Cryo-Electron Microscopy

Estimating Conformational Variances by Principal Motion Analysis

Inaugural-Dissertation

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

vorgelegt von
Benjamin Falkner
aus Wickede

Düsseldorf, 10. September 2012

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

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

Referent: Jun.-Prof. Dr. Gunnar F. Schröder
Korreferent: Prof. Dr. Stefan Egelhaaf

Tag der mündlichen Prüfung: 25. 10. 2012

Abstract

In der Einzelpartikel-Cryo-Elektronenmikroskopie (Cryo-EM) enthalten die Aufnahmen zweidimensionale Projektionsbilder einer Vielzahl von Kopien des gleichen Proteins. Diese Proteine befinden sich in leicht unterschiedlichen Konformationen, wodurch die Varianz der Daten erhöht wird. In der Regel wird aus den Projektionsbildern eine einzelne dreidimensionale atomare Dichte rekonstruiert, wobei allerdings die konformationelle Heterogenität der Probe vernachlässigt wird. In dieser Arbeit liegt der Schwerpunkt auf der Entwicklung einer Methode mit der die Varianz der Projektionsbilder als dreidimensionale Konformationsbewegungen des Proteins interpretiert werden kann.

Da die Varianz der Probe die Auflösung der 3D-Rekonstruktion beschränkt und bisher nicht genutzt wurde, um atomistische Informationen erhalten, wurde die Bootstrapping-Technik verwendet, um mehrere dreidimensionale Dichten aus einem Experiment zu rekonstruieren, die gemeinsam die Varianz der Probe enthalten. Die Principal Component Analysis (PCA)(dt. Hautptkomponentenanalyse) auf diesen 3D-Dichten, die korrelierte Konformationsänderungen der Volumen erkennt, wird hier durch die neu entwickelte Principal Motion Analysis (PMA) ergänzt, die atomistische globale Bewegungen des Proteins detektieren kann.

Die PMA ist empfindlicher gegenüber Konformationsänderungen als die Volumen PCA. Dieses neue Verfahren besteht aus drei wichtigen Schritten: Bootstrapping der Bilder, um ein Volumen Ensemble zu erhalten, atomistisches Refinement, um das Volumen-Ensemble auf ein atomistisches Ensemble abzubilden und schließlich eine PCA-Transformation auf dem atomistischen Ensemble.

Die PMA wurde auf zwei Chaperone (GroEL/ES und Mm-CPN) angewendet, welche als Teil ihrer Funktion großen Konformationsänderungen durchführen. In beiden Fällen kann die Varianz der experimentellen Daten in großen Teilen als Schwankungen interpretiert werden, die den bekannten Konformationsänderungen entsprechen. Um sicherzustellen, dass diese Ergebnisse zuverlässig sind, wurden verschiedene Validierungsverfahren entwickelt.

Um die Eigenvektorberechnungen auf Volumina und atomistischen Daten dieser Größe ausführen zu können, wurde darüber hinaus ein schneller inverser Eigenwert-Solver entwickelt.

Weiterhin wurde ein Kreuz-Validierungsverfahren für das Refinement von atomistischen Strukturen gegen niedrigaufgelöste Dichten entwickelt. Dieses Verfahren verwendet eine unabhängige Schale von Raumfrequenzen als freien Datensatz. Durch Berechnung der Kreuzkorrelation der sich ergebenden Struktur mit den

freien Daten wird ein Qualitätsfaktor gewonnen. Dieser kann weiter zur Optimierung von Parametern genutzt werden.

Abstract

In single particle cryo-electron microscopy (cryo-EM) the micrographs contain 2D projection images of a large number of copies of the same protein. These proteins are typically in slightly different conformations, which increases the variance of the data. In general a single 3D reconstruction is calculated from the projection images ignoring the heterogeneity of the specimen. In this work a method is developed, which interprets the variance of the projection images as conformational motions of the protein.

While the variance of the specimen is limiting the resolution of the 3D reconstruction and is not used to obtain atomistic information, the bootstrapping technique was applied to generate multiple 3D volumes which represent the variance of the specimen. The Principal Component analysis (PCA) on these volumes, which detects correlated conformational volumetric changes, is extended by the newly developed Principal Motion Analysis (PMA), which determines global atomistic motions of the protein. The PMA is more sensitive to conformational changes than a volume PCA. This new method consists of three important steps: 1) bootstrapping of the images to obtain a volume ensemble, 2) atomistic refinement to translate the volume ensemble into an atomistic ensemble and 3) a PCA on the atomistic ensemble.

The PMA was applied to two chaperonins (GroEL/ES and Mm-CPN), that are known for high flexibility and which undergo large conformational changes upon executing their function. In both cases the variance of the projection images can be interpreted as conformational changes, that are to a large extent in agreement with known or suggested motions of these proteins.

To ensure that the results are reliable several validation approaches have been developed.

To perform these eigenvector calculations on these very large volumes and atomic models a fast inverse Eigenvalue solver was developed for this special kind of problems.

