. They are calculated that each bead has at least 11 neighbors. This number was determined to give the best results by fitting a correct structure without Concord restraints, calculated in the next subsection 5.7.2 The number of beads and DEN restraints are shown in Table 5.3.

sampling	distance	# beads	# DEN restraints
1	1.5Å	5,507	579,250
2	3.0Å	691	4,840
3	4.5Å	207	913

Table 5.3: Number of beads and DEN restraints for bead models of different sampling.

5.7.2 Refinement of Bead Model

In the next step the correct protein structure is used, but no structure based restraints (Concoord restraints). The refinement is done like a bead model refinement would work. The atoms were just connected to their neighbors by DEN restraints. The goal was to determine the number of DEN restraints which are necessary in order to get a motion which is correlated enough to determine the eigenvector of the main movement. During the refinement only the number of DEN restraints per atom was changed.

The calculations show that the best results were achieved by using 11 neighbors per atom, see Figure 5.13. The scalar product for 11 DEN restraints per atom is 0.93. It is as good as before where the refinement was done in the standard mode with Concoord restraints.

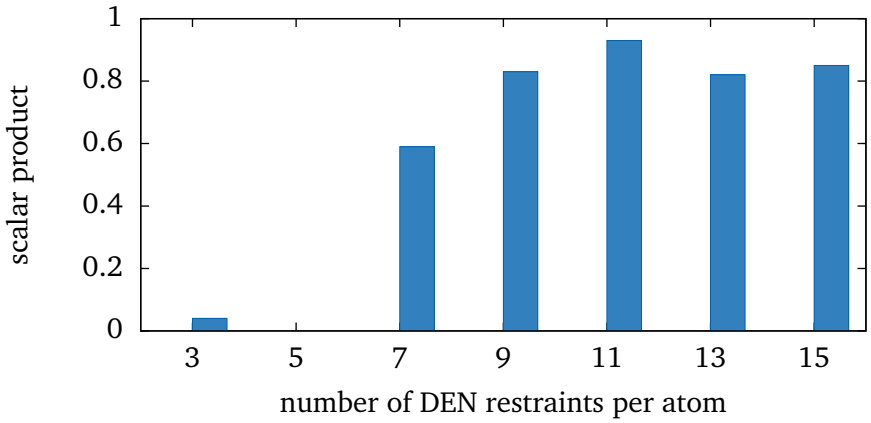


Figure 5.13: Determining the number of next neighbor DEN restraints needed. Scalar Product of the first eigenvector from different refined structure ensembles with the “real” eigenvector. The ensembles distinguish by the number of DEN restraints used during the refinement.

Now the real bead models with sampling on every second and third grid point are flexibly fitted into the bootstrap maps. The bootstrap maps which contain 50 images per map were used since this led to the best outcome by determining the principal motions on the atomic structure. The target model density inside DireX is calculated by weighting the density map by the occupancy values. This leads to a more realistic density map.

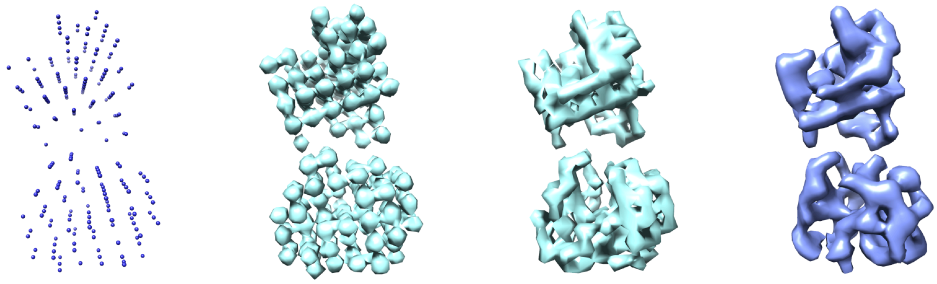


Figure 5.14: Effect of the occupancy value weighted density map calculation. The first image shows the bead model (sampling=3). The second image shows the density map created on the bead model with 5Å resolution and in the third image with 7Å resolution. The last image shows the density map created by using the occupancy values.

In Figure 5.14 the bead model is shown and the difference between a density map calculated with and without taken the occupancy value into account. The map looks better and more smooth.

5.7.3 Principal Motions of Bead Model

The PCA was calculated on the 100 refined bead models in order to determine the principal motions of the system. The outcome of the PCA are 99 eigenvectors and corresponding eigenvalues. The eigenvalues which describe the variance of the two bead models were normalized and plotted in Figure 5.15. Eigenvalue distribution looks quite promising since in the one case the first eigenvector describes 50% of the variance and in the second case 60%.

The movement seems to be well defined by only a small number of eigenvectors.

5.7.4 Sorting Results

Results for Coarse Bead Model Sampling

It was not possible to sort the images along the first eigenvector of the principal motion. A sorting along more than one eigenvector is very time consuming since there are a lot of combinations possible. Therefore

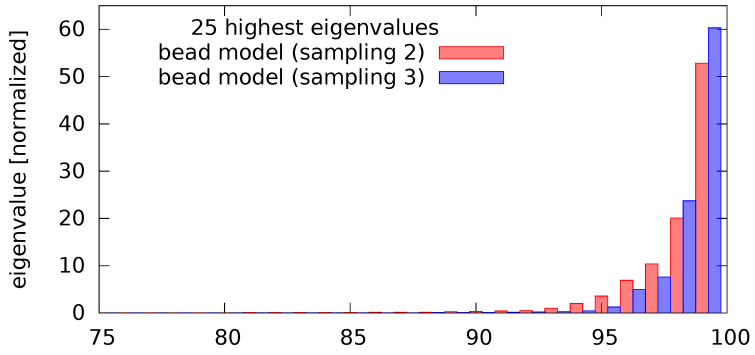


Figure 5.15: Distribution of the 25 highest eigenvectors on the two bead model refined structure ensembles.

the factorial coordinates of the refined bead models were calculated. In Figure 5.16 the coordinates along the first two eigenvectors are shown. The structures can be divided into two clusters. Nine structures lie further outside and have not been taken into account when determining the cluster centers.

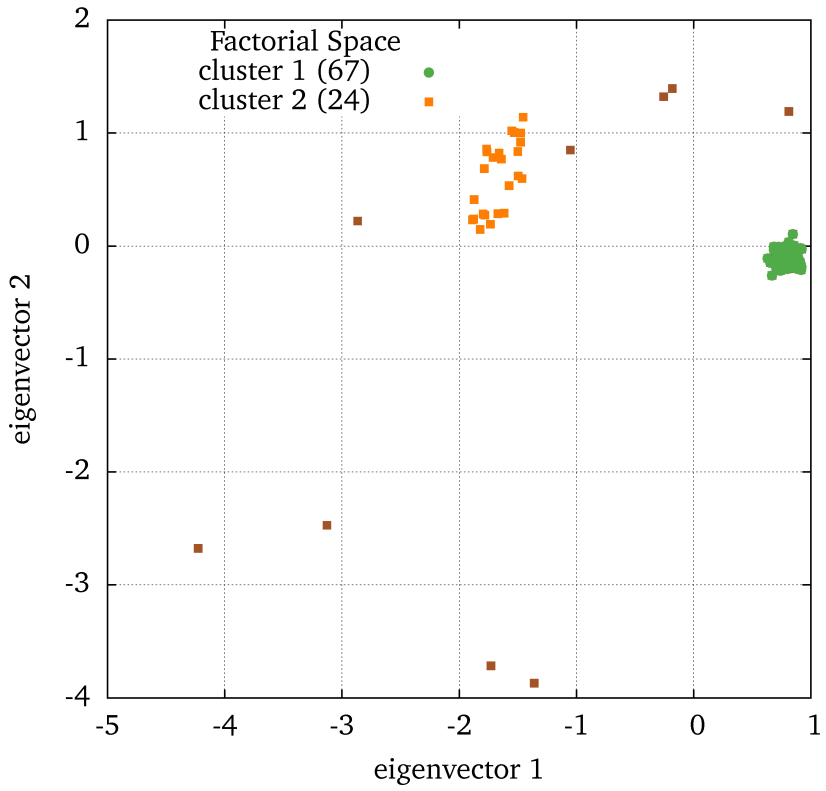


Figure 5.16: Factorial Coordinates (projection along eigenvectors) for all images with two cluster centers at (0.8, -0.1, 0.04) and (-1.66, 0.62, 0.28).

The first cluster center which is very well defined contains 67 structures

and the second contains 24 structures. The cluster centers were calculated to (0.8, -0.1, 0.04) and (-1.66, 0.62, 0.28). This information gives us an idea of the proportion between the reference structures along the first three eigenvectors. The absolute values can not be calculated from this. One could lay one straight through the two points and bend the structure along this straight line. This was not done. Instead, the average structure was shifted by a factor of ten in both cluster center directions which gets close to the correct shift. In Figure 5.17 the two density maps calculated out of these shifted bead models shifted along the first three eigenvectors are shown. Next to this the two conformations are shown for density maps reconstructed of perfect sorted images. If you look at the density maps and compare with the two "correct" maps on the right, it looks as if they are very close to the right solution.

The open conformation belongs to the first cluster and the closed conformations to the second not so well defined cluster center.

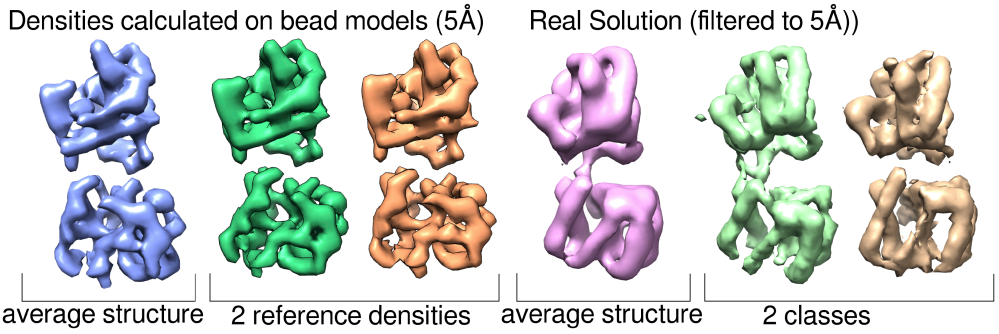


Figure 5.17: Reference structures calculated on the bead model compared with the correct density maps.

The images were sorted to the two bead model reference density maps as before by projection matching. 2/3 of the images were assigned to the first class (open) and only 1/3 of the images to the second class. A fraction of 62% of images were correctly sorted back. Since this was not very satisfying images were then sorted by using a finer bead model.

Results for Fine Bead Model Sampling

For this denser bead model six reference structures along the first eigenvector were calculated. in Figure 5.18 we show the distorted bead models and the resulting density maps. These density maps look much smoother and closer to real density maps than the density maps we calculated before on the other bead model.

An opening and closing motion becomes clearly visible. The images were assigned to these six classes as before by projection matching on low-pass filtered images. The reference density maps were simulated to 5Å.

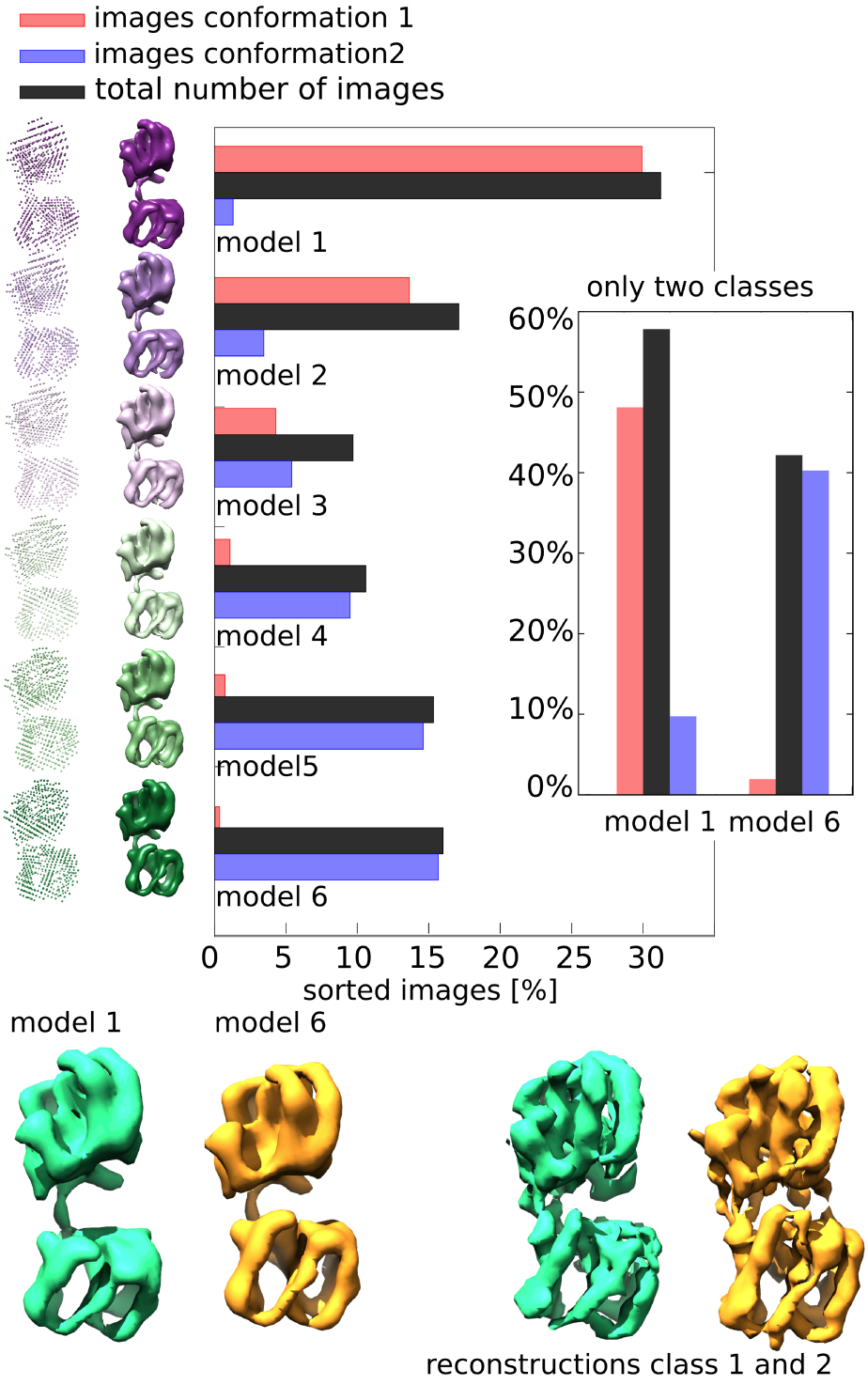


Figure 5.18: This image shows six bead models and their corresponding density maps along the first eigenvector. At the same time, the distribution of the images after projection matching is shown and therefore the amount of images depending on which starting conformation they belong to.

We also show how many pictures we have assigned to which density. If we break these images up in images of the open and closed conformation, we can see that this has worked well. In the next step the images were assigned only to model 1 or model 6. The assignment was correct for 88,4% of the images. On the bottom of Figure 5.18 we show the two model densities again a little bit bigger and next to them the two reconstructed density maps after the bead model sorting.

In several bead model tests more images were always assigned to the open conformation. Also the initial model and therefore reconstructed map of all images is closer to the open conformation than the closed one. This arguments and the results in Figure 5.18 suggest that the bead model has to be shifted further in the closing motion direction.

5.8 Discussion

We were able to create a test system with realistic cryo-EM images. When simulating the test system, an attempt was made to create a system which is as simple as possible, in which only two different conformations were used. In this case, images from five different micrographs were simulated. However, a real data set contains images from several thousand micrographs. But for testing some methods the simple system is still sufficient.

Images were sorted by the density PCA à la Penczek and by Principal Motion Sorting. The simulated images can of course also be sorted using other methods. A 3D classification in Relion [64] was tested and provides excellent results for this data with an assignment of 97% accuracy. Then we sorted the simulated images with the "codimensional PCA" from Penczek. It almost worked perfectly and also 97% of the images could be sorted correctly. Subsequently, we sorted the images according to the principal motions with atomic structure and bead model which resulted in a result of 91% and 88.4% correct assignment. Unfortunately, it became only later clear that these results could not be directly compared. In the projection matching by sorting with principal motions, the orientation parameters used in Relion were taken. In the calculation of the factorial coordinates by multiplication of images with the projections of eigenvolumes in SparX the correct orientation parameters were used since we forgot to write the newly determined angles in the header of the image stack file. The bootstrap maps were calculated in both cases with Relion.

We assume that our method is, in many cases, more sensible because a movement of the conformational change is the basis and this is better given by a structural change than by a density change. On the other hand, for example, a conformational difference given by the presence and absence of a ligand can be described very well by a density PCA and not by our method. If the motion is too large it becomes hard to describe it by principal

motions. Since this is only a linear displacement, a very strong shift leads to a distorted structure.

Our method can be applied in certain cases and is not always the best choice. Particularly continuous movements can be described well with principal motions. We could also show that the principal motions with the atomic structure could be correctly determined. With the bead models we unfortunately had no way to quantify how correctly our calculated principal motions were. However, in both cases the opening and closing of the model was clearly recognizable for the first eigenvector and the density maps calculated from the bead models also looked as if this were going in the right direction.

Later we applied the method to other examples and were able to sort small local changes on another test set and ribosome data on which we wrote a paper, see Chapter 7. We created a more realistic test system by using more different conformational states for creating the system. The reference densities were calculated with a resolution which is much lower than we expect in the class, so that we avoid a model bias in high frequency details. In the projection matching step we applied the CTF of the image we want to compare to the projection images and improved with this the correct assignment further. We show how the refinement can be validated. Unfortunately, we have not managed to sort these new data with a bead model.

A problem which has not yet been solved successfully is the question how many reference densities are optimal and in what distance from each other they should be placed. This can not be determined from the data. One should therefore proceed in such a way and move the structure in positive and negative direction so far that the structure is actually already strongly distorted. Then the images are sorted. After a successful projection matching, no images or less images as possible should be assigned to the broken classes. The distribution of the images indicates how far a shift of the atoms is useful. The number of classes can then be tested. This described path was also applied to the ribosome data in Chapter 7.

6 Principal Motions of Molecular Machines from Cryo-EM Data

A manuscript in preparation follows. It contains a method by which principal motions of an atomic structure can be determined by a heterogeneous cryo-EM dataset. This was tested on GroEL/ES data, contributed by Dong-Hua Chen, Junjie Zhang and Wah Chiu. A powerful tool to calculate correlated motions of MD simulations from proteins is the principal component analysis (PCA). In this work the PCA is applied on a structure ensemble which was created by flexible fitting an atomic model to an ensemble of slightly different density maps. The density maps were calculated by a statistical bootstrapping method and therefore, representing slightly different conformational states. The PCA provides us the principal motions of the structure which are hidden in the heterogeneous cryo-EM data. The program for the calculation of the PCA on the atom positions was written by Benjamin Falkner. Other software which also compute the principal components of structures could not be used. These programs require a lot of memory since a matrix is calculated with the dimension of the (number of atoms)^2. In the new sparse PCA, a matrix of the dimension (number of structures)^2 is diagonalized. This is not as computationally expensive as we are dealing with a rather small ensemble of structures but with a large amount of atoms. Additional the accuracy of the principal motions was validated. Therefore, a randomized density ensemble was created and the principal motions were compared to the principal motions of the randomized structures on which the densities were calculated. This part was done by me.

6.1 Abstract

Understanding molecular motions is necessary to reveal the mechanism by which a protein or protein complex works. Single-particle images from cryo-electron microscopy (cryo-EM) provide information on the conformational distribution, and therefore on conformational dynamics of these molecular machines. We present the principal motion analysis (PMA) method to determine conformational motions of macromolecules from single-particle cryo-EM images and in addition demonstrate how correlations of domain motions and allosteric couplings can be revealed.

6.2 Introduction

Single-particle cryo-EM determines the structure of large proteins and macromolecular complexes. In cryo-EM a three-dimensional density map is reconstructed from a series of single-particle images. As a single-molecule technique, cryo-EM yields information on the entire compositional and conformational distribution of macromolecules. A common way of analyzing this conformational (or compositional) distribution is to sort the single particle images into classes of similar conformation (or composition). 3D density maps can then be reconstructed from each class independently. This yields a collection of 3D reconstructions which describes the conformational (or compositional) distribution of a macromolecular complex. Several methods have been proposed to classify images, which are either performed in 2D [91] or in 3D [4, 59, 64, 65]and possibly focused on only the region of interest [26, 57, 84].

Image classification works well, if the conformations are clearly distinct and largely different from each other, but is less feasible if the conformational motions are continuous. But even sorted classes could still show significant conformational variations which cannot easily be further classified.

In case of a continuous distribution a statistical bootstrapping technique has been found to be useful. A bootstrap approach has been proposed by Penczek and coworkers [58, 75, 97] to estimate the density variance that originates from the underlying conformational variance of the particle. In bootstrapping, a number n of images are randomly picked from the stack of all m particle images. These n images are then used to reconstruct a density map. This procedure is then repeated k times which yields an ensemble of k density maps. This ensemble reveals the density variance as fluctuations around a mean structure assuming a continuum of accessible conformations and is typically analyzed by principal component analysis (PCA) to visualize the dominant global motions. Other methods have also been proposed to analyze the density variance [94] and covariance [4, 45, 78] and to determine continuous conformational changes [18]

However, density variance does not fully reflect the conformational motions of the macromolecule. There could be large conformational motions resulting in only small density changes and vice versa. Consider for example a cigar shaped domain. A motion along the long axis produces density variance only at the tips of cigar, whereas a same-sized motion perpendicular to the long axis produces much more density variance all along the entire shape. For understanding macromolecular motions it is therefore necessary to determine conformational variance, that means the variance of the atomic coordinates.