Further a cross-validation method for the refinement of atomistic structures against low resolution densities was developed. This method uses an independent shell of spatial frequencies as free data that are not used in the refinement. By calculation the cross-correlation of the resulting structure with the free data an independent quality measure is obtained. This allows to further optimize the parameters in the refinement.

Contents

Vorwort

Diese Thesis fasst einen Großteil meiner Arbeit aus den letzten drei Jahren zusammen und dient der Erlangung des Doktorgrades Dr. rer. nat. an der Heinrich-Heine-Universität Düsseldorf. Sie ist ein Résumé meiner Forschung und deren Grundlangen im Bereich der Analyse von Cryo-EM (Elektronenmikroskopie) Bildern. Die Grundlangen sind die Funktionsweisen der EM, die Rekonstruktion der Bilder zu 3D Volumen und das Refinement von atomaren Strukturen an diese Volumen. Diese werden in dem ersten Teil beschrieben, im zweiten Teil folgen di von mir entwickelten Methoden und die Ergebnisse der Analyse von den Chaperone GroEL/ES und Mm-CPN.
Die ersten Kapitel sind sehr mathematisch gehalten und präsentieren die Grundlagen stark kondensiert. Dies ist nicht immer leicht zu verstehen, deshalb habe ich versucht möglichst häufig Abbildungen hinzuzufügen. Dennoch ist es kein Lehrbuch geworden sondern eine kurze Beschreibung der gängigen Methoden und Approximationen in diesem Feld. Im ersten Teil ist auch die Sparse PCA enthalten, eine von mir getroffene Umformulierung der Hauptkomponentenanalyse (PCA - Principal Component Analysis), welche eine deutlich schnellere Berechnung der PCA in bestimmten Fällen ermöglich und mir die Arbeit sehr erleichtert hat.
Der zweite Teil ist deutlich ausführlicher geschrieben und beginnt mit einem Kapitel zur Validierung des atomaren Refinements. Nach diesem Kapitel beginnt der Hauptteil der Arbeit die Principal Motion Analysis (PMA). In diesem Teil sehe ich den Schwerpunkt der Arbeit und glaube eine Technik entwickelt zu haben, welche Potential für die Zukunft bietet. Ich habe versucht diesen Teil möglichst verständlich zu schreiben und die grundlegenden Ideen Schritt für Schritt zu erklären. Hier gilt mein besondere Dank meinem Betreuer Gunnar Schröder, der sich die Zeit genommen hat meine wirren Gedanken zu ordnen und in verständliche Sätze zu betten.
In den letzten drei Jahren habe ich viel über die Biophysik, Proteine, Statistik und deren interdiziplinäre Verknüpfung gelernt. Es war eine sehr spannende Zeit und eine gute heterogene Gruppe im FZ-Jülich. Die Unterstützung aus der Gruppe war immer hervorragend. Somit möchte ich allen danken, Kumaran Baskaran, André Wildberg, Wang Zhe und Gunnar Schröder, da sie alle Anteil an dieser Arbeit haben.
Die Cryo-EM Experimente wurden von Junjie Zhang and Chen Donghua am Baylor College of Medicine in Houston im Labor von Wah Chiu durchgeführt. Sie waren bei Nachfragen zu dem Experiment immer hilfreich und bereit umfassend über die Methoden zu informieren.
Zum Schluss möchte ich mich noch bei meiner Frau und Tochter bedanken, da sie meine unendlichen Geschichten über Physik und Mathematik ohne Widerworte

ertragen haben, obwohl sie sie niemals hören wollten. Ebenso haben sie sich nie beschwert, wenn sie wegen meiner Arbeit zurückstecken mussten.
Ich bin sehr glücklich, dass ich die Chance zu dieser Arbeit hatte und hoffe, dass sie einige Leser interessieren wird.

Benjamin Falkner

1 Chapter
Introduction

1.1 Cryo Electron Microscopy

Cryo-electron microscopy (Cryo-EM) is an emerging technique to determine the structure of large macromolecular complexes. The resolution limit has been constantly pushed to higher resolutions, where in some cases resolutions below 4 Å were achieved in recent years. Cryo-EM shows great promise to be able to routinely determine atomic structures of macromolecules, and it can be expected that its importance as a structure determination technique will continuously grow.
In the cryo-EM experiment micrographs are recorded which contain projection images of single particles typically in different orientations. From the projection images a three-dimensional density distribution can be reconstructed, which necessarily averages over these individual particles[53]. In cryo-EM a large number (typically 10^4 to 10^6) of individual protein projections are imaged in different random orientations (while the orientations are not necessarily equally distributed). From these different views of the protein a 3D density distribution is back-projected. Today it is typically to reach resolutions in the range from about 6 to 20 Å. At such resolutions it is typically not possible to directly build atomic structures. This is a limiting factor for cryo-EM compared to X-ray crystallography or NMR spectroscopy, where atomic structures can be determined directly from the data.
In most cases, cryo-EM experiments are interpreted by placing high resolution structures determined by either X-ray crystallography or NMR into the cryo-EM density map. At low resolution of less than 15 Å the density map defines only the overall shape but no internal details. In such a situation the proteins are placed only as rigid bodies into the density. Therefore several methods have been developed that perform a rigid body refinement of structural elements into a predefined shape [68, 97].
At higher resolutions, conformational differences to the high-resolution structures could become apparent. In that case flexible fitting methods can be used, that are able to deform the atomic structures by shifting atoms individually to optimize a measure of fit to the reconstructed volume. For this purpose several methods