Here we present a method to determine global conformational motions of macromolecules and show how density variance can be translated into conformational variance. The method is based on image bootstrapping,

which yields an ensemble of density maps capturing the sample variance. This density variance is then translated into conformational variance of the macromolecule by refinement of an atomic model against the bootstrapped density maps to yield an ensemble of models, from which the main molecular motions are obtained through principal component analysis. We demonstrate our approach first on a simulated data set and then on experimental data of the chaperonin GroEL/ES.

6.3 Results

6.3.1 Test with Simulated Data

he structure of the ribose-binding protein (RBP, PDB accession code 1URP), which consists of two-domains (see Figure 6.1 and Extended Data Figure 6.4b), was used to generate an ensemble of 100 conformations using the program CONCOORD [34]. In this example, the conformational variance can be directly determined by a principal component analysis of the conformational ensemble and serves as the true answer that is to be compared with our method. The first eigenvector describes a rotation of the two domains (Figure 6.1c) and the second eigenvector describes an opening and closing of the cleft between the domains (Figure 6.1d).

Particle images were simulated from the ensemble of atomic models (see Methods in section 6.4). A bootstrap was performed by randomly choosing 300 particle images for the reconstruction of each of 500 density maps for each half set. The resolutions of the obtained maps are between 8.3 and 9.8Å as determined by FSC plots. The PDB model was then refined to each of the 500 bootstrapped density maps for each half set using the program DireX [72], which yielded an ensemble of 500 refined atomic models for each half set. The average RMSD between these models of (0.51 ± 0.05) Å is rather small and shows that the bootstrapping yields similar density maps. Cross-validation was used to ensure that the models were not overfitted [24]. Details of the refinement are described in Methods (section 6.4). To determine whether the atomic models capture the variance of the ensemble of density maps, we compared each model with all density maps (see Extended Data Figure 6.5).

A PCA of the ensemble of refined models yields eigenvectors, which are sorted by decreasing eigenvalue (i.e. variance along the eigenvector). We developed an efficient inverted PCA method that scales with the number of samples instead of using the spatial dimensions and allows for analyzing large structures as well as large volumes (see Methods in section 6.4). The first few eigenvectors agree very well with the eigenvectors obtained directly from the analysis of the CONCOORD ensemble. The scalar products between the corresponding eigenvectors for half set 1 (half set 2) are: 0.984 (0.986), 0.919 (0.876), and 0.848 (0.824) for the first, second, and

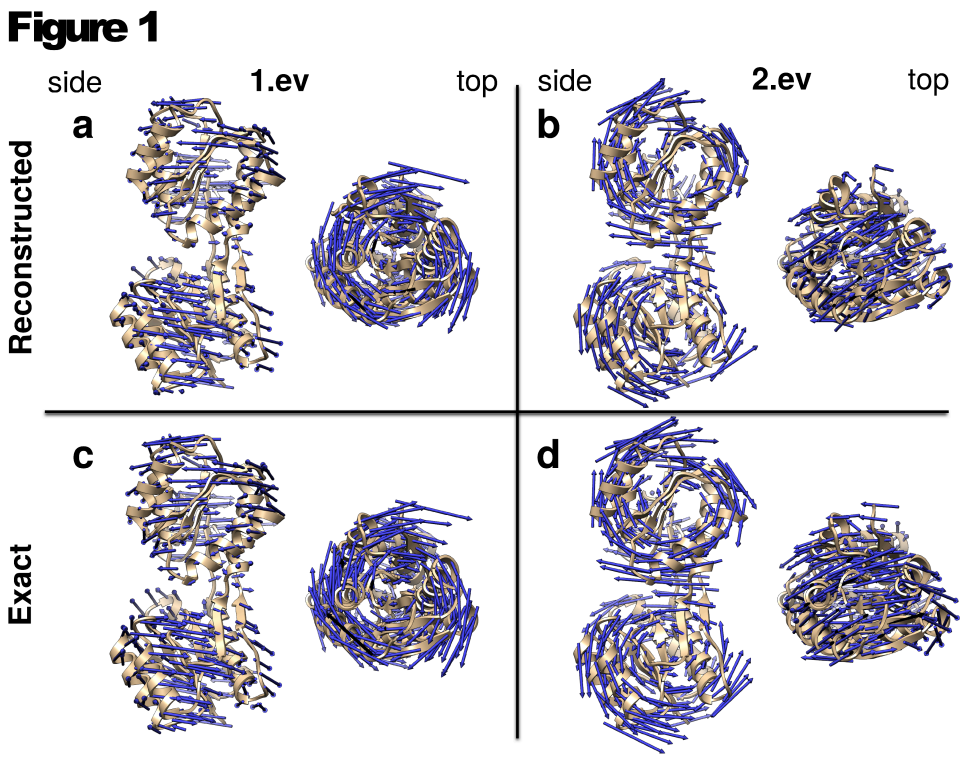


Figure 6.1: Principal motions obtained for the ribosose-binding protein (PDB ID: 1URP) from a bootstrapping analysis of simulated projection images. Projection images were computed from an ensemble of structures generated with the CONCOORD algorithm. An atomic model was refined against bootstrapped density maps which yielded an ensemble of atomic models. Top row: the first (left column) and second (right column) eigenvectors from a principal component analysis of this ensemble are shown (blue arrows). Bottom row: for comparison the first and second eigenvectors obtained for the original CONCOORD ensemble of structures is shown. The high similarity between these eigenvectors indicate that our method can reliably reconstruct conformational motion from the ensemble of bootstrapped density maps.

third eigenvector, respectively. These results confirm that the main conformational variance can be reconstructed accurately by bootstrapping.

6.3.2 Application to GroEL/ES

Our approach was further applied to a data set of the chaperonin GroEL/ES at a resolution of 8Å [13]. This data set was collected with CCD camera and was chosen to demonstrate the robustness of our method. Bootstrapping of the images and calculation of the density maps was done as described above for the simulated data. An X-ray structure of GroEL/ES (PDB ID

1AON3) was refined against all 100 density maps (as described above), which yielded an ensemble of 100 atomic models. The average RMSD between these models of $(0.42/\mathrm{pm}0.05)$ Å is again small and the same correlation matrix analysis as done for the ribose-binding protein test case indicates that the ensemble captures the sample variance (Extended Data Fig. 5).

The first eigenvector obtained from a PCA (Figure 6.2a–c) shows as the main motion a rotation of the apical domains in the trans-ring as well as a screw motion of GroES. The cis-ring apical domains move upwards when GroES is moving downwards (cf. Supplementary Movies 1–3). The second eigenvector (Figure 6.2d-f) shows the same screw motion of GroES as in the first eigenvector but with larger relative amplitude and the rotation of the trans-ring apical domains has opposite sign, which indicates that the motion of GroES and GroEL is mostly uncorrelated.

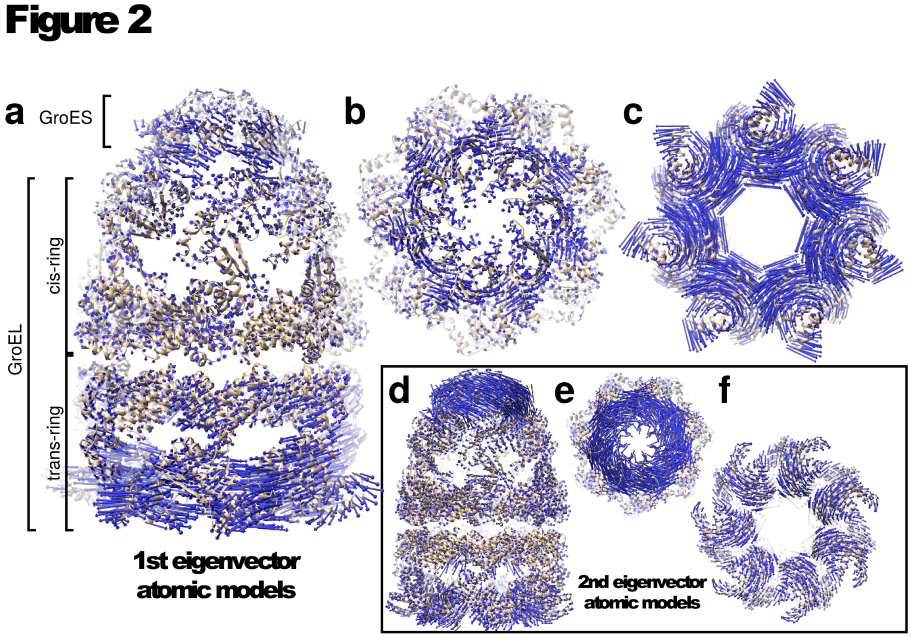


Figure 6.2: Principal motion analysis of GroEL/ES. First (a–c) eigenvectors obtained from a principal component analysis on the ensemble of fitted atomic models. The models have been refined to 100 bootstrapped density maps. The side (a), top (b), and (c) bottom view show that the first eigenvector is dominated by a rotation of the cis-ring apical domains but also contain a contribution from GroES. (d–f) The second eigenvector is mainly a rotation and upward motion of GroES ((a) side, (b) side, and (c) bottom view). The length of the eigenvector is arbitrarily scaled for clarity

Interestingly, the same type of motion but with larger amplitudes is found when studying two different conformations of GroEL/ES in two different

nucleotide states [14, 61]. The difference between these two conformations is shown in Extended Data Figure 6.9a-c. The fact that the same motion is observed means that an onset of the large-scale rearrangement between different nucleotide states is already present in the small conformational fluctuations that we analyzed by bootstrapping.

Since the principal motions determined by bootstrapping are small we further validated them by comparison with results from randomized density maps (see Methods in section 6.4). For additional validation, we compared the internal motion within one GroEL subunit in the cis-ring to the libration axis obtained by TLS refinement against X-ray diffraction data [12]. The rotation axes from this work are very similar to the libration axes from TLS analysis for the GroES and the apical domains (see Figure 6.3), with angles between the corresponding axes of 4° and 42°, respectively.

6.3.3 Correlations Between Domains

ince we determined an ensemble of atomic models which capture the sample variance, it is possible to calculate correlations between molecular motions such as domain motions. This is achieved by calculating a PCA only on the domains of interest and then projecting the corresponding coordinate subspaces of the ensemble onto an eigenvector. The correlation coefficient, r, between the projected motions of the individual domains of GroEL/ES is computed for the first eigenvector and shown in Figure 6.3c-f (see also Extended Data Figure 6.10).

The strongest correlation is observed between the cis-equatorial and trans-apical domains (r=0.44), indicating a significant coupling between the rings (standard error of r is 0.10). The motion of GroES is correlated with the cis-ring but not with the trans-ring. For comparison, the projections of the randomized ensemble (see above) onto the first eigenvector yields no correlations (values in brackets in Figure6.3c-h), except for the motion of GroES, which seems weakly correlated with its direct neighbor domain (cis-apical). These very low correlation coefficients determined from the randomized maps show that the large correlation coefficients from the bootstrap maps (in particular between the distant domains) are caused by correlations between voxels in the bootstrap maps.

The PMA method is a unique tool to study conformational motions of macromolecules as well as correlated motions in atomic detail by single-particle cryo-EM, which enables for example to reveal allosterically coupled motions.

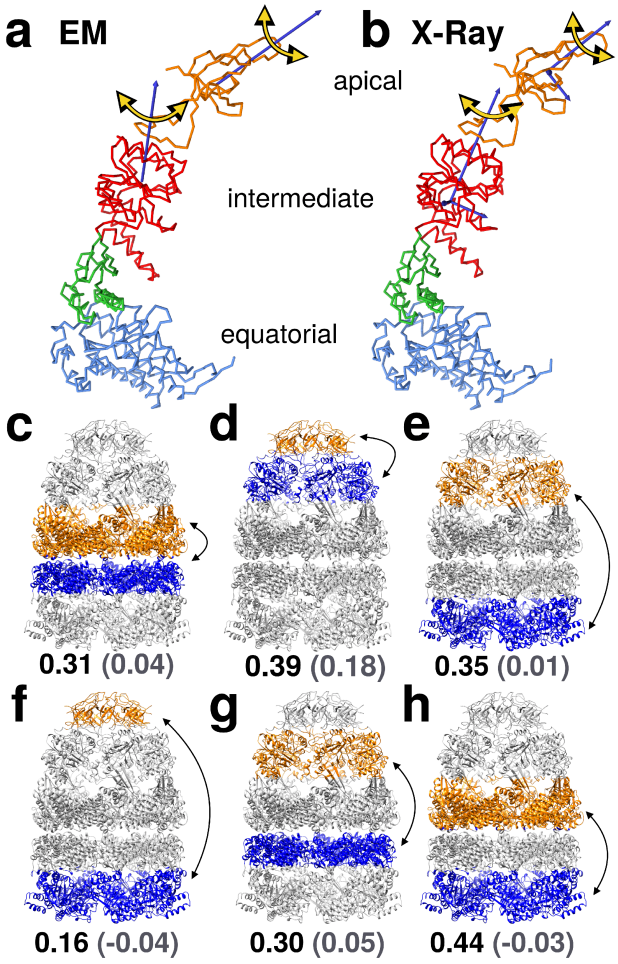


Figure 6.3: Comparison of rotational axes (blue arrows) of GroEL determined from (a) cryo-EM data and (b) TLS refinement against X-ray diffraction data. The ensemble of models fitted against the bootstrapped maps were used to calculate the rotation axis of a single subunit of GroES (orange) and the apical domain (red). TLS refinement of crystallographic data (cf. Chaudhry et al. [12]) with rotations of the respective domains (cis-apical and GroES). The rotational axes are very similar. (c–h) Correlations between domain motions. Correlation coefficients were computed between coordinate subspaces projected onto the first eigenvector for different domains: (c) cis-eq vs trans-eq, (d) GroES vs cis-apical, (e) trans-apical vs cis-apical, and (f) GroES vs trans-apical, (g) cis-apical vs cis-eq, and (h) cis-eq vs trans-apical. Values in brackets are obtained for projections of the randomized ensemble onto the same first eigenvector.

6.4 Methods

6.4.1 Simulating Images for Ribose-Binding Protein (RBP) Test Case.

Here we aimed to simulate a realistic image data set of a protein that shows conformational dynamics. An ensemble of 100 structures was computed using the CONCOORD program with default parameters [34]. The CONCOORD method samples structures around a given structure and was used to generate a realistic conformational ensemble, which at the same time is also well defined and reproducible. The average RMSD between the 100 RBP structures is (3.0 ± 1.1) Å. CONCOORD ensembles have been shown to yield a quite realistic description of the conformational fluctuations and are often very similar to ensembles obtained from few nanosecond molecular dynamics simulations. Density maps were then calculated from the CONCOORD models with DireX, from which projection images were computed with Relion [67], using a pixel size of 1.5Å. For each of the 100 CONCOORD models 20 images were computed for 5 different defocus values in the range of 0.8–$1.6\mu\mathrm{m}$, yielding 10,000 images in total. Noise was added before and after applying the CTF and a B-factor of 130Å^2 to the projection images [6]. The 10,000 images were then split into two independent half sets.

6.4.2 Generation of Bootstrapped Density Maps.

or validation of the conformational eigenvectors in the RBP test case, we needed two independent bootstrap ensembles and therefore the 10,000 images were split into two independent half sets. Each bootstrap map was calculated using Relion with 300 images randomly resampled from these 5,000 images for each of the half sets. The resolution of the bootstrap maps ranges from 8.3 and 9.8Å according to the FSC=0.143 criterion [70]. For each half set 500 bootstrap maps were computed. The bootstrapping of the GroEL/ES particle images was performed with the EMAN [80] program calculateMapVariance.py, which reconstructed 100 density maps as described in [13].

6.4.3 Real-space Structure Refinement.

DireX (available from `www.simtk.org/home/direx`) was used for the refinement of the starting structures. For the RBP case the starting structure (PDB ID 1URP) was refined individually against the 1000 bootstrap density maps (500 from each half set) for 80 steps. The refinements were done on both half sets to estimate the error of the principal motions determined from our approach. The refinement was performed with maps filtered

to 10 Å and the interval 7–10 Å was used for cross-validation [24]. The number of Deformable Elastic Network (DEN) restraints [72] was set to two times the number of atoms. DEN restraints randomly chosen for atom pairs with a distance within 3–15 Å. For the refinements against the 100 bootstrap GroEL/ES density maps, 120 steps were performed and DEN restraints were randomly chosen for atom pairs with a distance within 3–15 Å and no restraints were defined between GroEL/ES subunits. The number of DEN restraints was set to 3 times the number of atoms. For the refinement the maps were filtered to 9 Å and the interval 8-9 Å was chosen for cross-validation.

6.4.4 Principal Component Analysis (PCA).

The standard PCA requires to diagonalize the covariance matrix, which can become very large for high-dimensional spaces. The dimensionality of the coordinate spaces in cryo-EM is typically very large for both density maps (10^6–10^7) as well as atomic models (10^4– 10^6) and the corresponding covariance matrices are therefore too large for regular diagonalization. We developed a fast method to compute the eigenvectors for the case of a high-dimensional coordinate space but when only a small number of samples are used [23], as is the case in this work where only 100–500 bootstrapped density maps (or refined atomic models) are considered.

The idea is outlined briefly in the following.
Consider m data points, $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_m) \in \mathbf{R}^{n\times m}$ (density maps or atomic models) in the n-dimensional coordinate space, which we assume to have zero mean (i.e. the mean value has been subtracted). In the standard PCA, eigenvectors of the covariance matrix,

$$\mathbf{C} = \mathbf{X}\mathbf{X}^{\mathrm{T}} \in \mathbf{R}^{n\times n}, \tag{6.1}$$

need to be calculated such that

$$\mathbf{C}\mathbf{V} = \lambda\mathbf{V}, \tag{6.2}$$

where $\mathbf{V} = (v_1, v_2, ..., v_n)$ is the matrix of eigenvectors and $\lambda = (\lambda_1, ..., \lambda_n)$ is the vector of eigenvalues. To solve this eigenproblem the covariance matrix $\mathbf{C}$, needs to be diagonalized. If $m < n$, the covariance matrix is rank deficient and there are only m-1 non-trivial eigenvalues.

We instead solve the following eigenproblem first

$$\mathbf{C}'\mathbf{V} = \lambda\mathbf{V}, \tag{6.3}$$

where $\mathbf{C}' = \mathbf{X}^{\mathrm{T}}\mathbf{X} \in \mathbf{R}^{m\times m}$. This requires to store and diagonalize only a $m \times m$-Matrix. For the basis transformation to the full n-dimensional

coordinate space we consider first the eigenvalue equation

$$\mathbf{C}'v_i = \lambda_i v_i \tag{6.4}$$

and then multiply data matrix $\mathbf{X}$ from the left side,

$$\mathbf{X}\mathbf{C}'v_i = \lambda_i \mathbf{X} v_i \tag{6.5}$$

and with $\mathbf{C}' = \mathbf{X}^T\mathbf{X}$ it is

$$\mathbf{X}\mathbf{X}^T\mathbf{X}v_i = \lambda_i \mathbf{X} v_i \tag{6.6}$$

and

$$\mathbf{C}(\mathbf{X}v_i) = \lambda_i(\mathbf{X}v_i) \tag{6.7}$$

which shows that the vectors $(\mathbf{X}v_i)$ are solutions to the original eigenvalue problem. The speed of this approach depends now mostly on the number of bootstrapped density maps or refined atomic models and can easily be applied to compute all eigenvectors and eigenvalues for all-atom ribosome structures and large density maps (e.g. 300x300x300). This approach is implemented in the program pmtk (available on Github `https://github.com/bennof/pmtk`).

6.4.5 Validation of Principal Motions with a Randomized Density Ensemble.

Since the global motions determined by bootstrapping are small it is important to validate that these motions in fact originate from correlations between density values at different voxels. For this we created a set of randomized density maps that have the same expected value and the variance as the original bootstrapped density ensemble, but without correlations between voxels. The X-ray structure has been refined against all 100 randomized density maps which yielded another ensemble of 100 atomic models. The PCA on this ensemble yielded the eigenvalues shown in Extended Data Figure 6.7 together with the eigenvalues obtained with the original bootstrapped maps. The first two eigenvalues from the original ensemble are significantly larger than those from the randomized density ensemble, while the eigenvalues for eigenvectors higher than the second are almost indistinguishable from those of the random maps, which indicates that those motions are not significant and cannot be attributed to correlated collective motions, but instead should be interpreted as uncertainty. This uncertainty comprises both conformational fluctuations and experimental error.

6.4.6 Analysis of the Correlations.

Correlations between domains in GroEL/ES were computed by first projecting the coordinate subspaces of the individual domains onto the first

eigenvector, which was computed from the full system (i.e. all domains) and which is shown in Figure 6.2a-c. The projections for the individual domains were then plotted against each other (Extended Data Figure 6.10) and the correlation coefficient computed.

6.4.7 Analysis of TLS Data.

TLS refinement parameters were taken from the PDB file 1URP and were analyzed with the TLSANL program (part of CCP4) [38]. The rotation axes from the libration tensors were written to VRML format and visualized in Chimera. The rotation axes of GroES and the apical domain (residues 196–345) domains were determined from the bootstrapped ensemble with the task file get_ncs_matrices.inp from the program CNS [10] using two models shifted along the first eigenvector.

6.4.8 Fitting ADP and ATP States with GroEL Crystal Structure.

Models of GroEL/ES for the ADP and ATP bound states were obtained by DireX refinement of the GroEL-GroES-ADP crystal structure (PDB ID 1AON [95]) to the density maps EMD-1181 and EMD-1180, respectively [61]. The resolution of the density maps for the ADP and ATP bound states were 8.7 Å and 7.7 Å, respectively.