have been developed that are able to refine atomic structures into density maps [96, 90, 86, 59, 19].
To ensure a chemically reasonable structure it is necessary to introduce either a forcefield or some kind of restraints. In addition, since the molecules that are studied with cryo-EM are usually very large, a large number of parameters have to be fitted, which is always accompanied by a danger of overfitting. Forcefields and/or restraints are typically used to reduce the amount of over-fitting. Approaches are either partially [90], or completely based on restraints, like the Deformable Elastic Network (DEN) method[78, 80]. In all methods the number and strength of the restraints has to estimated to yield a good fit to the data but at the same time to avoid overfitting.
In a regular 3D reconstruction it is usually assumed that all proteins are in the same conformation. This is, however, in general not the case: large macromolecules have an inherent flexibility and even if the particle images are sorted into classes based on their mutual similarity, there will always be some residual variance among the particles. The power of cryo-EM lies in the fact that in contrast to other techniques the observables are actually single particles and not ensemble averages. That means we have at least in principle access to the full distribution of conformational states present in the sample. Therefore there is more information in the cryo-EM data than just a static average structure: the individual particle images show the molecules in slightly different conformations according to their equilibrium distribution in the sample as defined by the experimental conditions. While this structural heterogeneity is often considered a nuisance as it fundamentally limits the achievable resolution, appropriate analysis of this heterogeneity could potentially reveal functionally highly relevant motions[64, 63, 82]. Extracting these functional motions is however a significant computational challenge since the information content of a single particle image is low due to a low signal to noise ratio. Standard 3D reconstruction procedures typically average over all particles by which all information about the conformational flexibility is lost.
The amount of data that needs to be analyzed in cryo-EM experiments is very large. The development of efficient algorithms is therefore key to an exhaustive analysis and to maximize the information that can be extracted from the data.

1.2 Cross-Validation of the Refinement

At low resolution the parameter to observable ration is large, in particular for large macromolecules, which commonly causes overfitting and, thus, results in wrong or flawed models. To be able to detect overfitting in cryo-EM based refinement is an important prerequisite for the optimal interpretation of cryo-EM density maps.
It is usually necessary to use restraints during the structure refinement to avoid overfitting. The question is however how to optimally choose the restraints and their relative strengths? On the one hand, too few or weak restraints result in overfitting, and on the other hand side too many or strong restraint would yield an insufficient fit to the data.
A solution to this question is given by the concept of cross-validation, which has been introduced to the closely related problem of X-ray crystallographic refinement

almost 20 years ago [24] and has in the following drastically increased the reliability of refined crystal structures. The idea is to leave out part of the data (the 'test set') that is not used for the refinement but only for assessing the refined model. In crystallography typically 10% of the structure factors are randomly chosen as the test set, while the remaining 90% of the structure factors (the 'work set') are used for refining the structure.
A crucial prerequisite for the cross-validation is that the information in the test set is independent from the information in the work set. For diffraction data this assumption is usually justified. However, due to the very different nature of the experiment, for cryo-EM density maps this assumption does not generally hold. In this work the crystallographic cross-validation approach is adapted to structure refinement against cryo-EM data. The method is tested on three proteins with simulated data, where the target structure is known and furthermore, we apply the method to the refinement of a GroEL crystal structure against a 5.4 Å experimental cryo-EM density map.

1.3 Principal Motion Analysis

The heterogeneous ensemble of single particles in the specimen offers more information than the averaged density. Due to the fact that the observables are single particles the conformational space of the ensemble can be explored by statistical methods. This space is only a small part of the full conformational space of the protein. To reveal at least part of the sample heterogeneity, several approaches have been described, for example to sort particle images into classes that belong to distinct protein conformational states [25, 104, 28, 91]. A 3D density can then be reconstructed for each of these classes separately and the ensemble will represent the individual conformational states. These approaches can be successful if the conformations are clearly separable and enough data is available, but usually fail for relatively small continuous conformational fluctuations. For the case of continuous conformational variations a bootstrapping method has been proposed to calculate an ensemble of different density maps from which the density variance can be obtained[64]. Further this was used to rebuild conformationally different volumes by using the principal components of the bootstrapped ensemble called eigenvolumes [63, 82]. This will provide additional information about the conformational space but can only be referred to atomistic changes if the volumes are distinct. This approach can be easily be applied to all typical cryo-EM datasets.
In this work a method is presented which is able to determine large-scale correlated motions of a protein in near atomic detail in an atomistic representation from such a bootstrapped ensemble. The method is applied here to determine large-scale correlated motions (principal motions) of two large proteins. Both proteins are chaperonins which are well known for undergoing large conformational changes between an open and a closed state and are therefore a appropriate system for motion analysis. Chaperonins are multimeric barrel-like protein complexes. They consist of two rings that are stacked back-to-back forming two large cavities. They play an essential role in mediating protein folding, which is assumed to take place

inside these cavities. Furthermore chaperonins are also assumed to be involved in multiple diseases like cancer and neurodegeneration.
In general two groups of chaperonins are distinguished: group I chaperonins (e.g. GroEL) use a co-factor (e.g. GroES) as a lid to close the cavity during substrate folding, whereas group II chaperonins (e.g. Methanococcus maripaludis chaperonin or Mm-CPN) can close the cavity without an additional co-factor by undergoing a large conformational rearrangement of the ring structure.
Group I chaperonins are found in bacteria as well as organelles of endosymbiotic origin, while group II chaperonins are the chaperonins of the eukaryotic cytosol and the archaea.
We present a study of the motions of the group I chaperonin structure of GroEL which was in the so called bullet-shaped state, where one side was closed by the co-factor GroES. The structure of GroEL alone and in complex with GroES have been determined by X-ray crystallography [9, 98]. GroEL/ES is responsible for folding about one third of all proteins in bacteria.
As second molecule the group II chaperonin Mm-CPN (from Methanococcus maripaludis) was investigated. Mm-CPN is analyzed in a state where both cavities are open. In this open state the subunits make only few contacts with neighboring subunits and, thus, and can be expected to be very flexible. This large scale flexibility will be a challenge for the principal motion analysis. Mm-CPN is a close homologue of the thermosome whose crystal structure has been determined by Ditzel et al. [20].

2

Chapter

Recording and Reconstruction of Cryo-EM data

2.1 Experiment

2.1.1 Electron Microscope

The Electron Microscope is a type of microscope that can be used to image the surface of a specimen, using backscattered electrons and photon emissions, or the inner of a specimen by measuring the transmission of electrons. In cryo-EM only transmission microscopes are used to record the data, because the inner part of the specimen becomes visible and in general the resolution is higher than in reflection based microscopes.
The transmission electron microscope was invented in the early 1930 by Knoll and Ruska and had a resolution that was not better than light microscopes[46]. But already in this time there have been speculations about atomic resolutions. The basic setup of an electron microscope is almost always the same and starts with an electron source, that can be a thermionic tube or a field emission source. The next elements are a coherence filter based on B-field separation of electrons by their velocities and an E-Field based accelerator[26].
The simple magnetic electron lens consists of a coil of wire surrounded by a magnetic material, which is shaped to modify the magnetic field and will effect the focusing of the beam. In the middle of the lens is a gap for the electron beam to pass through the magnetic field. In general these lenses can be compared to optical lenses an are used in the same fashion[44].
The next element in the electron microscope is the specimen holder followed by objective lenses and a recording device.
The electrons in the beam have a small mass so even electrons with a small energy travel with roughly half the speed of light and have to be treated in a relativistic way. The de-Broglie wavelength λ of the electron is:

$$\lambda = \frac{h}{p} \tag{2.1}$$

where h is the Planck's constant and p is the impulse. The energy E can be calculated as:

$$E^2 = \left(m_0c^2 + eV\right)^2 = p^2c^2 + m_0^2c^4 \tag{2.2}$$

where c is the speed of light in vacuum, V is the acceleration potential, e the charge, m the mass of the electron, m_0 the rest mass of the electron, $p = mv$ the momentum and v the velocity. Substituting the energy function yields:

$$\begin{aligned} \left(m_0c^2 + eV\right)^2 &= \left(\frac{hc}{\lambda}\right)^2 + m_0^2c^4 \\ \Rightarrow \lambda &= \frac{hc}{\sqrt{eV(2m_0c^2 + eV)}} \\ &= \frac{12.3[keV\mathring{A}]}{\sqrt{eV(1022[keV] + eV)}} \end{aligned} \tag{2.3}$$

The frequency can be used to estimate the maximal resolution R of the electron microscope using the Abbe's Equation:

$$R = \frac{1.22\lambda}{2n\sin\theta} \tag{2.4}$$

where $n = 1$ is the refraction index and θ is the half-angle of maximum cone of light, which is very small on electron microscopes and its sinus can be approximated as 10^{-2} or smaller. All together the resolution can be approximated by:

$$R = 61\lambda \tag{2.5}$$

The theoretical resolution[1] is much lower than the resolution obtained in the experiments because of the aberration and the magnetic lenses will lower the resolutions dramatically and are one big problem in electron microscopy. This aberrations have to be corrected in the experiment by focusing the beam and tilting the specimen layer or later by filtering the micro graphs. Today the best resolutions in cryo-EM after reconstruction are up to $6 - 3$ $\mathring{A}$ while the theoretical resolution is a hundredth part of todays best resolutions.
In structure biology the electron microscope is an alternative to the crystallography to obtain images and structural informations of particles in a more native state. The biological material is introduced to the electron microscope as a thin film of amorphous ice . By this it is possible to obtain images of fully hydrated macromolecules. Usually liquid ethane is used for rapid freezing an aqueous solution dispersed by the specimen. Rapid freezing is necessary to avoid the water from forming several crystals instead of a more homogeneous ice layer. In general the ice layer has to be less than 500 nm to avoid multiple scattering events, for thicker specimen more complicated techniques have to be use, like special freezing methods and cutting those blocks into thin sections. This is no problem for proteins because their size is smaller than the scattering limitation, which is more important for bacteriophages or even larger biological structures. But even for thin ice there will be an effect on the noise of the images[26].