6.5 Extended Data Figures

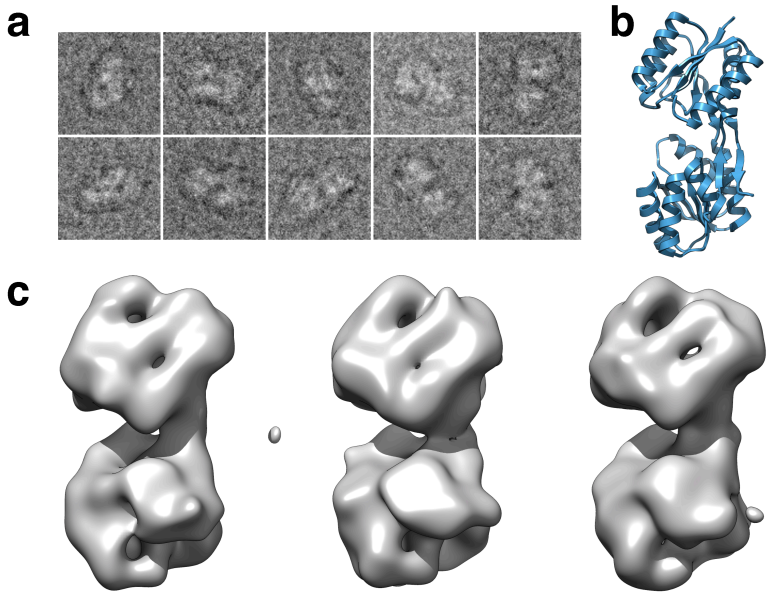


Figure 6.4: (a) Simulated projection images with added noise and CTF used for the ribose-binding protein test case. (b) Cartoon representation of the ribose-binding protein (PDB ID: 1urp) (c) Three examples of bootstrapped density maps, that were used as target maps for the refinement of the atomic model, which yielded the ensembles of models. The map contour level is set to 2σ for all three maps.

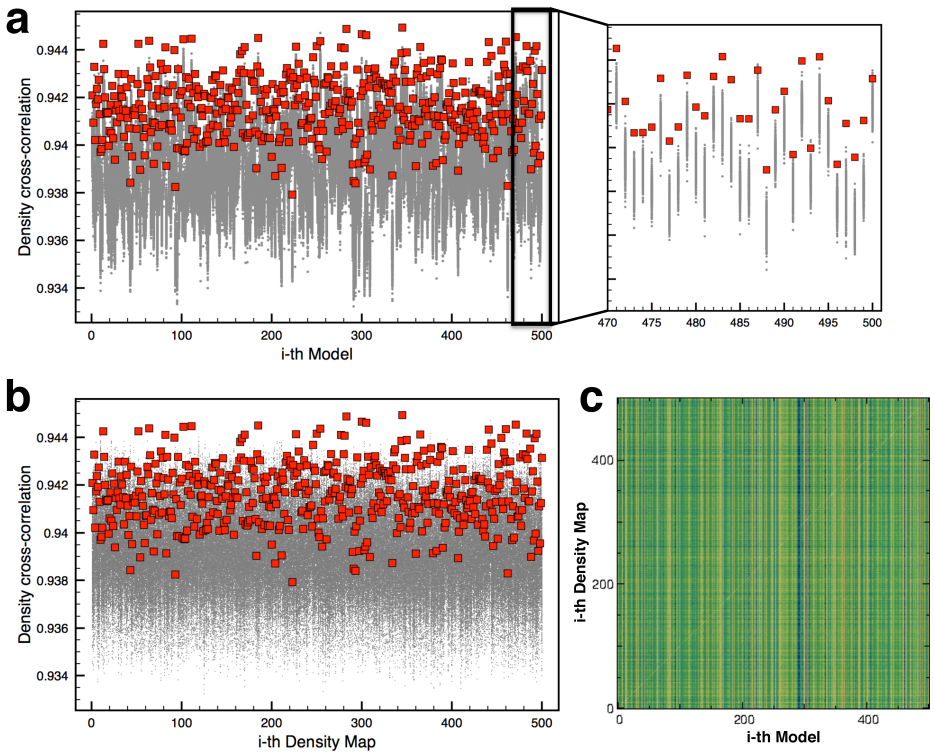


Figure 6.5: Comparison of all fitted models with all bootstrap density maps for the RBP test case with simulated data. Density cross-correlation values were computed between the bootstrap maps and the model maps, which were calculated from the fitted atomic models. Shown are projections of the cross-correlation matrix (a) on the model maps and (b) on the bootstrap density maps. Diagonal elements of the cross-correlation matrix are highlighted (red squares). (a) Cross-correlation values plotted for each model map. In most cases, the model fits best to that bootstrap density it has been fitted to (red squares). We found that 94% of the models fit best to the density map it was refined against, with an average Z-score of the cross-correlation value of 4.4. (b) Plotting cross-correlation values for each bootstrap map shows that the corresponding model fits significantly better than the average over all models. When comparing each density map with all models (see b), we find an average Z-score of 1.6. This means the fitted models indeed capture the underlying distribution of conformations. (c) Cross-correlation matrix comparing all models with all density maps.

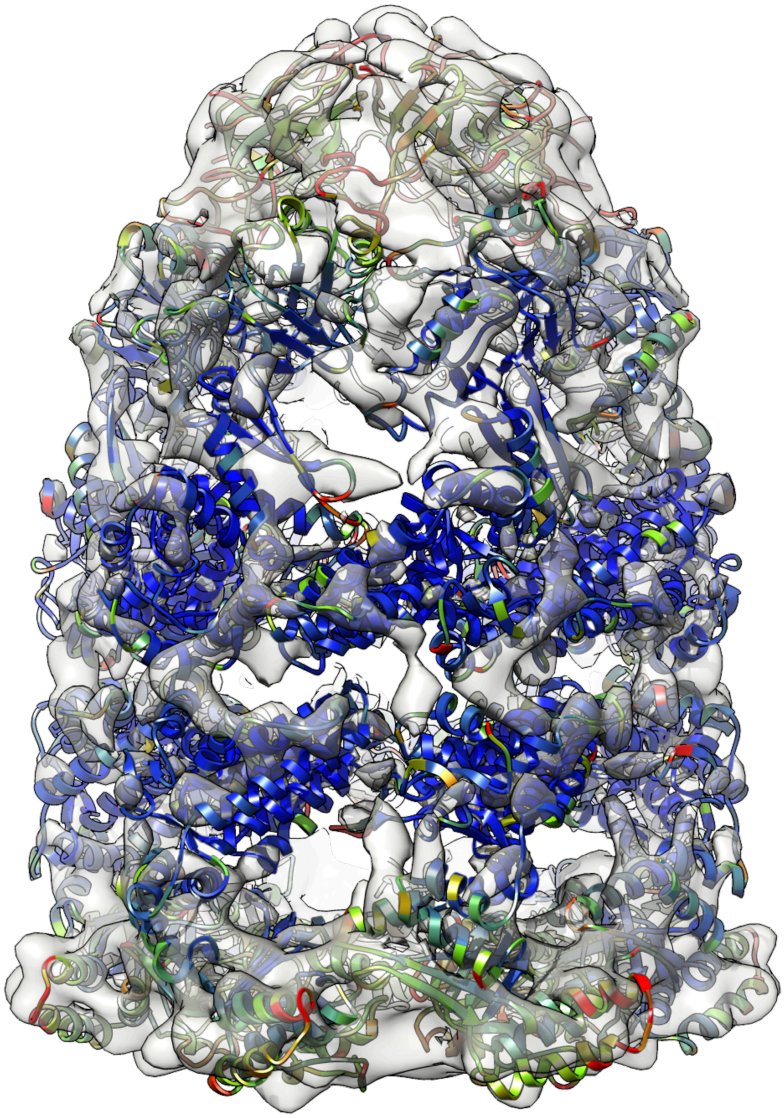


Figure 6.6: Variance map computed from the ensemble of bootstrapped density maps of GroEL/ES. Superimposed is the atomic model color-coded by the atomic fluctuations observed in the ensemble of fitted atomic models. The regions of large fluctuations in the atomic model correlate with regions of large density variance. Lowest fluctuations and lowest density variance are found in the equatorial domain (blue).

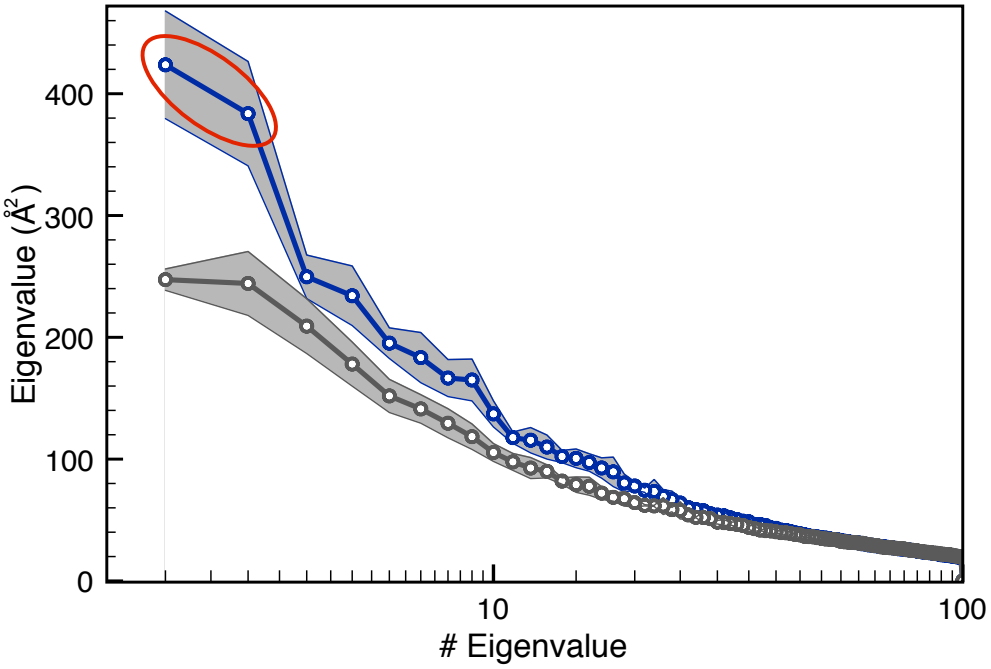


Figure 6.7: Eigenvalue spectrum obtained from the principal component analyses compared for the original bootstrap ensemble (blue curve) and for the ensemble of random density maps (black curve). The random density maps were generated with the only restraint to have the same mean and variance as the original bootstrap ensemble, i.e. the correlations between the voxels in random map ensemble are all zero by definition. The larger eigenvalues for the original bootstrap ensemble is therefore the result of correlations between voxels. Errors on the eigenvalues have been calculated using a jackknife resampling approach.

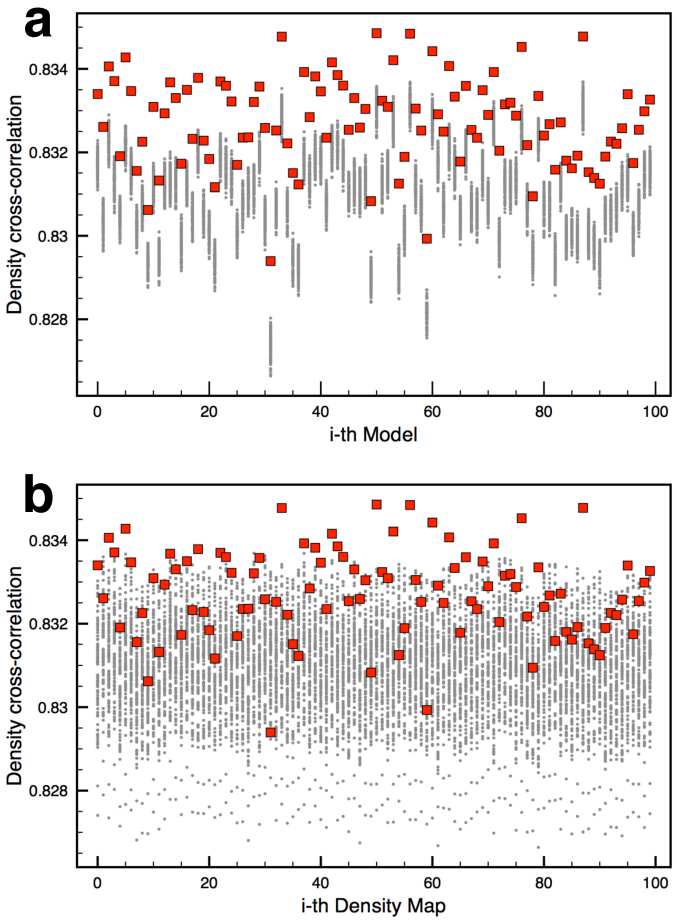


Figure 6.8: Comparison of all fitted models with all bootstrap density maps for the GroEL/ES data. Density cross-correlation values were computed between the bootstrap maps and the model maps, which were calculated from the fitted atomic models. Shown are projections of the cross-correlation matrix (a) on the model maps and (b) on the bootstrap density maps. Diagonal elements of the cross-correlation matrix are highlighted (red squares). (a) Cross-correlation values plotted for each model map. In most cases, the model fits best to that bootstrap density it has been fitted to (red squares). (b) Plotting cross-correlation values for each bootstrap map shows that the corresponding model fits significantly better than the average over all models.

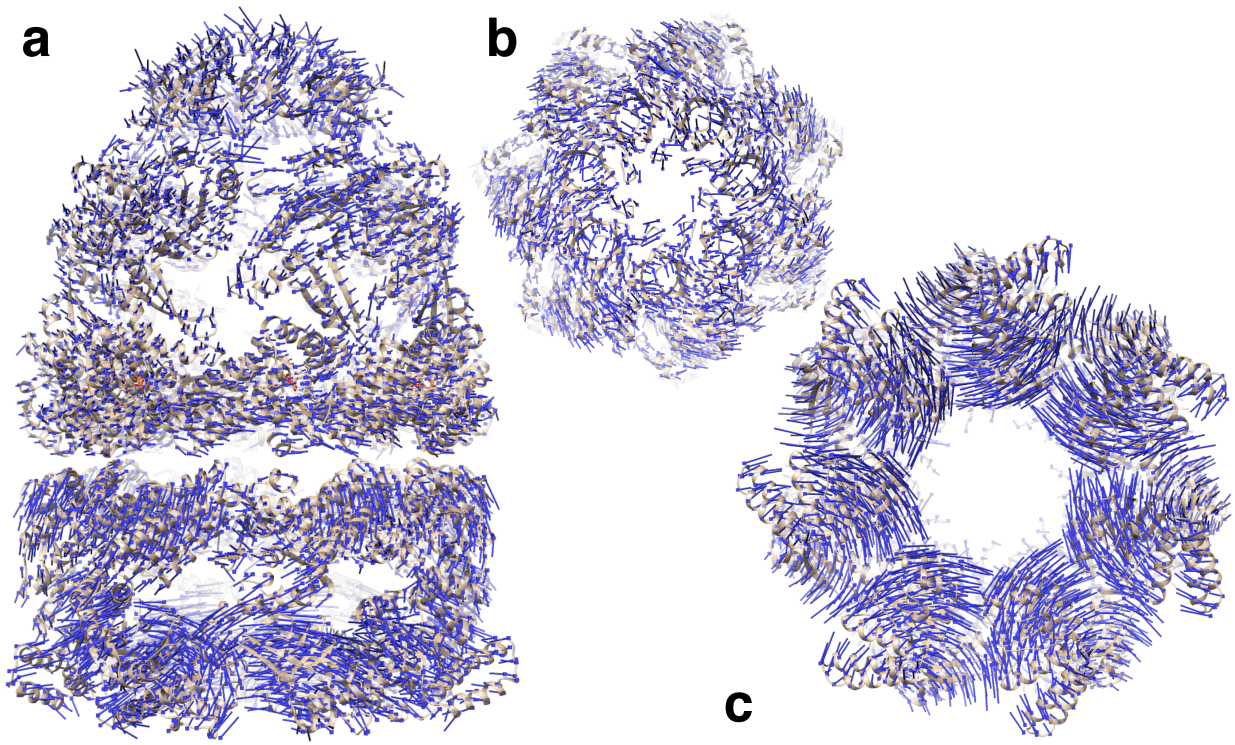


Figure 6.9: The conformational difference between two different nucleotide states of GroEL obtained from models fitted to two different EM maps. (a) Side, (b) top, and (c) bottom view of the structural difference. The difference is very similar to the first eigenvector from the principal motion analysis and shows that an onset of this large-scale motion is already present in the small fluctuations within one conformation.

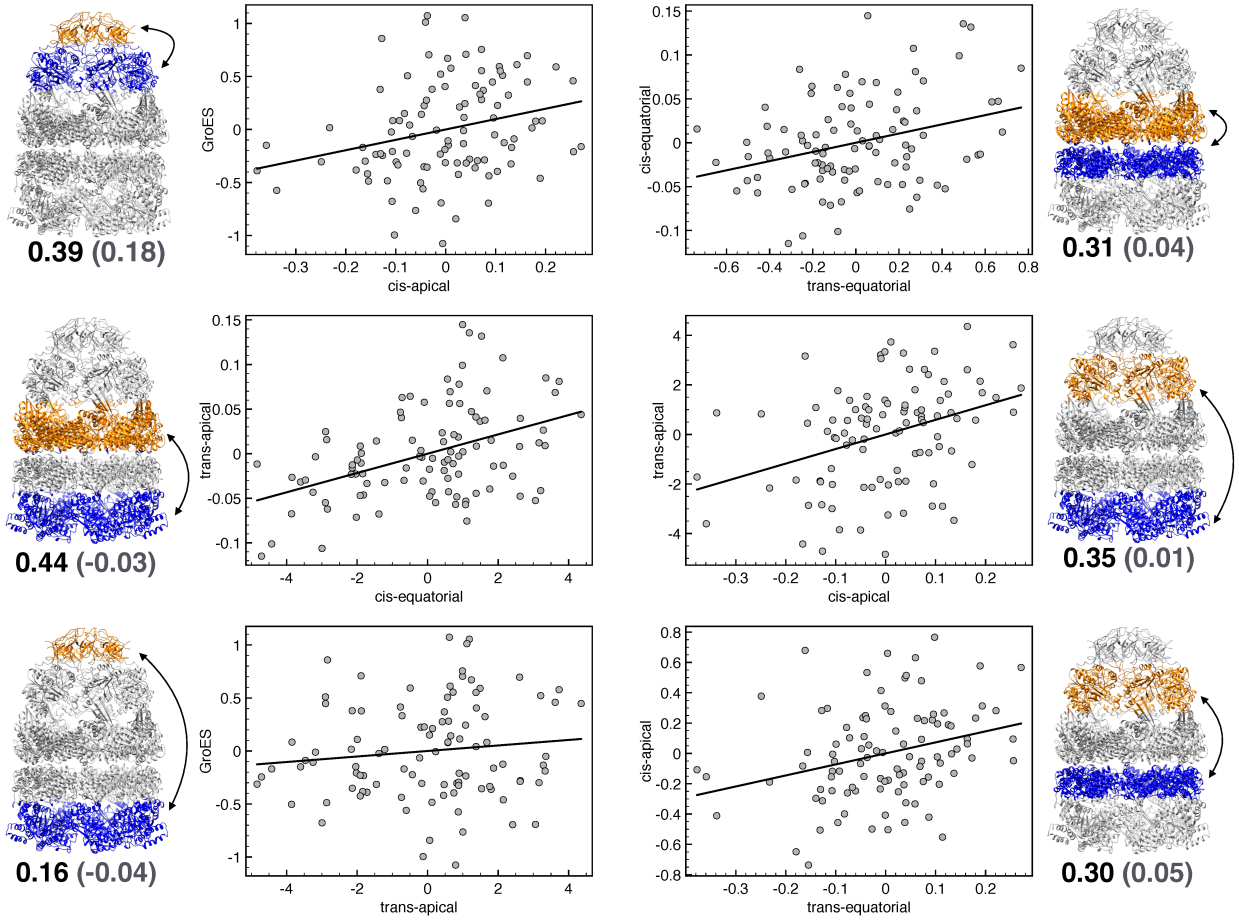


Figure 6.10: Correlations between domain motions. Projections of individual domains onto the first eigenvector are plotted against each other for different combinations: (a) cis-eq / trans-eq, (b) GroES / cis-apical, (c) trans-apical / cis-apical, and (d) GroES / trans-apical. Correlation coefficients are given below each structural image. Numbers in parentheses are the obtained for projections of the randomized ensemble onto the same first eigenvector.

7 Sorting Cryo-EM Images into Classes of Similar Conformations by Principal Motion Analysis

This Chapter contains the manuscript in preparation about the sorting method by principal motion anaylsis. The principal motions are calculated as described in the previous chapter and used for classifying the images. This was done by using a projection matching procedure. The method was applied to a simulated test set and ribosome data which was provided by Niels Fischer and Holger Stark (MPI Göttingen).

7.1 Summary

Typical image classification methods for single-particle cryo-EM images aim to achieve large density difference between and high density similarity within classes. However, to reach high-resolution within the classes it is necessary to sort images into the same class that belongs to particles of the same conformation. Density variance and conformational variance is not necessarily the same. We present a new image classification method that maximizes conformational differences with the goal of better capturing conformational motions and reaching higher resolution within classes. This is achieved by reconstructing the principal motions of the macromolecule from the image data set and constructing reference densities along the directions of the principal motions. We show that our method is useful to detect continuous conformational changes and that we can classify images in order to improve density in mobile regions. The method was tested on simulated data and on experimental data of the ribosome with bound co-factor SelB, which we sorted for small local changes of the L1 stalk and $tRNA^{sec}$.

Keywords Electron Microscopy / cryo-EM, Single-Particle analysis, Principal Component Analysis (PCA), 3D Classification, Conformational Dynamics

7.2 Introduction

A major advantage of structure determination by single-particle cryo-electron microscopy is the possibility to reveal many different conformational states of a macromolecule. Knowledge about these different states yields important information about functional motions and is therefore oftentimes crucial for understanding the function and mechanism of a protein or protein complex. To determine the structures of different conformational states it is necessary to sort the single-particle images into classes of similar conformations. The assignment of an image to a class is a challenging task due to the low signal-to-noise ratio (SNR) of the cryo-EM images and can easily lead to wrong assignments. In addition, the estimates of the angular orientation of the images with respect to each other needs to be determined at the same time and can also change depending on the assignment to the class.

The accuracy of the classification, i.e. the probability that an image is assigned to the correct class, depends on the extent of the conformational difference between the different classes. Conformational states with large differences in the density distribution such as states with and without a bound ligand are easier to sort than density distributions with small differences. It is especially difficult to sort for continuous motions.

A popular method for sorting single-particle images is the maximum likelihood approach implemented in Relion [67]. This approach also provides a solution for 3D classification by performing a multi-reference refinement where the user chooses the number of desired classes [64, 66]. The starting N reference volumes (classes) are created by randomly splitting all the images into N classes and reconstructing the corresponding densities. The differences between the maps will grow by iterations of this multi-reference refinement [66]. Another common sorting approach is supervised classification where references are chosen from multiple known conformations. This type of sorting has been applied for example for the ribosome for which several density maps and atomic models exist [29]. Models for multi-reference refinement were built from predicted structures by Normal Mode Analysis (NMA) [9, 79]. This idea has been extended by describing density maps by generic point-mass models, which allows to treat density maps as coarse-grained elastic density volumes [53] and use them in an improved classification procedure by an elastic 3D-to2D-alignment method [40].

The bootstrapping method is a statistical resampling method for analyzing the variance in a given data set and has been introduced to the cryo-EM field by Penczek [56, 58]. Later, this approach was used in a sorting procedure by applying a principal component analysis (PCA) onto an bootstrap ensemble of density maps [59]. The eigenvectors of a given set of density maps are called “eigenvolumes” since their dimension is that of a density map. The projections along the first few eigenvolumes are calculated for

the EM images and are referred to as factorial coordinates. Classification is finally done by k-means clustering of the factorial coordinates. Recently developed approaches enable the calculation of the covariance matrix directly from the set of 2D projection images [41, 45]. For this purpose, a mathematical relation between variance in 2D and covariance in 3D has been derived [78].

All the methods described above sort images into classes with maximum density variance. However, since we are ultimately interested in conformational motions and aim to maximize the resolution within the classes, the images should instead be sorted for maximum conformational variance between the classes. It is important to note that density variance is not necessarily the same as conformational variance; large conformational motions could lead to only little changes in density and vice versa, depending on the shape of the structure and the type of motion.

Here, we describe a method to sort single-particle images into classes with maximum conformational variance. For this, the density variance obtained by bootstrapping is translated into conformational variance by fitting an atomic model to the ensemble of bootstrap density maps and determining the principal motion from the ensemble of models. For classification, reference structures are built along the largest eigenvectors and the assignment of the images into to the classes (defined by the references) is done by regular projection matching. In contrast to predicting the main motions as done by other methods such as normal mode analysis (NMA) we do not use any prior knowledge about the conformational motions. Instead the principal motions are determined directly from the experimental data. Our approach is therefore an unsupervised classification method.

Our method was first tested on a simulated test case where we show that it is possible to accurately reconstruct the conformational variance of the structural ensemble we put in. Additionally a 3D classification in Relion was done and the results of both methods are compared. Then an application of the method to a cryo-EM dataset of a 70S ribosome is presented. The ribosome data are extensively pre-sorted and describe a state of the unratched ribosome with bound $tRNA_{sec}$ and co-factor SelB. We were able to do a classification to resolve different states along the continuous motion of the L1 stalk and the bound SelB. This is again compared to a local classification performed with Relion.

7.3 Results

7.3.1 Sorting Method

Sorting cryo-EM images by principal motions is recommended for data sets for which small continuous conformational changes are expected. It might also be helpful to apply this method on a sub-state density map after a

hierarchical 3D classification to divide the images into further classes and gain extra knowledge about the dynamics of the system. We consider the case, where large conformational differences have been already removed by standard sorting procedures, such that the remaining conformational variance is small and it can be assumed that orientation parameters (Euler angles and translational shifts) do not need to be further refined during the classification procedure. The conformational variance is approximated by linear movements which are resolved by the use of an atomic model which has been built into the original density map reconstructed from all images. Before starting the set of images is split into two subsets which were obtained from a Relion refinement and apply the method to both half sets independently. The method is implemented as a combination of different programs including DireX [**?**], EMAN2 [80] and Relion [67]. At first, the set of images is split randomly into two subsets and the method is applied to both half sets independently for assessing the reliability of the sorting procedure..

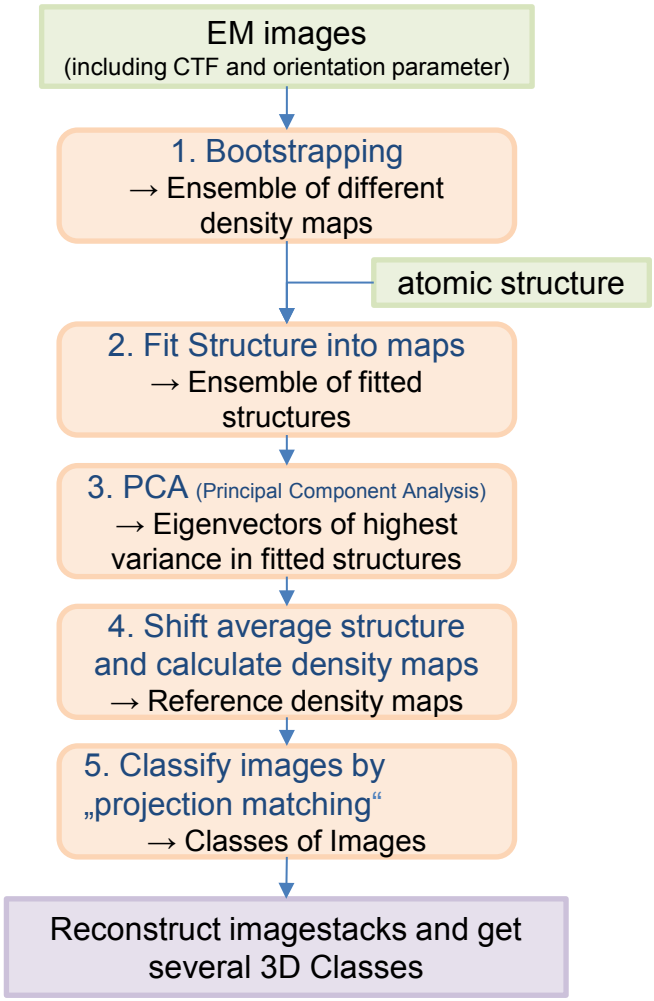


Figure 7.1: Showing a schematic overview of our image classification method.

The sorting by principal motion analysis comprises five steps, which are

shown in Figure 1 and described in the following:

1. **Bootstrapping.** An ensemble of bootstrap density maps is created, which captures the density variance of the sample. Each map is reconstructed by randomly selecting (with replacement) m images out of all N images, that means some images might be used more than once, other might be left out. The choice of of images m is a trade-off such as to be large enough to lead to a satisfying resolution of the density maps, but also not too large to ensure a sufficiently large variance between the bootstrapped density maps. In order to have a good statistical sampling we typically choose the number of density maps to be about 200 (Figure 7.8d).

2. **Model Fitting.** An atomic model built or refined into the density map reconstructed from all images is used for translating the density variance of the bootstrap density maps into a structural, conformational variance. This translation is achieved by fitting the atomic structure into the different bootstrapped density maps using the program DireX, which yields an ensemble of atomic models. Due to the small differences between the bootstrap density maps with integrated FSC values of 8–9, the differences between the fitted atomic models is small, typically about 0.3-0.7 Å. Proper cross-validation is required to avoid overfitting [24].

3. **Principal Component Analysis (PCA).** The ensemble of atomic models approximates the conformational variance of the "real" underlying conformations and is analyzed by a PCA. The main motions of the structure are then represented by the eigenvectors and eigenvalues of the covariance matrix of the atom positions. The eigenvalues indicate the contribution of the corresponding eigenvectors to the total variance. If there are only few very large eigenvalues, then most of the total conformational variance is described by only a few eigenvectors. It is important to note that the variance we calculated by the PCA is much smaller than the real underlying variance and depends on the number m of images used for the reconstruction of the bootstrapped density maps. The absolute size of the variance is however not of interest here, only the directions of eigenvectors are relevant for the sorting. We refer to the first eigenvectors as "principal motions", as they can be interpreted as the principal motions of the structure.

4. **Generation of References.** In the next step, reference density maps are calculated from atomic models that are obtained by shifting the atoms of the average model along the first eigenvector (in positive and negative direction). The number of steps along the eigenvector determines the desired number of classes. The reference density maps are filtered to low resolution (typically below 10 Å) in order to avoid any model bias. Very large shifts result in strongly distorted stereo-chemistry and unrealistic conformations, since the eigenvector is linear. Such extreme classes are only useful for validation, because

they should not be populated in the sorting projection matching step. These reference density maps are filtered to low resolution (typically below 10 Å) in order to avoid any model bias.

5. **Projection Matching.** The particle images are then sorted to the references (classes) according to the maximum cross-correlation value between reference density projection and filtered particle images. This projection matching step depends strongly on the image quality. Ideally, no images or few images should be sorted to the extremely distorted reference classes.

Since the image data were split at the beginning, this procedure yields two half maps for each class. These maps are sharpened and masked by a postprocessing step in Relion.

7.3.2 Test with Simulated Data

Our method was tested on simulated data where the underlying conformational distribution is known exactly. The D-ribose binding protein was chosen as a test system, which contains 127 residues in two domains (see Figure 7.2A) and which is known to undergo large conformational changes. The structure has been solved in three different conformational states by X-ray crystallography [7]. An ensemble of 100 conformations was generated with the program CONCOORD [34], using the PDB structure (PDB ID: 1urp) as starting model. The CONCOORD program uses a geometry-based conformational sampling algorithm to mimic realistic conformational fluctuations around a given protein structure. The root-mean-square deviations (RMSDs) of the calculated models to the original PDB structure vary between 1.27 Å and 5.41 Å (see Figure 7.3a). For all 100 models cryo-EM projection images were simulated with the EMAN2 software as described in Methods **??** and are shown in Figure 7.7b). These projection images were used in an automated reconstruction in Relion using as initial model a density map calculated from the atomic model that was low-pass filtered to 15 Å. The reconstructed density map and FSC curves obtained with Relion are shown in Figures 7.8a) and b).

Determining Principal Motions

For each of the two half sets 500 maps were reconstructed by the bootstrapping technique. An atomic model was flexibly fitted to all bootstrapped density maps. Principal motions were calculated from the ensemble of fitted models and compared to the principal motions calculated from the original CONCOORD ensemble (the true answer). Table 7.1 shows the scalar product of the first three CONCOORD eigenvectors with the first eigenvectors from the fitted models for both half sets.

eigenvector	half 1	half 2
1	0.984	0.986
2	0.919	0.876
3	0.848	0.824

Table 7.1: Scalar product between the first eigenvector of the CONCOORD ensemble (start structures) and the first eigenvector of the ensemble of structures which were fitted to the bootstrap density maps, for both subsets.

The high correlation value between the correct first eigenvector and our determined first eigenvector of 0.98 shows that we can in fact reconstruct the correct motion from the variance of the projection images. In Figure 7.2D)it is shown how the scalar product of the first eigenvector motion is increasing with the number of fitted structures. Figure 7.2B) visualizes the first two principal motions of the protein. The first eigenvector describes a motion in which the lower and the upper domain of the protein rotate with respect to each other. The second eigenvector describes an opening and closing movement of the two domains, which is highly similar to the movement between the three resolved X-ray structures (PDB IDs 1ba2, 1urp, 2dri). The first eigenvector contributes 24.3% (half 2: 22.3%) of the total variance (cf. Figure 7.8C) and therefore captures most of the conformational motion.

Image Classification

To quantify the calculations we sorted the CONCOORD structures by projection along the first eigenvector and calculated the RMSD to the average structure (see Figure 7.3A). We can see that structures, whose projections along the eigenvector have a high value, are also farther away from the average structure (which is the start protein structure (PDB ID: 1urp) as measured by RMSD. This indicates that the eigenvector describes the principal motion quite well. Fluctuations in the curve arise from orthogonal motions into directions of other (higher) eigenvectors. The largest RMSD values to the average structure are around 5 Å. We were able to sort the images into several different substates.

To create the references, the atomic model was shifted 9 steps in positive and 9 steps in negative direction along the normalized first eigenvector with a stepsize 0.67 Å, which corresponds to the standard deviation, σ, of fitted structures. The maximum displacement is therefore 6 Å (9σ) and in total 19 classes. After sorting the images to these 19 reference density maps, we obtained a class distribution shown in Figure 7.3B. This Figure shows the image class distribution along the first eigenvector. No images were assigned to the first class and the first three classes and the last three classes are almost empty (below 50 images), which means in total 13

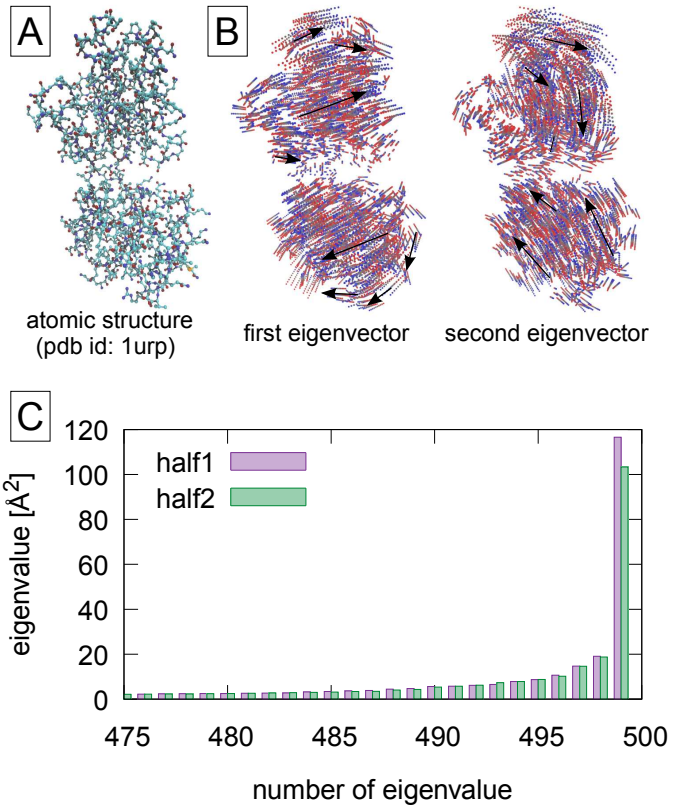


Figure 7.2: (A) The atomic structure of the D-ribose binding protein is shown. (B) The first eigenvector rotates the upper and the lower part against each other. The second eigenvector describes an opening and closing of the protein. c) Showing the distribution of the 25 highest eigenvalues.

different density maps are obtained. Five of these different conformations are shown in Figure 7.8C. Most of the images are assigned to classes that were shifted by less than 4 Å. This finding is in good agreement with the distribution shown in Figure 7.3A. It has to be considered that the RMSD of the CONCOORD structures also contains shifts in other eigenvector directions.

It is also possible to take more than one eigenvector into account for sorting the images. This means the structure is shifted by a linear combination of the first and the second eigenvector. To test whether the sorting can be further improved by sorting along the first two eigenvectors, the average structure was shifted along the first eigenvector in 9 steps RMSDs of 1 Å and along the second eigenvector in 5 steps by RMSDs of 0.8 Å, which yielded 45 reference density maps. The distribution of images dependent on the reference structure is shown in Figure 7.3C for the half 1 subset. The shifts are given in RMSD values.

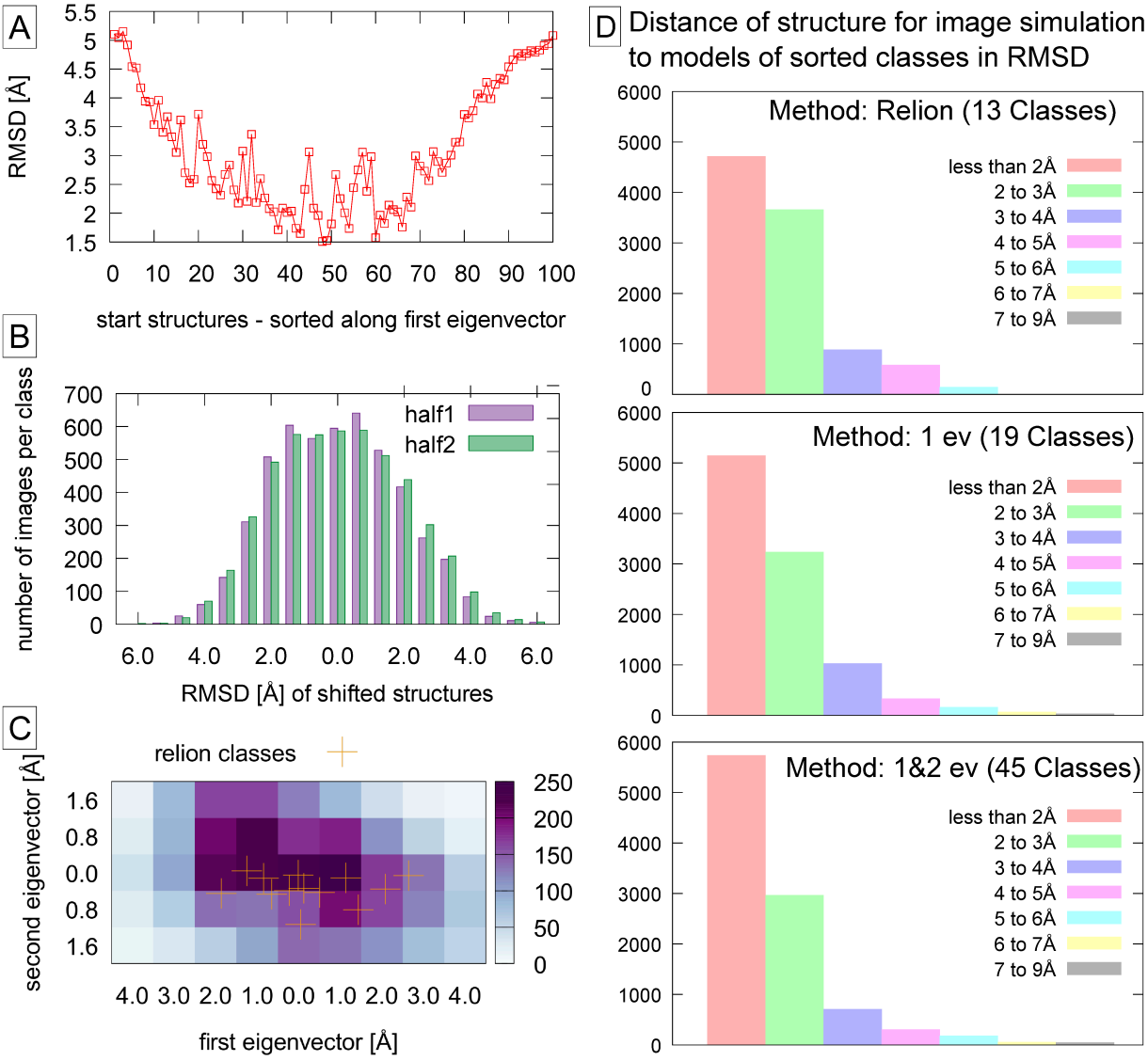


Figure 7.3: A) RMSD values from 100 CONCOORD structures sorted along the first eigenvector to the average structure. B) Images were sorted to 19 reference densities along the first eigenvector. Here we show the number of images assigned to each class and the shift of the eigenvector is transformed in RMSD values. C) We calculated the same distribution by assigning the images to references of structures shifted by a linear combination of first and second eigenvector (only subset "half 1" shown). Panel D) shows the distribution of all 10,000 images in terms of the distance to the correct class for the Relion classification (top), sorted along the first eigenvector (middle) and along the first and second eigenvector (bottom).

Validation of Classification

For comparison a 3D classification in Relion was performed. A different number of classes was tested. If the number of classes was chosen above 15, empty classes remained after the refinement converged. Validation of classifying in Relion and sorting the images by principal motions was done by tracking back from which starting structure the images of the classes originated and how close they are to the reconstructed density map. This was done in the case of sorting by principal motions by calculating the RMSD between the structure from which the image was simulated and the atomic model shifted by the eigenvector which has been used as reference for classification. In the Relion case the RMSD was calculated between the reference structure that belongs to the image and the atomic model that was flexibly fitted into the class density. All the distances were then clustered by distance intervals and summed up. This was done for a Relion classification with 13 classes, the sorting by principal motions along the first eigenvector which resulted also in 13 classes (if we consider only the classes which contain at least 50 images) and the sorting by the first and second principal motion. The results of these calculations are shown in Figure 7.3D). The number of images with a distance smaller than 2 Å is higher in the principal motion sorting than in Relion. It has to be considered that the Relion classes are more centered and much closer to the start structure. This can be seen in Figure 7.3C) where we plotted the factorial coordinates of the Relion structures to our 2D sorting distribution. This means it is not possible to assign images from an original start conformation above 7 Å RMSD distance. So it is difficult to directly compare the plots in Figure 7.4.

7.3.3 Ribosome

The sorting method was further applied to a cryo-EM data set containing 56,881 images of a bacterial Ribosome bound to $tRNA^{sec}$ and the co-factor SelB. The fully processed ribosome data was provided by Niels Fischer [27]. The motion analysis uses a model of the entire ribosome/SelB complex (PDB ID: 5lzd). The average resolution of whole density map is 3.9 Å, however some parts are better resolved than other parts. Especially the interesting $tRNA^{sec}$/SelB region has significantly lower resolution.