[1] For 300 keV electrons is the theoretical Wavelength about 0.038 Å.

In the electron microscope an image is taken from frozen specimen containing several particles and each single particle is recognized as a part of the information. In fact this induces one of the issues for the work with the images and the preparation of the experiment, because several conformations of the object will be in the dataset, which can have different effects:

- conformations are distinct: the specimen can be spilt into multiple conformations, which will reduce the amount of information per conformations;
- conformations are not distinct: slightly different conformations will be mixed up and an additional noise introduced to the dataset.

On the other hand the single particles vary in their orientations and show different views of akin particles. This can be used to regain the three dimensional shape of the specimen.
Another problem in the experiment is that biological macromolecules are extremely sensitive to radiation, which will haven an effect on the time duration of the electron beam and the energy of the beam. This limits the experiment and will increase the noise of the images and finally reduce the resolution. At this point the low temperature will save the biological specimen from the radiation and will help to increase the dose and the energy of the electrons[26].
All this together is mostly a reason for a low signal-to-noise ratio (SNR) in the micro graphs and will effect the resolution of the conformational information of the specimen.
Today the focus is on bright field electron microscopy because of its high contrast at lower doses, It allows a very easy way to calculate the effects in the electron microscope. In this context there is no difference between a bright field conventional transmission electron microscope (BF-CTEM) and a bright field scanning transmission electron microscope (BF-STEM). A very important feature is that the image model can be assumed to be linear[26, 32].
The process of the electron scattering in the specimen is an elastic scattering or Rutherford scattering on the Coulomb potential $f(x, y, z) = f(\mathbf{x})$ of the specimen. A positive potential will accelerate the electrons and the wavelength will decease which is a reason for a phase shift. The beam is parallel to the z-axis with wave function $\exp(2\pi i z/\lambda)$. When the specimen is also weak, so that the phase shift φ will be proportional to the Coulomb potential :

$$\varphi(\mathbf{x}) = \sigma f(\mathbf{x}) \tag{2.6}$$

where σ is a scaling factor and the resulting wave function is approximated by:

$$\psi(\mathbf{x}) \sim \exp\left(2\pi i \frac{z}{\lambda}\right) \exp\left(i\sigma f(\mathbf{x})\right). \tag{2.7}$$

The high energy electrons (100keV-300keV)[2] will path through the thin sample with only small deviations, so that the potential can be flattened first to an 2D projection $p(x, y) = p(\mathbf{x})$ along the optical axis z[32]:

$$p(x, y) = \int f(x, y, z) dz \tag{2.8}$$

[2] in general the energy is approximately 100 keV to 1000 keV, but too high energies will destroy the biological specimen.[44]

On the other hand it can be computed by solving the system with Bloch waves which goes beyond the scope of this work [44].
Due to the linear image model the transmission function $\phi = \exp(i\varphi(\mathbf{x})) \sim \exp(i\sigma p)$ can be described as an occlusion or absorption process and the incoming electron wave function ψ_{inc} will be modified:

$$\psi_{spec}(\mathbf{x}) = \phi(\mathbf{x})\psi_{inc}(\mathbf{x}) \tag{2.9}$$

where $\mathbf{x} = (x, y)^T$ is a two dimensional vector in the projection plane of the beam. The incoming wave function in a CTEM can be assumed to be a plan wave with constant intensity propagating in z direction. Further the wave function is monochromatic, so that the amplitudes can be estimated as 1 without loss of generality. In weak phase object (WPO) approximation the sample has to be very thin, that is the case for the biological material in cryo-EM, and so the the specimen will create a only a small phase shifts in the wave functions of the electrons[58]. So the wave functions yields:

$$\psi_{spec}(\mathbf{x}) = \phi(\mathbf{x})\psi_{inc}(\mathbf{x}) \sim \phi(\mathbf{x}) \sim \exp(i\sigma p(\mathbf{x})) \tag{2.10}$$

Using a power series approximation will be helpful for further simplifications:

$$\exp(i\sigma p(\mathbf{x})) \sim 1 + i\sigma p(\mathbf{x}) + \cdots \tag{2.11}$$

The effect of the magnetic lenses is a phase shift $\chi(\mathbf{k})$[3] where $\mathbf{k}$ is the spatial frequency vector. This aberration is a convolution ($\circ$) in real space by the point spread function(PSF) H and can be easily expressed as the product in Fourier space by:

$$\Psi_i(\mathbf{k}) = \Psi_{spec}(\mathbf{k}) \cdot \exp(-i\chi(\mathbf{k})) \tag{2.12}$$

where $\Psi_{spec}(\mathbf{k}) = \mathcal{F}(\Psi_{spec})(\mathbf{x})$ and $\mathcal{F}$ is the Fourier transform. The recorded images are the intensities $I(\mathbf{x})$ of the wave functions behind the objective lens $\psi_i(\mathbf{x}) = \mathcal{F}^{-1}(\Psi_i)(\mathbf{k})$:

$$I(\mathbf{x}) = |\psi_i(\mathbf{k})|^2 \tag{2.13}$$

Using all approximations for the intensity function, the new term adds up to:

$$\begin{aligned} I(\mathbf{x}) &= |\exp(i\sigma p(\mathbf{x})) \circ H(\mathbf{x})|^2 \\ &= |(1 + i\sigma p(\mathbf{x})) \circ H(\mathbf{x})|^2 \\ &= |1 \circ H(\mathbf{x}) + i\sigma p(\mathbf{x}) \circ H(\mathbf{x})|^2 \end{aligned} \tag{2.14}$$

The First convolution can be easily solved using Fourier convolution theorem:

$$\mathcal{F}(1 \circ H(\mathbf{x})) = \delta(\mathbf{k}) \exp(-i\chi(\mathbf{k}) \Rightarrow \exp(-i\chi(\mathbf{0}) = 1 \tag{2.15}$$

so that the back transform is 1 because the values for all frequencies except 0 are 0.

$$\begin{aligned} I(\mathbf{x}) &= |1 + i\sigma p(\mathbf{x}) \circ H(\mathbf{x})|^2 \\ &= 1 + \sigma p(\mathbf{x}) \circ (iH(\mathbf{x}) - iH^*(\mathbf{x})) \end{aligned} \tag{2.16}$$

[3] See section 2.1.2 on page 19.

where * denotes complex conjugation. If the Fourier transform of H is known the difference can be evaluated in phase space:

$$\begin{aligned}\mathcal{F}(iH(\mathbf{x}) - iH^*(\mathbf{x})) =& i\exp(-i\chi(\mathbf{k}) - i\exp(i\chi(\mathbf{k})) \\ =& 2\sin(\chi(\mathbf{x}))\end{aligned} \tag{2.17}$$

Finally the image function is defined as:

$$I(\mathbf{x}) \approx 1 + 2\sigma p(\mathbf{x}) \circ \mathcal{F}^{-1}(\sin(\chi(\mathbf{x}))) \tag{2.18}$$

The Image in the cryo-EM is more and more recoded by using CCD[4] cameras instead of photographic film. The advantage is the direct output of pixel images that can be processed by computers.
In a first step all segments of the micro graphs are selected where particles are expected. This is a semi-automated process by preselecting the interesting positions by a computer program and than checking those data by a human. Depending on the signal-to-noise ratio in further steps of the data analysis images can be removed from the stack, because of the noise one could not identify if an image contains information or just noise [106][87].

2.1.2 Contrast Transfer Function

The magnetic fields of a magnetic lens is determined from Maxwell's equations, which prevent those from acting like an ideal lens known in optical microscopy. It is a sophisticated field by itself to optimize magnetic lenses by modifying the shape or increasing the amount of poles in the magnetic field. This will introduce the aberrations in the electron microscope, already mentioned by Scherzer in 1949[77]. Aberrations can be modeled in a variety of basis functions, Zernike polynomials, which are based on radial polynomials on the radial deviations and the azimuth are used in adaptive optics, when the more obvious and bottom to the line power series of positional and angular deviations are used in electron microscopy. As the wavelength of the electrons is much smaller than the dimension of the lenses and the specimen the system can be described like optical systems by refraction indices[44].
To reduce the complexity of the system the beam will be assumed to be parallel to the optical axis and all off axis aberrations can be neglected, further all positional deviations can be ignored. A perfectly symmetric lens will ignore the directions of the angular deviations and will simplify the rotational symmetric description of the aberration, where the angular deviations (α_x, α_y) can be described by the radius $\alpha = \sqrt{\alpha_x^2 + \alpha_y^2}$. This reduces the series to the even powers and the phase shift χ described by:

$$\chi = \frac{2\pi}{\lambda}\left(\frac{1}{2}C_1\alpha^2 + \frac{1}{4}C_3\alpha^4 + \frac{1}{6}C_5\alpha^6 + \cdots\right) \tag{2.19}$$

where the coefficients C_i have units length. Analogous to Zernike polynomials for radial symmetric aberrations the power series implies $-C_1 = \Delta f$ as the defocus

[4] charge-coupled device

and $C_3 = C_s$ as the spherical aberration. Higher order terms will be ignored because their entry is too small on the assumed beam configuration.
Scherzer figured out that a static, rotationally symmetric magnetic field like magnetic lenses will always induce a spherical aberration greater than zero, so the second order term will be important and has to be corrected to increase the contrast of the micro graphs[77]. The defocus term can be used to offset the effect of spherical aberration to increase the the bandwidth were low spatial frequencies are transferred with a similar phase.