Determining Principal Motions

First principal motions of the whole Ribosome were determined as before by applying a PCA on the atom positions of 200 structures fitted to bootstrapped density maps. Most flexible regions which contribute to the first principal motion are color coded in Figure 7.4A). Regions that stay rigid are colored in red and flexible regions in gray and blue. Outer parts are

more flexible than inner parts but the most flexible part is the L1 stalk. By turning the ribosome structure around by 180°, the high flexibility of the L11 protein with corresponding RNA becomes visible. The ribosome is in the unratched state and correlated motions between the subunits cannot be detected anymore. Therefore, we analyzed local regions of interest independently and want to focus on two examples: The first example is the L1 stalk, since the first eigenvector is dominated by its motion and the second example focuses on the region of the $tRNA^{sec}$ and SelB which has a lower local resolution of the density map due to higher mobility of these domains and which therefore has still potential for improvement.

L1-Stalk Dynamics

The L1 Stalk is known to be a very flexible part of the ribosome and which movement plays an important role during the translation cycle. Its movement is correlated to the ratchet motion of the ribosome. The L1 stalk is supposed to pull out the deacylated tRNA out of the tRNA tunnel [8, 81].

In our atomic model the L1 protein and E-site tRNA are not included. We calculated the principal motions only on the atoms of the L1 stalk. The motion of the first eigenvector can be described as a combination of a translational motion towards the E-Site, and additionally a rotation around an axis through the stalk, shown in Figure 7.4B.

The distribution of eigenvalues in Figure 7.4C) shows that we are able to describe the movement of the L1 stalk by a superposition of less than 5 eigenvectors, which means the motion is rather well defined and not random. The first eigenvector was used for classification, which contributes 39% to the total variance of the L1 stalk in half 1 subset and 44% in half 2.

The images were sorted into classes by shifting the L1 stalk atoms along the first eigenvector in 19 steps. Only the L1 stalk atoms were shifted but the reference density maps are computed for the whole ribosome structure to a resolution of 10 Å. For already distorted (too far shifted atoms) L1 structures, we found only few images which correlate best. But most of the images were assigned to reasonable more central reference structures. The distribution of the image assignment is shown in Figure 7.10A. We also show how the distribution changes by using eigenvectors with much lower impact.

In Figure 7.5 we show five highly populated classes which contain about 2/3 of the images (yellow to red) and additionally the density map of all images (purple). The dark gray L1 stalk structure is part of the atomic model which was fitted into the density map reconstructed from all images. We show this model to visualize the differences between the classes. All density maps shown here were multiplied by a soft spherical mask around

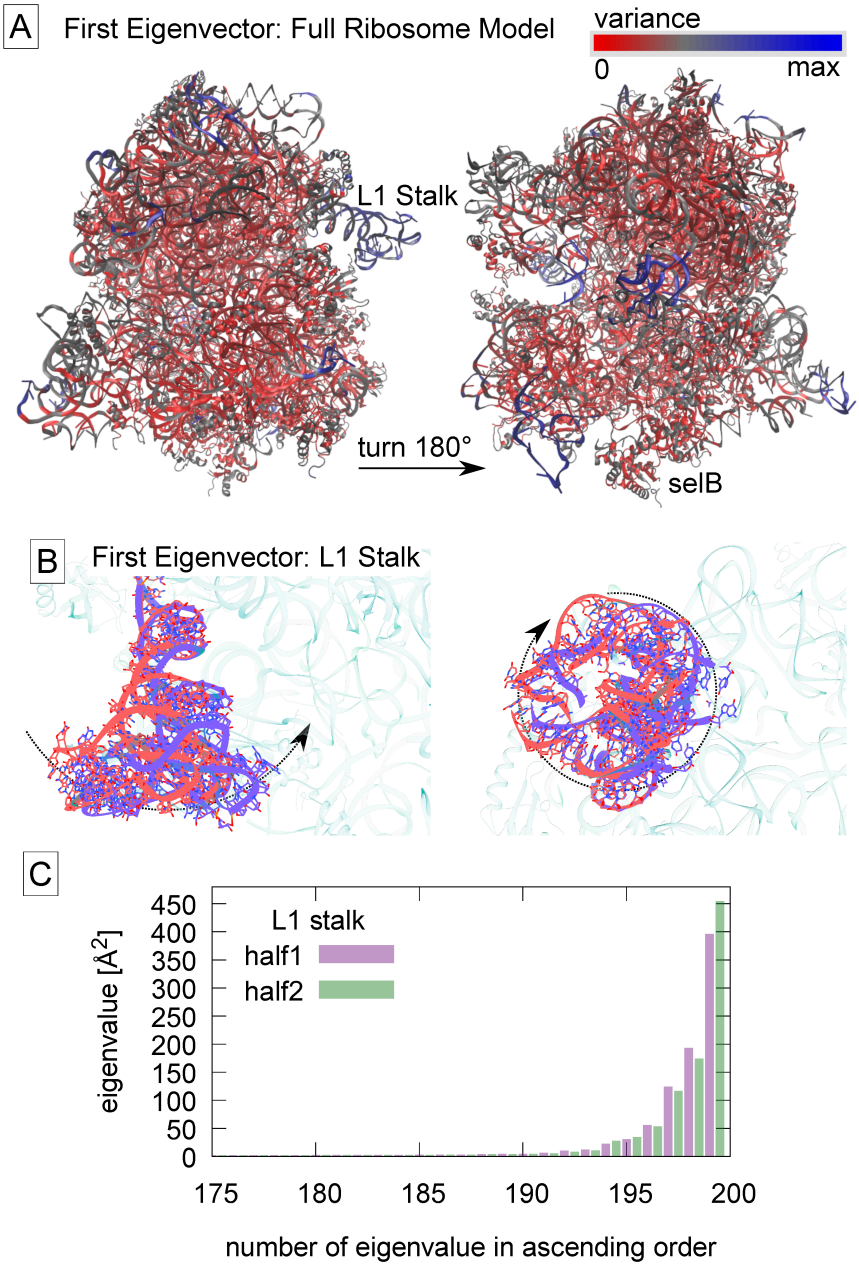


Figure 7.4: (A) Showing the variance of the first eigenvector from our refined structure ensemble. Rigid parts are shown in red, flexible parts transition in color over gray to blue. The L1 stalk shows the highest variance and the L11 protein plus the corresponding RNA the second highest. (B) Movement of the L1 stalk. For visualization of the first eigenvector on L1 stalk we show two structures in which the atoms were shifted in positive and negative vector directions. The movement is a rotation towards the tRNA tunnel (left picture) plus (right picture) a rotation around the L1 stalk axis. The distribution of the 25 highest eigenvalues is shown in C).

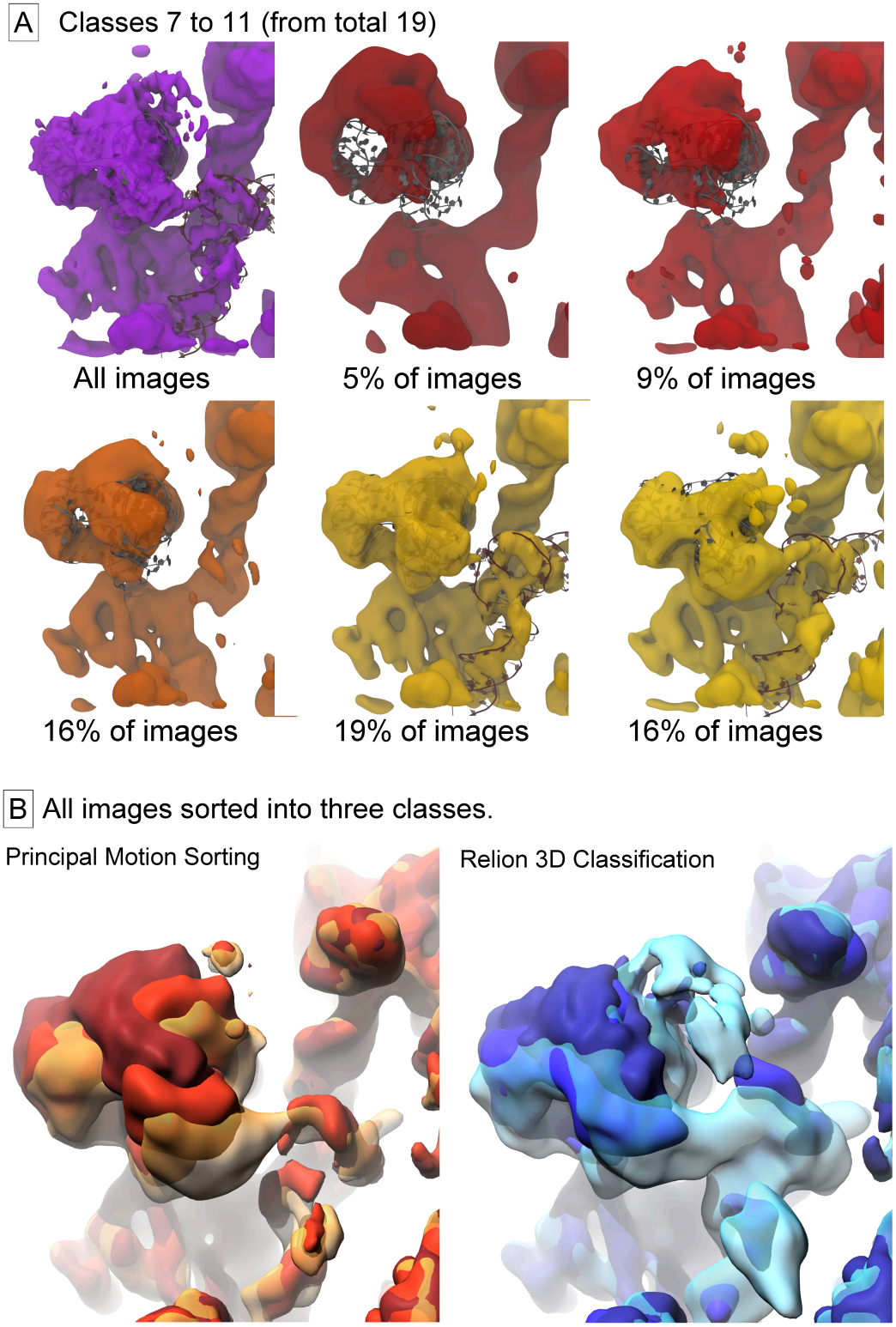


Figure 7.5: (A) Postprocessed densities of L1 stalk. First L1 stalk reconstructed by all images (violet) with atomic model (black) and tRNA (black) and then five classified conformations of L1 stalk. (B) Densities of all images sorted into three classes by principal motion sorting (left) and Relion 3D classification (right).

the L1 stalk, clipped to a smaller box size and then postprocessed in Relion. A postprocessing of a whole unmasked map would lead to an overestimation of L1 stalk resolution. The density threshold of the maps was adjusted such that each density encloses the exact same volume. In the density map of all images we can see that the E-Site tRNA appears, right from the L1 stalk. This is also the case for the maps of class 10 and 11. For these density maps we also show the tRNA structure in dark red. We can see a clear correlation between presence of tRNA density and the movement of the L1 stalk. In the density map of class 7 the tRNA density is also for lower density threshold values nearly fully vanished. The more the L1 stalk moves away from the E-Site the less tRNA density is present in the density maps. It is important to note that we did not use in any information about the presence of E-site tRNA, the reference models were shifted only for the L1 stalk.

We compared our focused classification of L1 with a focused classification in Relion. For this, we sorted all images into three classes along the first eigenvector. We calculated three classes in Relion and merged our 19 classes together to also yield three classes. We show the resulting density maps all at the same density threshold in Figure 7.5B. The mask we used in Relion for local classification has the size of the boxes shown in this figure. The local 3D classification in Relion generates one class with very well pronounced E-Site tRNA and two very similar classes without. The separation between images that contain tRNA and images which do not contain works quite well in Relion 3D Classification. However, the movement of L1 stalk is not resolved, its position is basically the same in all three classes. In our method we got three classes with very different L1 position. The L1 stalk moves towards the tRNA tunnel and away. We can see a correlation with L1 stalk position and appearance of E-Site tRNA. Therefore separation of tRNA density does not work as well as in Relion. Both methods yield different results, and we can learn something from both. Some images contain E-Site tRNA, some do not and the appearance of it is correlated with the L1 stalk movement.

SelB and tRNA Classification

The most interesting part of this ribosome density is the A-Site with bound $tRNA^{sec}$ and co-factor SelB. This part of the density is not well resolved. To analyze this region we applied a PCA on all atoms of the mRNA, $tRNA^{sec}$ and SelB. In Figure 7.6A we show five structures along the first eigenvector. The domain 4 of SelB is the most flexible part. During our refinement we also used DEN restraints between the SECIS motif, where domain 4 binds, and winged helices 3-4 of domain 4. So they were coupled during the refinement and moved together. The total variance of this part is only 1.2 times smaller than for the L1 stalk, but the largest eigenvalue is 8 times smaller. In the distribution of the 25 highest eigenvalues, shown in Figure 7.6B) we can recognize that the different states of SelB/$tRNA^{sec}$

are not well described by only one or a few linear movements. However, sorting along the first eigenvector results in three classes. Their density maps are shown in Figure 7.6C). Since the first eigenvector of the two subsets are more different from each other (correlation=0.88) than for the L1 stalk (correlation=0.95) the images show only the maps for the second subset half 2. Then we fitted the mRNA and the $tRNA^{sec}$/SelB structure into the density maps without DEN restraints between SECIS motif and SelB. The fitted structures are also shown in Figure 7.6C. In class three it seems that the SECIS motif and domain 4 detached and the domain 4 has moved away from tRNA.

7.4 Discussion

We have presented a method for sorting images into similar conformations. We were able to identify several classes from our simulated test data. The projection matching for the simulated data works very well. None of the images were assigned to very wrong reference density maps. We calculated average image/refinement parameters for each class like defocus, MaxValueProbDistribution and number of significant samples out of the Relion refinement star file. As shown in Figure 7.11, a small defocus (=low contrast) and low MaxValueProbDistribution (=low image quality) and high number of significant samples (=low image quality) are correlated with wrong assignments. The projection matching procedure is ill-posed and the assignment error is highly correlated with the signal-to-noise ratio of the images (Sigworth et al., 2010). In a further development the "projection matching" steps could be improved. Validation of our sorted test data leads to a higher accuracy of "correct" image assigning compared to a classification in Relion. The two methods work quite differently which becomes particularly clear when comparing results for the L1 stalk classification of the ribosome data. Relion is able to detect large differences in density like presence or absence of tRNA at the E-Site. Our method does not detect this difference as clearly, however it can reveal the movement of the L1 stalk and its correlation with the presence of the tRNA in the E-site. In our method we need an atomic model first for classifying the images. However, in some cases we do not have an accurate model available. We are therefore currently working on substituting the atomic model by a model which we call a "bead model". By that we mean a system of mass points we place randomly (or on a grid) into the average density map above a certain threshold. The mass points are connected among each other by springs. This model can be fitted into the bootstrap maps. We save our density values of the map in the mass points to recalculate reference density maps out of the distorted "bead models". The advantage is that we do not need to have an atomic model and we do not have a model bias. Sorting cryo-EM images by principal motions is recommended for data sets for which small continuous conformational changes are expected. It might

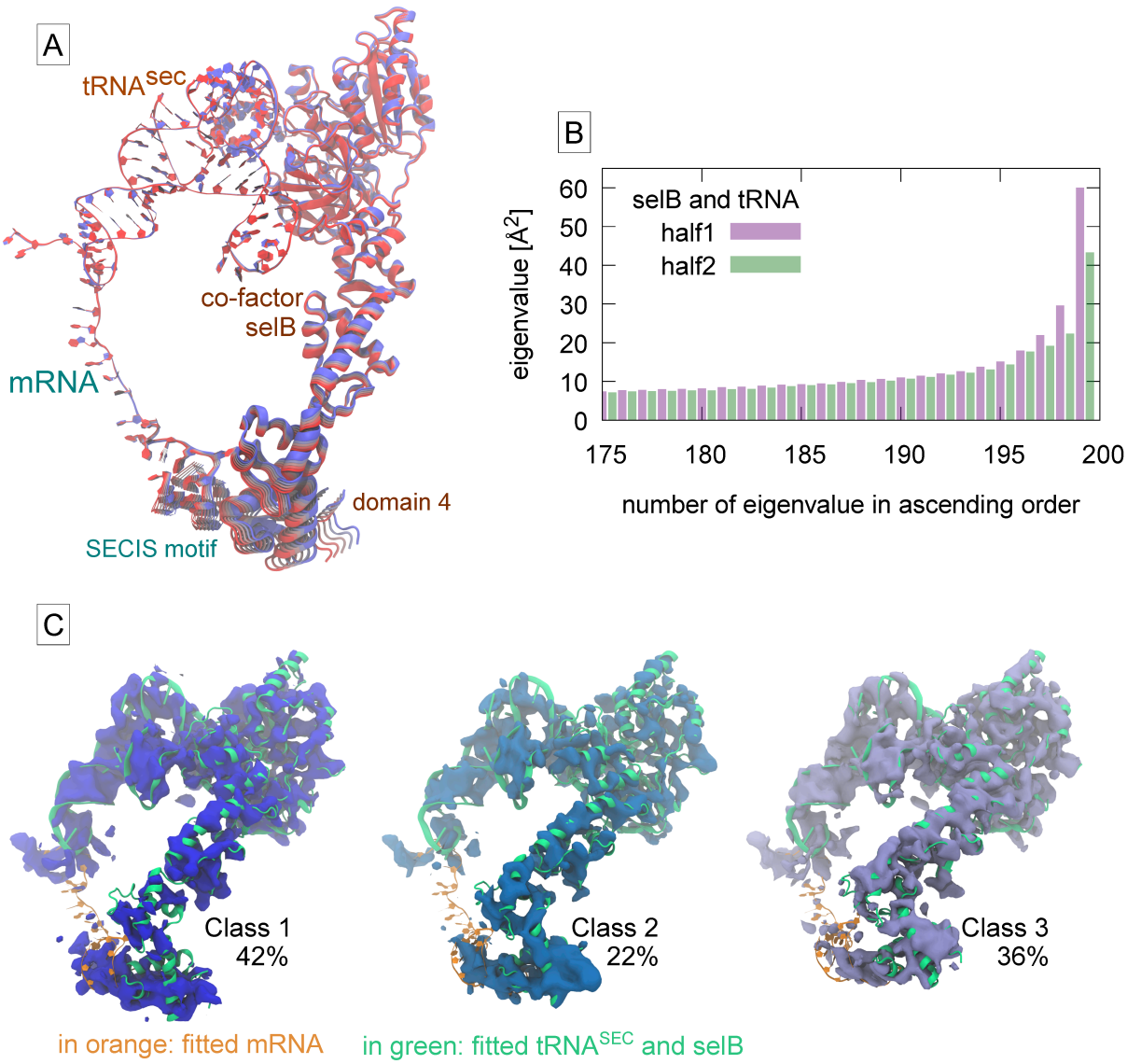


Figure 7.6: Sorting images for SelB movement. In Figure A) we show five structures of $tRNA^{SEC}$/SelB and mRNA which were shifted along the first eigenvector. The tRNA and domains 1 to 3 of co-factor SelB stay rather rigid. The domain 4 of SelB which interacts with the SECIS motif of the mRNA is the most flexible region. In B) we plotted the distribution of the 25 highest eigenvalues and in C) we show three different classes we calculated for this region. We fitted the atomic structure into all three density maps (selB and tRNA in green and mRNA in orange).

also be helpful to apply this method on a sub-state density map after a hierarchical 3D classification to divide the images into further classes and gain extra knowledge about the dynamics of the system.

7.5 Experimental Procedures

7.5.1 Efficient Calculation of PCA

The PCA is usually applied on the output of MD simulations, which means a large number of frames but a small number of atoms. In our case we have to deal with the opposite distribution. We have to deal with large structures (e.g. ribosome 150,000 atoms (without hydrogen atoms)) but only a small number of different structures (e.g. 200). Instead of diagonalizing a covariance matrix with the dimensions of (atom number)^2, the matrix is transformed into a covariance matrix with the dimensions of (structure number)^2. The implementation of the program we use for PCA calculation was done by Benno Falkner during his PhD [23].

7.5.2 Simulation of Test Images

First we calculated an ensemble of different conformations for the D-ribose binding protein with CONCOORD with default parameters. The average RMSD between the structures is 3.7 Å. For each structure we simulated a density map with 3 Å resolution with DireX. Then we calculated 100 random projections for each of the 100 structures that we ended up with 10,000 projections in total. We applied noise, B-factor (130Å^2) and Contrast Transfer Function (CTF) by applying five different defoci between $0.8 - 1.6\mu\text{m}$ to the projections by EMAN2 functions to make the data more realistic. This can be described by the following formula:

$$X_i = e^{-Bs^2/4}\text{CTF} * P_{\varphi,i}(V_k + N_{1i}) + N_{2i} \tag{7.1}$$

Image X is a projection P from volume V of class k, in orientation φ, plus noise N_1 convoluted with CTF function with B-factor plus noise N_2. Noise N_1 simulates the amorphous water or other noise which is affected by CTF. In Figure 7.7 we show five randomly chosen simulated images and the spatial signal-to-noise ratio (SSNR) determined in Relion for the five simulated micrographs.

7.5.3 Refinement of Test Images

For the Relion Auto-Refinement we took all 10,000 simulated images and determined their orientation parameters. We did not change any of the default parameters and as initial model we used a density map that was calculated from the atomic 1urp structure, low-pass filtered to 15 Å. After the Refinement we ended up with one density map with a resolution of 4.3 Å and after post processing in Relion (sharpening and masking) 4.0 Å. We reconstructed both half subsets again, to get a resolution we can

compare with our classes, since we did not do a maximum likelihood for each class again. This resulted in a resolution of 4.5 Å. The FSC curves and the density maps are shown in Figure 7.8A) and B).

7.5.4 Sorting of Test Images

We applied our method to both subsets, in which Relion has divided the images during the Refinement. Each bootstrap map was reconstructed from 100 images and calculated 500 maps. The resolution of these maps varies between 8.3 and 9.8 Å. The atomic structure of the ribose binding protein (pdb id: 1urp) was used as a model which we fitted into the bootstrap maps with DireX. We fitted the structure to a resolution of 10 Å and chose the cross-validation interval from 7 to 10 Å. The reference density maps where calculated with a resolution of 10 Å and the projection matching was done by calculating projections of the reference densities in the correct orientation by Relion. For each image the cross-correlation to all projection images were calculated above 10 Å and assigned the images to the class with the highest correlation value.

7.5.5 Relion Classification of Test images

For the 3D classification we used the final density map from our single 3D refinement with an initial low-pass filter of 6 Å. We chose 13 classes, a τ-factor of 2 and ran the classification for 50 iterations. The image alignment was turned off. We did not change the orientation parameter since we are also not doing this during our sorting method. Then we reconstructed the two subsets for all classes and ran the post-processing procedure. All these steps were done in Relion.

7.5.6 Ribosome Sorting

The Refinement of the ribosome data was done by Niels Fischer. We split the data into the two half subsets defined by Relion. Then we reconstructed 200 bootstrap maps for each subset. We built ensembles by reconstructing the maps with different number of images per map (identical number for each ensemble). Then we chose to work with the maps reconstructed out of 4,000 images, which corresponded to maps with an average resolution of 7 Å. It is useful to find a good balance between resolution of the maps and variance between the maps. If the number of images per map is higher the bootstrap map resolution gets also higher but at the same time the variance between the maps becomes smaller (see Figure 7.9D). The Refinement was done by default parameterization in DireX, only the map_strength parameter was lowered to 0.02 and the cross-validation

range was chosen between 6.7 and 7.9 Å. The Refinement was rechecked by cross-validation implemented in DireX and comparing the model maps of the refined structures with the bootstrapped maps (see Figure 7.9A-C). The PCA was calculated only on atoms of the L1 stalk and in the second example only for atoms corresponding to SelB/tRNA^{sec} and mRNA. The reference density maps were calculated in DireX to a resolution of 10 Å.

7.5.7 Ribosome Sorting – Relion

In Relion we ran a classification with 2, 3, and 4 up to 5 classes. We did this locally by reading in a soft spherical mask, calculated with Relion_mask_create for region of L1 stalk and also for the region of selB. The reference map was the final density map of the normal Relion reconstruction with all images, filtered to 8 Å, skipped alignment and the parameter τ was set to 4.

7.6 Supplement

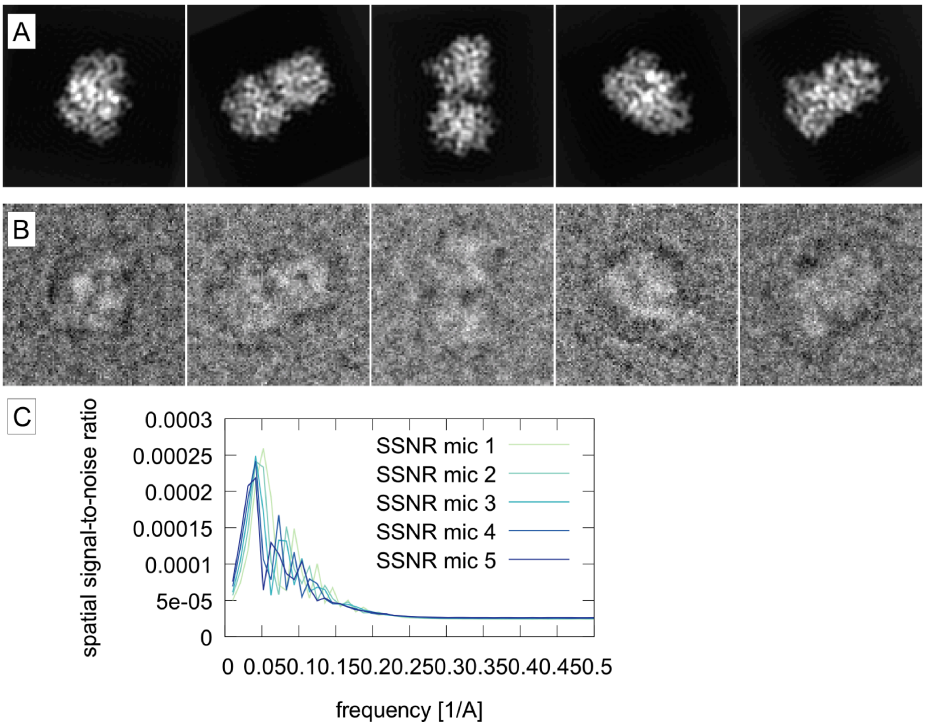


Figure 7.7: A) Shows five random projections and B) the projections plus applied noise, CTF and B-factor. We simulated images from five different micrographs (five different defoci) and in C) we plotted the spatial signal-to-noise ratio for each of them determined in Relion refinement.

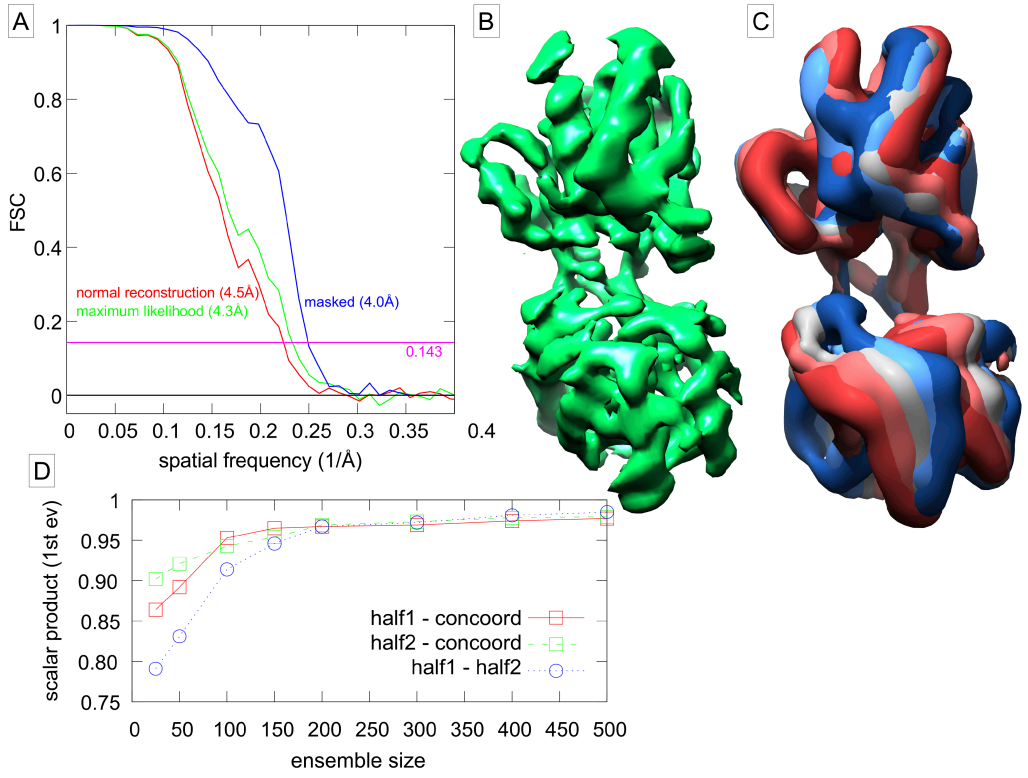


Figure 7.8: In plot A) we show the FSC curves after Auto Refinement of the simulated images. In green we show the FSC for the direct outcome after Refinement, in blue the FSC curve after sharpening and auto masking and in red the FSC curve by "normal reconstruction" out of the two star files. The sharpened reconstruction is shown in B). In Figure C) we show five reconstructions (classes 6, 8, 10, 12, 14) out of 16 classes we calculated in direction of the first eigenvector. In D) we calculated the scalar product of the first eigenvector for different numbers of fitted structures (ensemble size) with the correct Concoord eigenvector, for both half subsets. Additionally we calculated the scalar product between the first eigenvector of half 1 and the first one of half 2 subset.

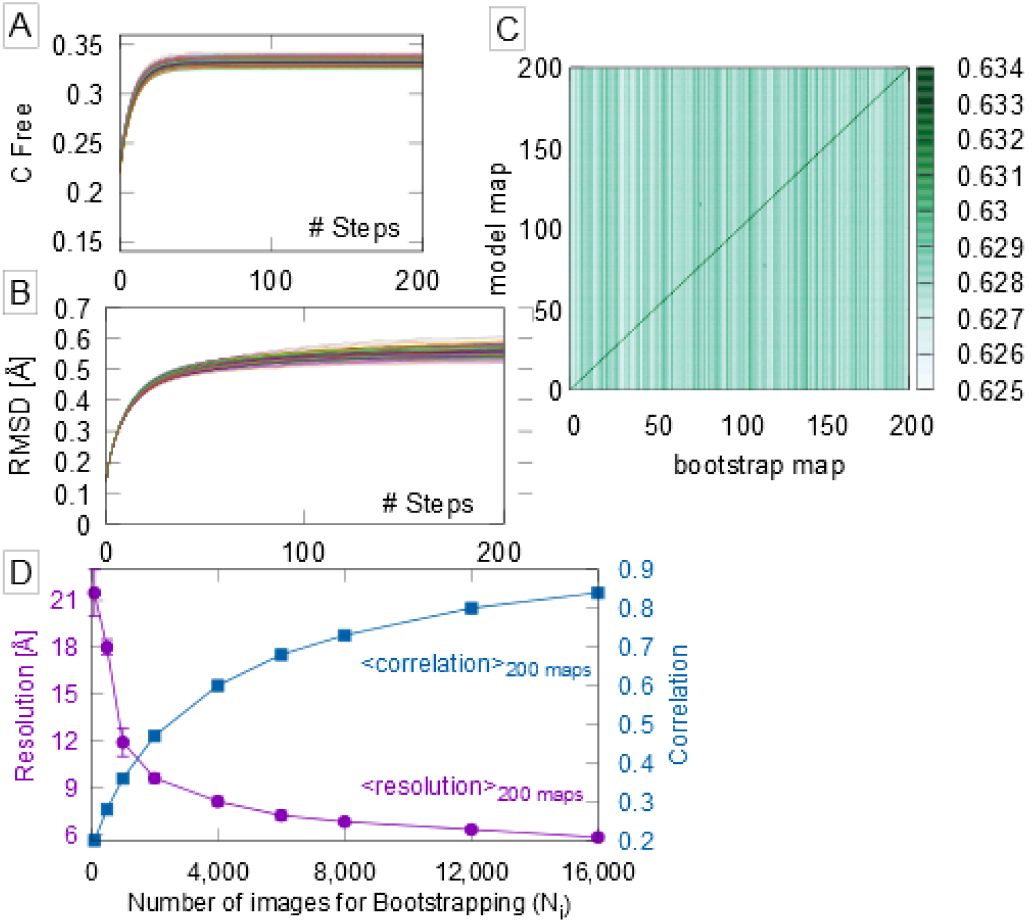


Figure 7.9: This Figure shows possibilities to analyze the Refinement of the atomic model into the bootstrap maps for half 1 subset of the Ribosome data. The Refinement was running in DireX for all density maps by the exactly same refinement parameter for 200 steps. Plot [A] shows the cross-validation value C_{free} which is increasing during the Refinement and converting after 50 steps which indicate there is no overfitting. One other value which can be measured to see if the Refinement has converged is the RMSD of the structure during the Refinement, which is shown in plot [B]. To see if we transformed the density variance successfully into a structural variance we calculate the correlation between all bootstrap maps with all model maps (calculated out of the refined structures). The correlation matrix is shown in Picture [C]. We can see that the correlation between the bootstrap map and the corresponding model map is higher than the correlation of the bootstrap map with all other model maps and vice versa.

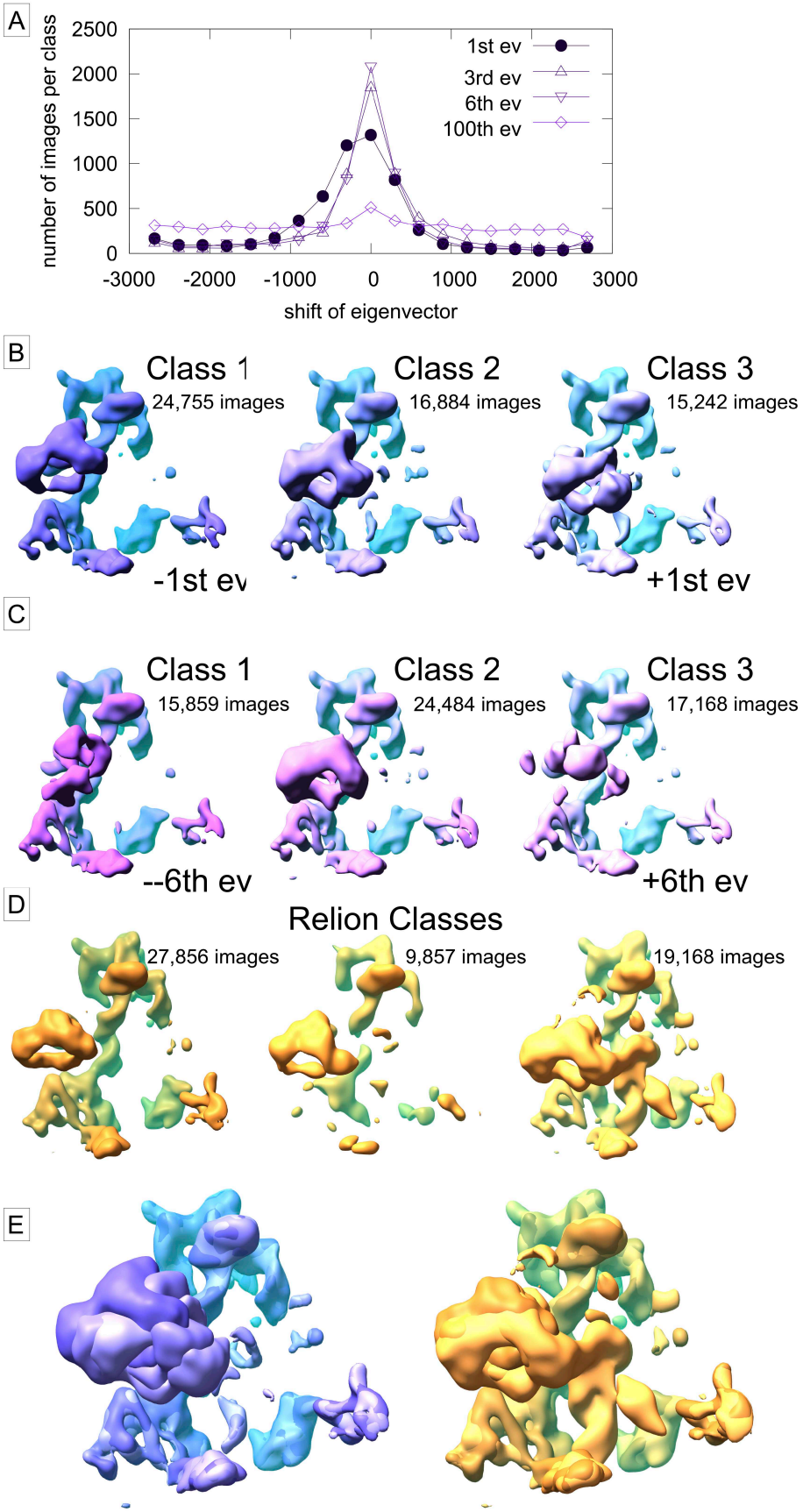


Figure 7.10: Caption next site.

Figure 7.10 Caption:
A subset of every fifth image was sorted to reference models along the first, third, sixth and hundredth eigenvector. The distribution of images per class is shown in the plot A). All reference models were calculated by shifting the atomic model by the different eigenvectors on the same magnitude which is given on the x-axis. Then we assigned the images to the three highest populated classes. The resulting three density maps for the first eigenvector are shown in B) and for the sixth eigenvector in C). For comparison, we show in D) the result of a 3D classification in Relion. The main difference between Relion and our method can be seen in picture E) where we show all three density maps for the principal motion sorting (1st eigenvector) and Relion classification at once to outline the differences.

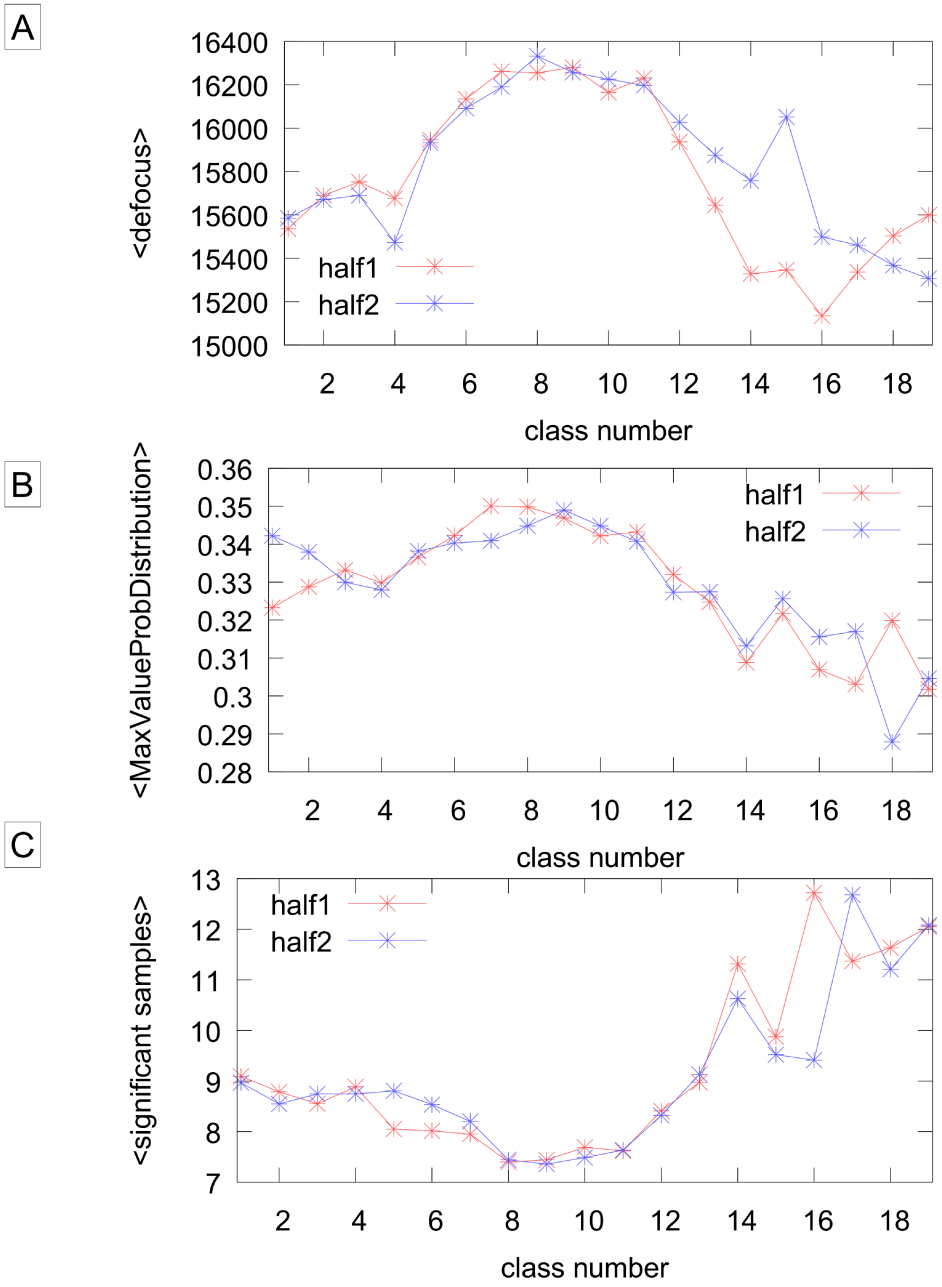


Figure 7.11: After sorting all images into 19 classes along the first eigenvector, we calculated average image properties like average defocus of the images assigned to the class in plot A), average MaxValueProbDistribution in plot B) and number of significant samples in C) from the Relion .star file. It is noticeable that the images taken in middle classes near the average structure are averaged under a higher defocus, have a higher MaxValueProbDistribution and lower number of significant samples. This means, on the average, the contrast of the images is larger, the correlation with the projection is higher, and the direction in which the image is weighted is clearer.

8 Bending of Density Maps into different Conformational States directed by Atomic Eigenvectors

8.1 Abstract

By using the knowledge of a conformational change in form of an eigenvector acting on an atomic model, we are able to bend a density volume describing one conformational state into another density volume describing a different conformational state. We apply our method to density maps in several different conformational states, which are usually the output of a 3D classification procedure of cryo-EM images. Especially classes we obtained by sorting cryo-EM images by principal motions, see Chapter 7 can be bent back to one single density map by applying the determined principal motions which were used for 3D classification to deform the maps. The bent maps can be finally averaged by an FSC weighting scheme to improve the overall resolution.