$$\Delta f = -1.2\sqrt{C_s \lambda} \tag{2.20}$$

Another important factor is that the defocus can be used to shift the sections with no phase informations and as a result a signal for all spatial frequencies will be received. So all images have an error in the first two terms of the aberration and have to be corrected subsequently by the inversion of the phase shift:

$$\chi = \frac{2\pi}{\lambda}\left(\frac{1}{4}C_s\alpha^4 - \frac{1}{2}\Delta f \alpha^2\right) \tag{2.21}$$

The angle α, the angle between the incident ray and its scattered direction, is related to $k = 1/d$ the spatial frequency in the image plane by the wavelength:

$$\alpha = \lambda k \tag{2.22}$$

and the phase shift can be written in terms of k:

$$\chi(k) = \pi\lambda k^2(0.5C_s\lambda^2 k^2 - \Delta f) \tag{2.23}$$

In general an envelope function E has to be applied to the CTF to adjust to finite source size, energy spread, drift effects and other effects etc. so that the final function is:

$$H(\mathbf{k}) = E(\mathbf{k})\exp(i\chi(\mathbf{k})) \tag{2.24}$$

According to the image function (Eq. 2.18) the dominant part is the phase modulation by the sinus and the approximated image function is:

$$I(\mathbf{x}) \approx 1 + 2\sigma p(\mathbf{x}) \circ \mathcal{F}^{-1}(E(\mathbf{k})\sin(\pi\lambda k^2(0.5C_s\lambda^2 k^2 - \Delta f))) \tag{2.25}$$

Further a chromatic aberrations has to be taken into account, which is described in the temporal coherence envelope function E_c:

$$\begin{aligned} E_t &= \exp\left(-\frac{1}{2}(\frac{\pi}{\lambda}\delta k^2)^2\right), \\ \delta &= C_c\sqrt{4\left(\frac{\Delta I}{I}\right)^2 + \left(\frac{\Delta E}{E}\right)^2 + \left(\frac{\Delta V}{V}\right)^2} \end{aligned} \tag{2.26}$$

where C_c is the chromatic aberration coefficient and $\Delta I/I$ fluctuations in the lens current, $\Delta V/V$ fluctuations in the accelerator voltage and $\Delta E/E$ the energy spread of emitted electrons[36][44].

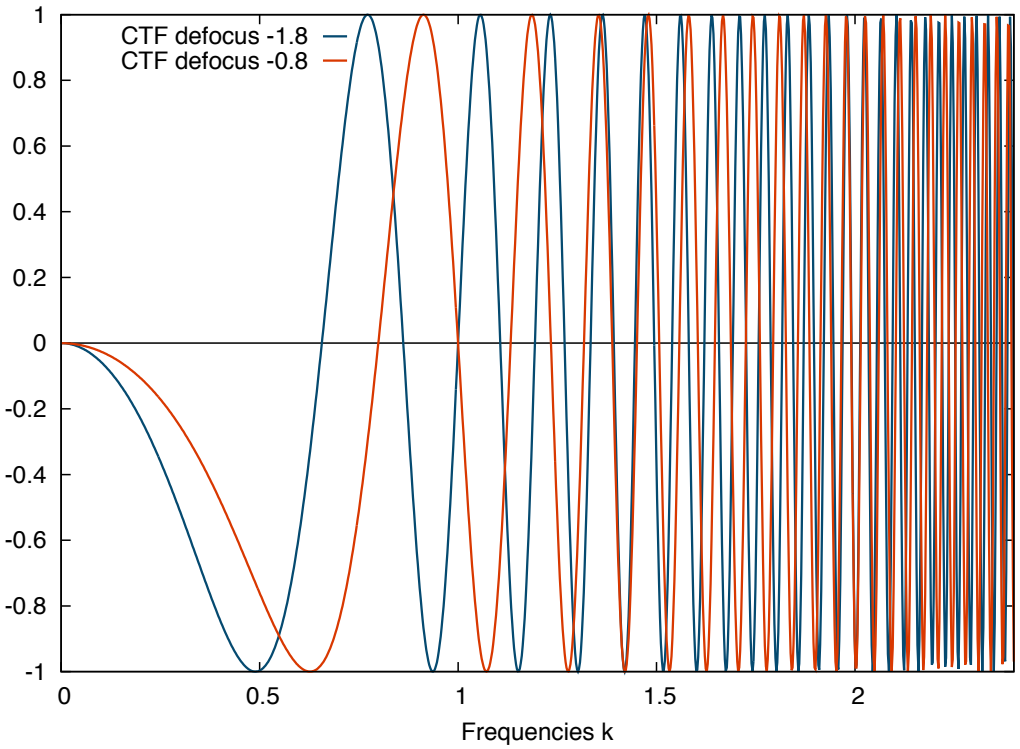

Figure 2.1: Two overlaid CTF functions with different defocus. The electron frequency is assumed to be unit sized $\lambda = 1$, the spherical aberration $C_s = 2.5$ and the defocus $\Delta f_1 = -1.8$(blue) and $\Delta f_2 = -0.8$(orange). The values have been chosen to illustrate the effect of changing the defocus and shifting the zeros. This is used to get information in areas of annihilation. If an entire defocus series is taken zero section will not appear in an averaged image [67].