8.2 Introduction

One big opportunity of cryo-electron microscopy is the possibility to detect the same particle in different conformational states. Therefore, it is necessary to be able to classify the detected images into the different conformations. After an extensive sorting cryo-EM images can be split into several different classes. The smaller the classes are the lower is the resolution information due to the reduced number of images per class. A high number of classes can be found especially for continuous motions of molecular assemblies. Our bending and averaging method can be done in three steps. First, the grid points of the density maps which include the density values are shifted as a superposition of the eigenvector direction from neighboring atoms. In the second step the bent density maps are interpolated back on a cubic grid, which is the new bent density map version. In the last step all maps are averaged by an FSC weighting scheme. We show that we can bent the maps back and improve the resolution of the averaged map for different data sets.

8.3 Results

8.3.1 Simulated Data D-Ribose Binding Protein.

For a first demonstration of our bending method, 500 density maps were calculated from simulated cryo-EM data of the D-ribose binding protein (PDB id: 1urp). The data contains simulated images of slightly different conformational states with a distance of 1.3 to 5.4Å RMSD from the start structure. We analyzed the principal motions of the system as described in Chapter 6. The density maps were calculated by applying the bootstrapping technique. Afterwards an atomic model of the protein was flexibly fitted to all 500 bootstrap maps. A further explanation about the data can be found in Chapter 7.

All the bootstrap maps were bent back to the average atomic model to optimally align all density maps. The idea was to improve the overall density by averaging over all the bent bootstrap maps. For comparison also an average over the "simple" non-bent bootstrapped maps was calculated. In Figure 8.1(a) we show the resolution of the averaged map per number of included maps and the corresponding FSC curves. The FSC curve for averaging "simple" bootstrap maps are converging towards the FSC curve for the normal reconstructed map out of all images. After averaging 500 maps they are almost identical and both maps have a resolution of 4.5Å. The density map averaged over all 500 bent bootstrap maps results in a significant higher resolution of 4Å. In Figure 8.1(c) these two density maps are shown. The averaged map in blue and the average over bent maps in pink. The pink density map looks much more defined and secondary structure elements became more visible. However, we can do it even better.

The principal motions were calculated on the fitted structure ensemble and the simulated data was classified into 13 density maps by the principal motion sorting procedure. This method was introduced in Chapter 5 and applied on this data and ribosome data in Chapter 7.

Here we use the sorted density reconstructions and the calculated principal motions. All these density volumes were bent back to the average density map. In Figure 8.1(b) we show the FSC curves for all thirteen classes and the averaged map calculated in the usual way between the half maps and additional the FSC curves calculated to the model map. The bent maps were averaged by applying different weighting schemes. All three different weightings improved the averaging further. The thirteen maps were normalized by the background noise to average out the noise in the best possible way. The maps were weighted by an FSC weighting scheme. In the third test case normalization and FSC weighting were both applied on the density maps.

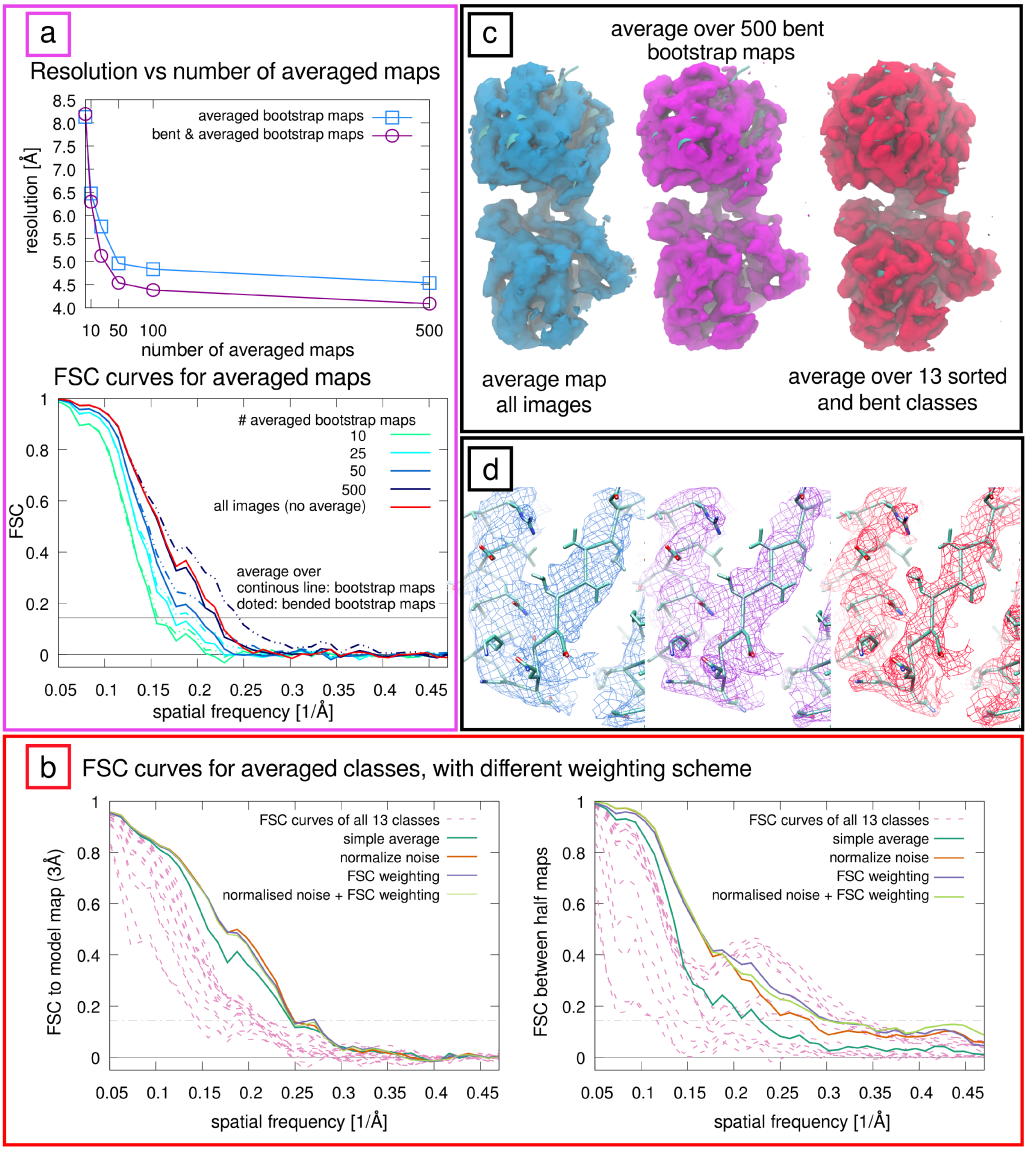


Figure 8.1: **Density Map Bending and Averaging applied on simulated cryo-EM images of D-ribose binding protein.** a) The upper plot shows the increase how of resolution by an average of a given number of bootstrap maps. Blue points show the average of "raw" bootstrap maps and the purple points of bent bootstrap maps. The plot below shows FSC curves for averaged maps containing different number of images, continuous lines show "raw" bootstrap map averages and dotted lines show bent bootstrap map averages. The same line color indicates the same number of density maps per average and red FSC curve is from a density map with all images. b) The left plot shows FSC curves between reconstructed density maps with a model map. The right plot shows normal FSC curves between the half maps. The pink dotted curves correspond to the 13 sorted classes. The average of the bent density volumes is shown in the dark green curve and different weigthings are shown in red, violet, light green (see key). c) Normal reconstruction is shown in blue and result of 500 bent and average bootstrap maps in purple. The average of the bent and noise weighted classes is shown in red. d) We zoomed in into one α-helix of the density maps of c).

As can be seen in Figure 8.1(b) the FSC curves between the half maps improved the most by FSC weighting as expected. The model FSC curves for noise normalization, FSC weighting and the combination of both give nearly the same results.

We started from a density volume with 4.5Å (blue), averaged over bent bootstrap maps resulted in a density volume with 4.0Å (pink) and then averaged over 13 sorted and bent density maps and ended up in a density map with 3.5Å resolution (red) shown in Figure 8.1(c). The improvement is clearly visible in the maps. For better visualization we zoomed into the density maps, and focused on one α-Helix, shown in Figure8.1(d). In this part one can see that the pink density is a bit better pronounced than the blue one. But the red density map looks quite good since it can even represent the side chains correctly.

8.3.2 Ribosome Data

As a second example the bending method was applied to Ribosome data. The data was classified, see Chapter 7 by a principal motion sorting on local areas since there were only small local motions left. Two interesting regions were studied and images were assigned to several classes along the first eigenvector. As an example we took three states of the very flexible L1 stalk. These three classes are shown in Figure 8.2(a) in yellow, orange and dark orange. These images actually show the L1 stalk in three different conformations. The change, which is clearly visible in the density maps, is difficult to recognize from this selected perspective in the figure. However, the actual variance can be seen much better in a Figure of the Supplement from Chapter 7.

Additionally we show in Figure 8.2(a) the L1 density map reconstructed of all images (blue), a model density map (gray) and the final density map which was averaged by the bent density maps (pink). The maps are filtered to their resolution measured by their corresponding FSC curves in Figure 8.2(b). The three starting density maps on the left (yellow to orange) are at a resolution around 8 and 9Å resolution. FSC curves for the density map of all images and the averaged map are pretty much identical. This means both corresponding maps (pink & blue) were filtered to the same resolution around 6.7Å. The blue density map contains more noise and the pink density map looks smoother and more defined. This density map also became more similar to the model density map (gray). We measured this by calculating model FSC curves for all maps. The central class two (orange) has a very high correlation with the model map since this is the class of the images having the highest correlation with the undistorted model. It is higher than the correlation for the density map reconstructed with all images (blue). After the bending and averaging the new density map (pink) has the highest correlation to the model of all

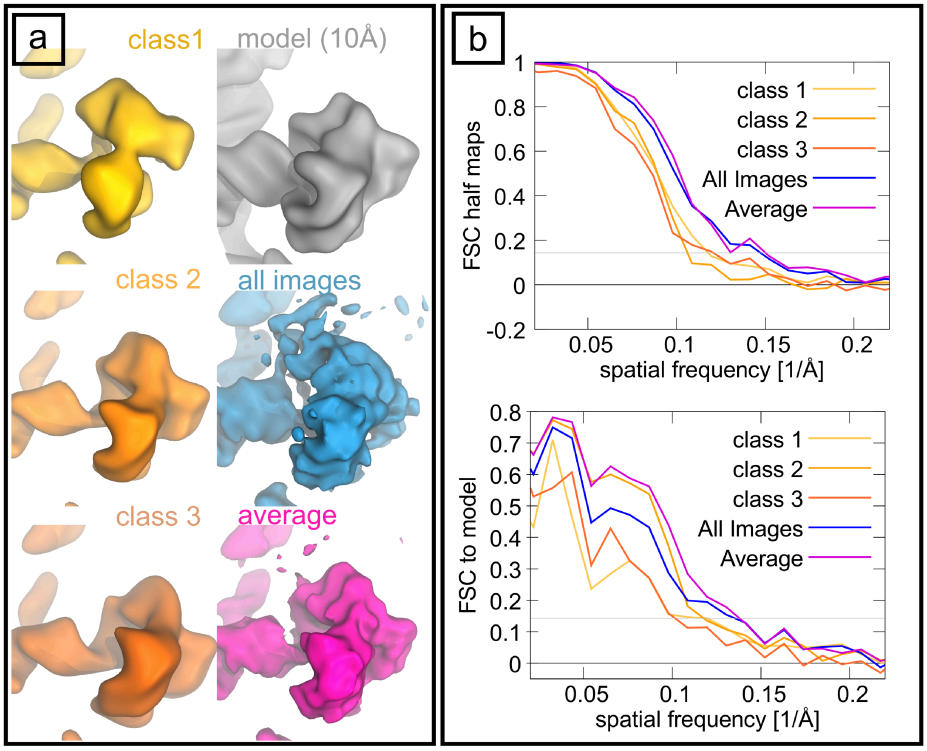


Figure 8.2: **Application of our bending method on three low resolution classes of L1 stalk (Ribosome) density maps.** a) We show the density maps for three different L1 stalk conformations (yellow, orange, dark orange). All images of the three classes together result in the blue density maps. In gray we show a model map calculated of the atomic L1 stalk model to 10Å resolution. Finally after bending density maps of class one and three on the central map of class two and averaging over these maps we ended with the pink density map. All maps are filtered to their corresponding resolution. b) The resolution of the different maps in a) can be read from their corresponding FSC curves shown in the first plot. Additionally we calculated the FSC curves for all the maps to a model map with 4Å resolution.

density maps. This method did not increase the resolution of the initial density map but the information in the map increases.

8.4 Discussion

In the averaging of density maps which are derived from a 3D classification, there is great potential. It is increasingly possible to separate different states of movement. Our method is a logical continuation after a principal motion sorting. The principal motion sorting method was applied to two data sets in Chapter 7. After sorting, we know different conformations of a dataset and at the same time the bending of the atomic structure from the

medium structure to the conformations. These should then also be used to increase the resolution and especially the information in the density map. We have shown that grid points of density maps can connected with neighboring atoms of an atomic structure and the lattice is bent with the displacement of the atoms. In this work we have shown how this can lead to promising results. After this bent lattice has been interpolated again on a cubic lattice one gets the same density map in another conformation. This is what we call bending of a density map. Afterwards bent maps can be averaged.

It turned out that it makes sense to weight the back bent density maps differently. This is particularly important when these density maps differ strongly in the resolution. For this purpose, we have weighted the density maps by applying an FSC filter. This should be the best method in order to achieve the same equal signal-to-noise weighting. It has also worked well to normalize the density maps before by the noise in the background.

If one now does not trust the sorting of images by the highest correlation coefficient with model projections, it is even possible to improve the density map. This can be done by fitting an atomic model to a huge number of bootstrap maps, bend the density maps back to the mean structure and average over these density maps. If an atomic model is not known this can theoretically also be done using a bead model, but we have not tested this yet. In this case it makes sense to store the density values as occupancy values in the bead model such that we end up with reasonable fitted models in the bootstrap maps. More information on the application of bead models can be found in Section 5.7.

8.5 Methods

8.5.1 Calculation of New Coordinates for Grid Points.

A density map is defined on a cubic grid with different density values at each grid point. A grid position $\vec{x}$ of grid point j is shifted by the eigenvector $\vec{a}$ components of surrounding atoms i weighted by factor w_{ji}, which is a function of the distance between grid point j and atom position $\vec{r}_i$.

$$\vec{x}_{j\text{new}} = \vec{x}_j + \sum_{i=1}^{N_i} w_{ij}\vec{a}_i \qquad w_{ji} = e^{-f|\vec{x}_j - \vec{r}_i|^2} \tag{8.1}$$

The factor f of the Gaussian weighting function has to be determined.

The way how we do it can be described by the example of an α-helix which has to be translated. This means each voxel has to be shifted by the same magnitude to preserve the form. A voxel which is located exactly at the atom position has to be shifted by this vector with weight 1.0, a voxel in

the middle of two atoms has to be shifted by a weight of 0.5 in direction of atom 1 and by a weight of 0.5 in direction of atom 2, to be shifted equally in respect to the other voxel. Atom 1 and atom 2 direction are in this example the same. This means the Gaussian function by which the weight is calculated has to use a factor f (see formula 8.1), that at half the distance between two neighboring atoms $0.5\langle d\rangle$ has a value of 0.5, which means:

$$w(0.5\langle d\rangle) = e^{-0.25f\langle d\rangle^2} = 0.5 \rightarrow f = \frac{-4\ln(0.5)}{\langle d\rangle^2} \tag{8.2}$$

If all the atoms would have exactly the same distance to each other this would work optimally and would be a smooth motion.

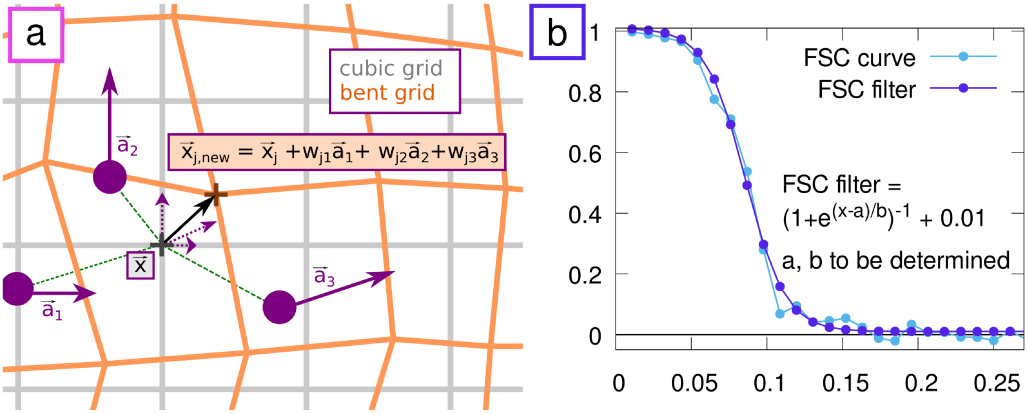


Figure 8.3: **Schematic overview of the map bending and example for the calculation of an FSC filter.** a) The density map grid points are shown in gray and after bending in orange. Three atoms of an atomic model are shown in purple. The first eigenvector acts on each atom and their components are shown by arrays. The shift of one grid point is a superposition of all neighboring atoms weighted by their distance.
b) A normal FSC curve is shown in light blue. The FSC filter function 8.3 is fitted on that curve and is shown in dark blue.

Figure 8.3 shows schematically how a density map bending works.The cubic starting grid is shown in gray. All grid points are shifted now to the orange grid. As an example one grid point is marked by a dark cross. The distances of this point to all atoms is now calculated. All atoms (purple) in a certain distance contribute to the shift of this point. The eigenvector components $\vec{a}_i$ of all atoms are added up by the weighting due to their distance and a new position for the grid point is found.

8.5.2 Interpolation of Density Values From new Coordinates on a Cubic Grid.

After the new grid positions are calculated, the deformed grid, which still contains all the density values, has to be interpolated back to a cubic grid. This interpolation is done by trilinear interpolation, which can

be understood by the assumption that every grid point has a volume of $(\text{grid constant})^3$ with a certain value. This value from a bent grid point is shared by the neighbor cubic grid points by the overlap volume fraction. In the case when there is one cubic voxel without any fraction with bent voxels, the value of the new voxel is set to zero.

Since this interpolation alone did not lead to satisfying results the bent grid is first interpolated on a finer grid with 2^3 times more grid positions and half a grid constant as the map had before. After the new cubic grid is calculated it can be transformed back to the normal cubic size by averaging eight voxels into one bigger voxel.

8.5.3 Averaging Techniques

The density maps that we want to average do not always contain the same number of images and therefore have different resolutions. A wrong averaging of a high and a low resolved map can increase the noise in the average map compared to the already highly resolved density map. Especially in the first test case the information in the 13 classes differed significantly. The first technique is a normalization of each density map by the background noise. This means the whole map is normalized by the average and the standard deviation of the surrounding noise. The second technique is the FSC weighting were a function $\mathrm{FSC}_{\text{approx}}$

$$\mathrm{FSC}_{\text{approx}} = (1 + \exp\frac{x - a}{b})^{-1} \tag{8.3}$$

with the free parameters a and b is fitted to the FSC of the density map and then the density map has been filtered with this function + 0.01. This function fits the FSC quite well as we show in Figure 8.3(b), where we fitted function 8.3 to the FSC curve of one of our density maps. We add a value of 0.01 to the FSC in order to not filter out all the background noise. This is useful as we are still able to calculate FSC functions after averaging.

9 Conclusion

Heterogeneity of macromolecules is a big challenge in the analysis of cryo-EM images. In single-particle analysis high-resolution can only be achieved, if the images show macromolecules in the same or at least very similar conformations. The images therefore have to be sorted into classes of similar conformations. Traditional image classification methods sort for similar density, assuming the same conformations also show the same density. However, this is not strictly the case.

In this thesis, a new method has been developed to sort images according to conformational variance. The method is based on the principal motion analysis (PMA), which analyzes the conformational variance by combining an image bootstrapping procedure with the refinement of models which was developed by Benjamin Falkner, a former PhD student in the group. This ensemble of models provides a picture of the conformational variability. Here the knowledge of this variability is used for sorting the particle images.

In order to be able to test our method correctly, we first considered how to build realistic test systems with images which have the same characteristics as real cryo-EM data. This allowed us to build test systems where we could develop and test our method for sorting the images. Principal motions were calculated for our simulated test system and the accuracy of this method could be assessed. Afterwards this data was sorted into classes along the first and the first plus second eigenvector by projection matching. The quality of the results could again be easily assessed since the data was simulated.

Most of the time I have spent working on the problem of sorting. One goal was also to use a bead model for the sorting method by calculating the principal motions on mass points which are randomly placed into the density map and then using the principal motion of them for sorting. This means no knowledge about an atomic structure is neccessary and the images can be sorted without any danger of a model bias. We could apply these bead model sorting to a test set where we simulated images for only two very distinct classes. Unfortunately, this did not work well when we tried to determine the principal motions for the experimental ribosome data. For this real cryo-EM data set we used an atomic structure and could identify the main regions of flexibility by the principal motion analysis. We decided to do local classification on regions of interest. Especially sorting the L1 stalk worked well. This was due to the fact that with the movement of the L1 stalks towards the tRNA exit tunnel, the tRNA also appeared more

and more in the density maps. This tRNA was not included in our atomic model and therefore also not in our reference densities. The fact that we see this density appear, correlated with the L1 stalk movement is for us a sign that our principal motions made sense and that we assigned the images according to real signal. We also tried to sort images to a smaller eigenvector with less impact on the variance which was not possible and approved that our assigment was done properly.

By assigning the images to classes along the first eigenvector we could see that some images were assigned to classes with a very distorted L1 stalk structure. When we looked at average image quality values like they are calculated inside the 3D refinement in Relion it became clear that images with lower quality tend to be assigned wrongly. With the ever-improving cryo-EM images, which are becoming clearer and more contrast-rich, the assignment of the images when sorting according to principal motions becomes better and less error-prone.

When comparing our method with Relion 3D classification, both methods yielded different results. But our method could also provide insights that one did not see at Relion, such as the correlation of L1 stalk movement and exit tRNA. Therefore, this method can be used to learn more about the conformal heterogeneity of a cryo-EM data set.

There are of course still many issues that could be investigated in the future. This includes, for example, the application of the bead model to determine principal motions and to sort heterogeneous cryo-EM data. One problem was that the cross-validation during the fitting of the bead model to the bootstrap maps was always decreasing indicating that the bead model is easily overfitted. This could be studied in more detail and a solution for this problem has to be found.

To improve the visualization of 3D reconstructions from real particle images, I have developed (together with Amudha Duraisamy) a method for sharpening density maps called VISDEM. In this method we use a bead model that is built into the density and that approximates the distribution of atoms in the structure. The bead model contains different proportions of C, N, O atoms, which corresponds to the average distribution of different atoms in proteins. The structure factor and density histograms from these beads models are used as restraints on the density map, which sharpens the map, filters out background noise, and balances density values of the overall density map further. The improvement could be demonstrated for experimental density maps over a wide resolution range.

As the next logical step after the sorting method has been shown to work, a method was developed to use the information gained from the sorting and subsequent density reconstructions to improve the resolution of our starting density map. This means the different 3D reconstructions representing different conformational states were elastically aligned (or "bent") back to the mean density map and were afterwards averaged. The different density

maps were bent by connecting the grid points of the density map with neighbouring atoms and by shifting along the inverted first eigenvector. This inverts the effect as the reference models were shifted before but in opposite direction. This was tested on simulated data and Ribosome data. Therefore, different averaging weighting techniques were tried; a procedure weighting the density maps by the FSC seems to be the most reasonable results and improved also the FSC of the averaged map significantly. Bending of density maps provided very promising results especially for the simulated test case in which the improvement was clearly visible in the density map but also measurable in from of a resolution improvement from 4.5 Å to 3.5 Å.

Danksagung

Ich möchte mich ganz herzlich bei allen Leuten bedanken die mir während der Doktorarbeit geholfen und mich unterstützt haben.

Als erstes möchte ich mich bei meinem Betreuer Gunnar Schröder bedanken für die wissenschaftliche und methodische Unterstützung. Bei Schwierigkeiten war er jederzeit ansprechbar und hat mir geholfen und bei auftauchenden Problemen immer wieder neue Ideen gehabt. Ich habe die Möglichkeit bekommen an einem topaktuellem Forschungsthema zu arbeiten und an Konferenzen und Workshops teilzunhemen. Ein besonderer Höhepunkt war hierbei die internationale Fachkonferenz (GRC Three Dimensional Electron Microscopy) in den USA.

Dann möchte ich mich bei unserem Institutsleiter Dieter Willbold bedanken der sich als Zweitkorrektor zur Verfügung gestellt hat.

Als nächstes möchte ich mich bei meinen Arbeitskollegen Amudha, Zhe, Philipp, Marianne, Michael und Dusan bedanken, die mich besonders in der schwierigen Anfangszeit unterstützt haben.

Da ich in meine Doktorarbeit viel Arbeit investiert habe wäre das natürlich nicht möglich gewesen ohne die Hilfe, Unterstützung und Ablenkung durch meine Eltern, meine Schwestern Alexandra und Frederike und meine Freunde Nicolas, Volkmar, Tobi, Liya und Tam.

Bibliography

[1] Abrahams, J. P. and Leslie, A. G. W. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallographica Section D*, 52(1):30–42.

[2] Adrian, M., Dubochet, J., Lepault, J., and McDowall, A. W. (1984). Cryo-electron microscopy of viruses. *Nature*, 308(5954):32–36.

[3] Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., and Adams, P. D. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D: Biological Crystallography*, 68(4):352–367.

[4] Andén, J., Katsevich, E., and Singer, A. (2015). *Covariance estimation using conjugate gradient for 3D classification in Cryo-EM*.

[5] Bai, X.-c., McMullan, G., and Scheres, S. H. W. (2015). How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, 40(1):49–57.

[6] Baxter, W. T., Grassucci, R. A., Gao, H., and Frank, J. (2009). Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. *Journal of Structural Biology*, 166(2):126–132.

[7] Björkman, J. and Mowbray, S. L. (1998). Multiple open forms of ribose-binding protein trace the path of its conformational change. *Journal of molecular biology*, 279(3):651–664.

[8] Bock, L. V., Blau, C., Schröder, G. F., Davydov, I. I., Fischer, N., Stark, H., Rodnina, M. V., Vaiana, A. C., and Grubmüller, H. (2013). Energy barriers and driving forces in tRNA translocation through the ribosome. *Nature structural & molecular biology*, 20(12):1390–6.

[9] Brink, J., Ludtke, S. J., Kong, Y., Wakil, S. J., Ma, J., and Chiu, W. (2004). Experimental Verification of Conformational Variation of Human Fatty Acid Synthase as Predicted by Normal Mode Analysis. *Structure*, 12(2):185–191.

[10] Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. *Nature protocols*, 2(11):2728–33.

[11] Campbell, M. G., Veesler, D., Cheng, A., Potter, C. S., and Carragher, B. (2015). 2.8 {{}/AA{}} Resolution Reconstruction of the Thermoplasma Acidophilum 20 S Proteasome Using Cryo-Electron Microscopy. *eLife*, 4:e06380.

[12] Chaudhry, C., Horwich, A. L., Brunger, A. T., and Adams, P. D. (2004). Exploring the structural dynamics of the E. coli chaperonin GroEL using translation-libration-screw crystallographic refinement of intermediate states. *Journal of Molecular Biology*, 342(1):229–245.

[13] Chen, D. H., Luke, K., Zhang, J., Chiu, W., and Wittung-Stafshede, P. (2008). Location and Flexibility of the Unique C-Terminal Tail of Aquifex aeolicus Co-Chaperonin Protein 10 as Derived by Cryo-Electron Microscopy and Biophysical Techniques. *Journal of Molecular Biology*, 381(3):707–717.

[14] Chen, D.-H., Madan, D., Weaver, J., Lin, Z., Schröder, G. F., Chiu, W., and Rye, H. S. (2013). Visualizing GroEL / ES in the Act of Encapsulating a Folding Protein. *Cell*, 153(6):1354–1365.

[15] Ciccarelli, L., Connell, S. R., Enderle, M., Mills, D. J., Vonck, J., and Grininger, M. (2013). Structure and conformational variability of the Mycobacterium tuberculosis fatty acid synthase multienzyme complex. *Structure*, 21(7):1251–1257.

[16] Cowtan, K. (2010). Recent developments in classical density modification. *Acta Crystallographica Section D*, 66(4):470–478.

[17] Cyrklaff, M., Adrian, M., and Dubochet, J. (1990). Evaporation during preparation of unsupported thin vitrified aqueous layers for cryoelectron microscopy. *Journal of Electron Microscopy Technique*, 16(4):351–355.

[18] Dashti, A., Schwander, P., Langlois, R., Fung, R., Li, W., Hosseinizadeh, A., Liao, H. Y., Pallesen, J., Sharma, G., Stupina, V. A., Simon, A. E., Dinman, J. D., Frank, J., and Ourmazd, A. (2014). Trajectories of the ribosome as a Brownian nanomachine. *Proceedings of the National Academy of Sciences of the United States of America*, 111(49):17492–7.

[19] De Rosier, D. J. and Klug, a. (1968). Reconstruction of Three Dimensional Structures from Electron Micrographs. *Nature*, 217:130–134.

[20] DeLaBarre, B. and Brunger, A. T. (2006). Considerations for the refinement of low-resolution crystal structures. *Acta Crystallographica Section D*, 62(8):923–932.

[21] EMDataBank (2015). www.emdatabank.org.

[22] Emsley, P. and Cowtan, K. (2004). Coot: Model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, 60(12 I):2126–2132.

[23] Falkner, B. (2012). *Cryo-Electron Microscopy - Estimating Conformational Variances by Principal Motion Analysis*. PhD thesis.

[24] Falkner, B. and Schröder, G. F. (2013). Cross-validation in cryo-EM-based structural modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 110(22):8930–5.

[25] Fernández, J. J., Luque, D., Castón, J. R., and Carrascosa, J. L. (2008). Sharpening high resolution information in single particle electron cryomicroscopy. *Journal of Structural Biology*, 164(1):170–175.

[26] Fischer, N., Konevega, A. L., Wintermeyer, W., Rodnina, M. V., and Stark, H. (2010). Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature*, 466(7304):329–333.

[27] Fischer, N., Neumann, P., Bock, L. V., Maracci, C., Wang, Z., Paleskava, A., Konevega, A. L., Schröder, G. F., Grubmüller, H., Ficner, R., Rodnina, M. V., and Stark, H. (2016). The pathway to GTPase activation of elongation factor SelB on the ribosome. *Nature*, 540(7631):1–20.

[28] Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annual review of biophysics and biomolecular structure*, 31(1):303–319.

[29] Gao, H., Valle, M., Ehrenberg, M., and Frank, J. (2004). Dynamics of EF-G interaction with the ribosome explored by classification of a heterogeneous cryo-EM dataset. *Journal of Structural Biology*, 147(3):283–290.

[30] Goddard, T. D., Huang, C. C., and Ferrin, T. E. (2007). Visualizing density maps with UCSF Chimera. *Journal of Structural Biology*, 157(1):281–287.

[31] Grassucci, R. A., Taylor, D., and Frank, J. (2008). Visualization of macromolecular complexes using cryo-electron microscopy with FEI Tecnai transmission electron microscopes. *Nature protocols*, 3(2):330–9.

[32] Grassucci, R. A., Taylor, D. J., and Frank, J. (2007). Preparation of macromolecular complexes for cryo-electron microscopy. *Nature protocols*, 2(12):3239–46.

[33] Grigorieff, N. (2007). FREALIGN: High-resolution refinement of single particle structures. *Journal of Structural Biology*, 157(1):117–125.

[34] Groot, B. L. D., Aalten, D. M. F. V., Scheek, R. M., Amadei, A., Vriend, G., and Berendsen, H. J. C. (1997). Prediction of Protein Conformational Freedom From Distance Constraints. *Proteins Structure Function and Genetics*, 29(2):240–251.

[35] Habeck, M., Hirsch, M., Scho, B., and Al, H. E. T. (2011). A Blind Deconvolution Approach for Improving the Resolution of Cryo-EM Density Maps. 18(3):335–346.

[36] Hawkes, P. W. (2009). Aberration correction past and present. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1903):3637–3664.

[37] Henderson, R. (2013). Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences of the United States of America*, 110(45):18037–41.

[38] Howlin, B and Butler, SA and Moss, DS and Harris, GW and Driessen, H. (1993). TLSANL: TLS parameter-analysis program for segmented anisotropic refinement of macromolecular structures. *Journal of applied crystallography*, 26(4):622—-624.

[39] Iancu, C. V., Tivol, W. F., Schooler, J. B., Dias, D. P., Henderson, G. P., Murphy, G. E., Wright, E. R., Li, Z., Yu, Z., Briegel, A., Gan, L., He, Y., and Jensen, G. J. (2007). Electron cryotomography sample preparation using the Vitrobot. *Nature protocols*, 1(6):2813–19.

[40] Jin, Q., Sorzano, C. O. S., de la Rosa-Trevın, J. M., Bilbao-Castro, J. R., Nunez-Ramırez, R., Llorca, O., Tama, F., Jonic, and Slavica (2014). Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure*, 22(3):496–506.

[41] Katsevich, E., Katsevich, A., and Singer, A. (2015). Covariance Matrix Estimation for the Cryo-EM Heterogeneity Problem. *SIAM journal on imaging sciences*, 8(1):126–185.

[42] Kishchenko, G. P. and Leith, A. (2014). Spherical deconvolution improves quality of single particle reconstruction. *Journal of Structural Biology*, 187(1):84–92.

[43] Kucukelbir, A., Sigworth, F. J., and Tagare, H. D. (2014). Quantifying the local resolution of cryo-EM density maps. *Nature methods*, 11(1):63–5.

[44] Kühlbrandt, W. (2014). The Resolution Revolution. *Science*, 343(6178):1443–1444.

[45] Liao, H. Y., Hashem, Y., Liao, H. Y., Hashem, Y., and Frank, J. (2015). Efficient Estimation of Three-Dimensional Covariance and its Application in the Analysis of Heterogeneous Samples in Cryo-Electron Microscopy. *Structure*, 23(6):1129–1137.

[46] Liao, M., Cao, E., Julius, D., and Cheng, Y. (2013). Structure of the TRPV1 ion channel determined by electron cryo- microscopy. *Nature*, 504(7478):107–112.

[47] Lopéz-Blanco, J., Ramón, J., and Chacón, P. (2013). IMODFIT: Efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *Journal of Structural Biology*, 184(2):261–270.

[48] Ludtke, S. J., Baldwin, P. R., and Chiu, W. (1999). EMAN: semi-automated software for high-resolution single-particle reconstructions. *Journal of structural biology*, 128(1):82–97.

[49] Lyumkis, D., Vinterbo, S., Potter, C. S., and Carragher, B. (2013). Optimod - An automated approach for constructing and optimizing initial models for single-particle electron microscopy. *Journal of Structural Biology*, 184(3):417–426.

[50] Mao, Y., Wang, L., Gu, C., Herschhorn, A., Désormeaux, A., and Finzi, A. (2013). Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. *Proceedings of the National Academy of Sciences*, 110(30):12438–12443.

[51] McMullan, G., Chen, S., Henderson, R., and Faruqi, A. R. (2009). Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy*, 109(9):1126–1143.

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

[53] Nogales-Cadenas, R., Jonic, S., Tama, F., Arteni, A. A., Tabas-Madrid, D., Va´zquez, M., Pascual-Montano, A., and Sorzano, C. O. (2013). 3DEM Loupe: Analysis of macromolecular dynamics using structures from electron microscopy. *Nucleic acids research*, 41(W1):363–367.

[54] Orlova, E. V. and Saibil, H. R. (2011). Structural analysis of macromolecular assemblies by electron microscopy. *Chemical Reviews*, 111(12):7710–7748.

[55] Park, W., Madden, D. R., Rockmore, D. N., and Chirikjian, G. S. (2010). Deblurring of Class-Averaged Images in Single-Particle Electron Microscopy. *Inverse problems*, 26(3):3500521–35005229.

[56] Penczek, P. A. (2002). Variance in three-dimensional reconstructions from projections. *Proceedings - International Symposium on Biomedical Imaging*, pages 749–752.

[57] Penczek, P. A., Frank, J., and Spahn, C. M. T. (2006a). A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *Journal of Structural Biology*, 154(2):184–194.

[58] Penczek, P. A., Frank, J., and Spahn, C. M. T. (2006b). Conformational Heterogeneity of Macromolecules Analyzed by Cryo-Electron Microscopy. *Microscopy and Microanalysis*, 12(S02):386–387.

[59] Penczek, P. A., Kimmel, M., and Spahn, C. M. T. (2011). Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images. *Structure*, 19(11):1582–1590.

[60] Peng, L. M., Ren, G., Dudarev, S. L., and Whelan, M. J. (1996). Robust Parameterization of Elastic and Absorptive Electron Atomic Scattering Factors. *Acta Crystallographica Section A*, 52(2):257–276.

[61] Ranson, N. A., Clare, D. K., Farr, G. W., Houldershaw, D., Horwich, A. L., and Saibil, H. R. (2006). Allosteric signalling of ATP hydrolysis in GroEL-GroES complexes. *Nature structural & molecular biology*, 13(2):147–152.

[62] Rohou, A. and Grigorieff, N. (2015). CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2):216–221.

[63] Rosenthal, P. B. and Henderson, R. (2003). Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 333(4):721–745.

[64] Scheres, S. (2016). *Processing of Structurally Heterogeneous Cryo-EM Data in RELION*, volume 579. Elsevier Inc., 1 edition.

[65] Scheres, S. H., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P., Frank, J., and Carazo, J. M. (2007). Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature methods*, 4(1):27–29.

[66] Scheres, S. H. W. (2010). *Classification of structural heterogeneity by maximum-likelihood methods*, volume 482. Elsevier Inc., 1 edition.

[67] Scheres, S. H. W. (2012a). A bayesian view on cryo-EM structure determination. *Journal of Molecular Biology*, 415(2):406–418.

[68] Scheres, S. H. W. (2012b). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530.

[69] Scheres, S. H. W. (2015). Semi-automated selection of cryo-EM particles in RELION-1.3. *Journal of Structural Biology*, 189(2):114–122.

[70] Scheres, S. H. W. and Chen, S. (2012). Prevention of overfitting in cryo-EM structure determination. *Nature Methods*, 9(9):853–854.

[71] Scherzer, O. (1949). The Theoretical Resolution Limit of the Electron Microscope. *Journal of Applied Physics*, 20(1):20–29.

[72] Schröder, G. F., Brunger, A. T., and Levitt, M. (2007). Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution. *Structure*, 15(12):1630–1641.

[73] Sigworth, F. (1998). A maximum-likelihood approach to single-particle image refinement. *Journal of Structural Biology*, 122(3):328–339.

[74] Sigworth, F. J., Doerschuk, P. C., Carazo, J. M., and Scheres, S. H. W. (2010). *An introduction to maximum-likelihood methods in cryo-EM*, volume 482. Elsevier Inc., 1 edition.

[75] Spahn, C. M. and Penczek, P. A. (2009). Exploring conformational modes of macromolecular assemblies by multi-particle cryo-EM. *Current opinion in structural biology*, 19(5):623–631.

[76] Spiegel, M., Duraisamy, A. K., and Schröder, G. F. (2015). Improving the visualization of cryo-EM density reconstructions. *Journal of Structural Biology*, 191(2):207–213.

[77] Subramaniam, S. (2013). Structure of trimeric HIV-1 envelope glycoproteins. *Proceedings of the National Academy of Sciences of the United States of America*, 110(45):E4172–4.

[78] Tagare, H. D., Kucukelbir, A., Sigworth, F. J., Wang, H., and Rao, M. (2015). Directly reconstructing principal components of heterogeneous particles from cryo-EM images. *Journal of Structural Biology*, 191(2):245–262.

[79] Tama, F., Valle, M., Frank, J., and Iii, C. L. B. (2003). Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proceedings of the National Academy of Sciences*, 100(16):9319—-9323.

[80] Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., and Ludtke, S. J. (2007). EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46.

[81] Trabuco, L. G., Schreiner, E., Eargle, J., Cornish, P., Ha, T., Luthey-Schulten, Z., and Schulten, K. (2010). The Role of L1 Stalk-tRNA Interaction in the Ribosome Elongation Cycle. *Journal of Molecular Biology*, 402(4):741–760.

[82] Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. *Structure*, 16(5):673–683.

[83] Tsai, J., Taylor, R., Chothia, C., and Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *Journal of molecular biology*, 290(1):253–66.

[84] Unverdorben, P., Beck, F., Sledz, P., Schweitzer, A., Pfeifer, G., Plitzko, J. M., Baumeister, W., and Förster, F. (2014). Deep classification of a large cryo-EM dataset defines the conformational landscape of the 26S proteasome. *Proceedings of the National Academy of Sciences*, 111(15):5544—-5549.

[85] Van Heel, M. (1987). Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*, 21(2):111–123.

[86] van Heel, M. (2013). Finding trimeric HIV-1 envelope glycoproteins in random noise. *Proceedings of the National Academy of Sciences*, 110(45):E4175–4177.

[87] van Heel, M., Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., and Patwardhan, A. (2000). Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly reviews of biophysics*, 33(4):307–369.

[88] van Heel, M. and Harauz, G. (1986). Resolution criteria for three dimensional reconstruction.

[89] van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. *Journal of structural biology*, 116(1):17–24.

[90] van Heel, M. and Keegstra, W. (1981). IMAGIC: A fast, flexible and friendly image analysis software system. *Ultramicroscopy*, 7(2):113–129.

[91] van Heel, M. and Stöffler-Meilicke, M. (1985). Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope. *The EMBO journal*, 4(9):2389–95.

[92] Vargas, J., Abrishami, V., Marabini, R., de la Rosa-Trevín, J. M., Zaldivar, A., Carazo, J. M., and Sorzano, C. O. S. (2013). Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *Journal of Structural Biology*, 183(3):342–353.

[93] Wang, B.-C. (1985). Resolution of phase ambiguity in macromolecular crystallography. In *Diffraction Methods for Biological Macromolecules Part B*, volume 115 of *Methods in Enzymology*, pages 90–112. Academic Press.

[94] Wang, Q., Matsui, T., Domitrovic, T., Zheng, Y., Doerschuk, P. C., and Johnson, J. E. (2013). Dynamics in cryo EM reconstructions visualized with maximum-likelihood derived variance maps. *Journal of Structural Biology*, 181(3):195–206.

[95] Xu, Z., Horwich, a. L., and Sigler, P. B. (1997). The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. *Nature*, 388(6644):741–750.

[96] Zhang, K. Y. J. and Main, P. (1990). Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Crystallographica Section A*, 46(1):41–46.

[97] Zhang, W., Kimmel, M., Spahn, C. M. T., and Penczek, P. A. (2008). Technical Advance Heterogeneity of Large Macromolecular Complexes Revealed by 3D Cryo-EM Variance Analysis. *Structure*, 16(12):1770–1776.

[98] Zhu, Y., Carragher, B., Glaeser, R. M., Fellmann, D., Bajaj, C., Bern, M., Mouche, F., Haas, F. D., Hall, R. J., Kriegman, D. J., Ludtke, S. J., Mallick, S. P., Penczek, P. A., Roseman, A. M., Sigworth, F. J., Volkmann, N., and Potter, C. S. (2004). Automatic particle selection: results of a comparative study. *Journal of Structural Biology*, 145:3–14.