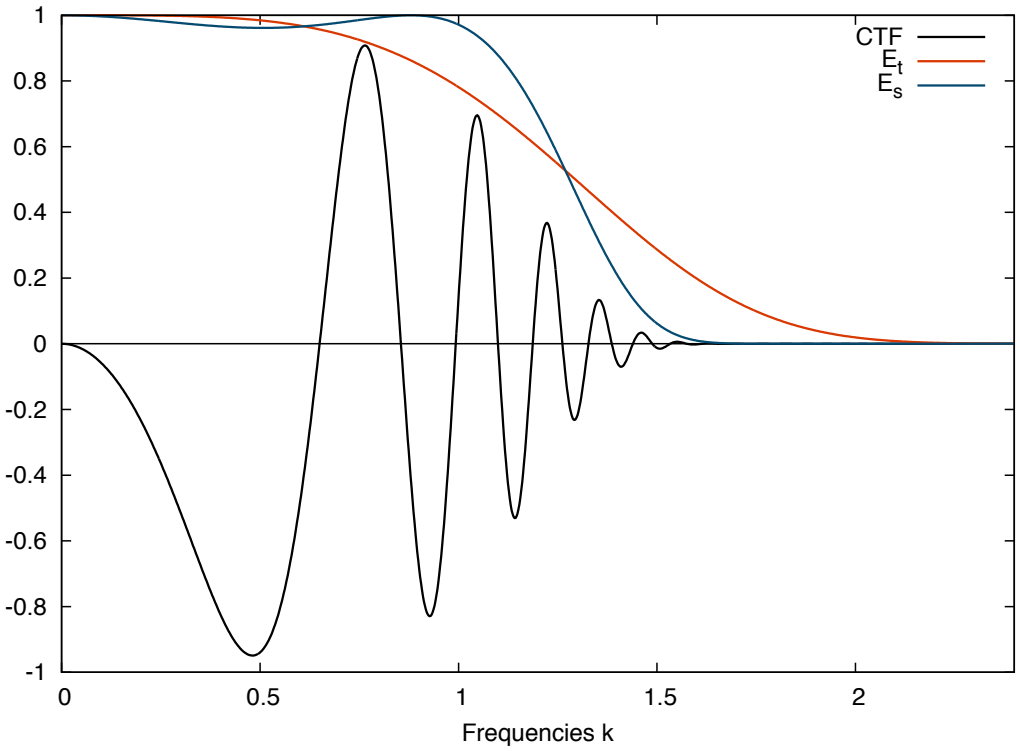

Figure 2.2: The CTF with envelope functions applied to has a large low frequency band, where a signal can be recorded at several defocus series. The effect of the envelope functions is suppression of high frequencies, while low frequencies can pass. This is the major resolution limiting effect in EM which is responsible for thousands times less resolution than expected from calculations (2.5). The function parameter are chosen corresponding to Figure (2.1).

On the other hand there will be spatial coherence based on the defocus which is described in the spatial envelope function E_s:

$$E_s = \exp\left(-(\pi\theta_s)^2(C_s\lambda^2k^3 + \Delta fk)^2\right) \tag{2.27}$$

where θ_s is the beam divergence angle.[26]
To remove the CTF from the images the Wiener filter, a least square error filter, is used:

$$W(\mathbf{k}) = \frac{H^*(\mathbf{k})}{|H(\mathbf{k})|^2 + 1/SNR} \tag{2.28}$$

where H is the contrast transfer function in Fourier space and SNR is an approximated signal to noise function rotationally symmetric and will be multiplied with the Fourier transformed signal. If there are image stacks with different defocus, the final resolution can further be improved and the zeros of the oscillating CTF will be filled with additional informations[95][67].
The importance of this CTF filtering in cryo-EM was shown by Penczek et al. 1997. Later more and more methods have been published applying the CTF correction on the 3D volumes reconstructed from the same defocus to reduce computation time and to get better approximations for the SNR [67].
Today the CTF is used in a less approximated form:

$$\mathrm{CTF}(\lambda, k, \Delta f, C_s) = -w_1 \sin(\chi(\lambda, k, \Delta f, C_s)) - w_2 \cos(\chi(\lambda, k, \Delta f, C_s)) \tag{2.29}$$

with $w_1 = \sqrt{1 - A^2}$ and $w_2 = A$ and A ranges from 0.07 [89] to 0.14 [81]. To simplify the refinement of the CTF Mindell suggests to use corrected power spectrum P_c by a smoothed power spectrum and then maximizing the correlation between P_c and the CTF [57].

(2.30)

2.2 Reconstruction of 3D Volumes

2.2.1 Alignment of Images in 2 Dimensions

First of all the images selected from the micro graphs have to be aligned, which depends on the used methods can be very complicated and it seams to be useful to first understand the alignment in 2 dimensions, this means that all copies of the molecule are in the same view. In cryo-EM the images can be shifted, rotated and isotropically scaled, if they have been taken from different micro graphs. The most important function to compare two images of same size is the correlation function, which will be maximal if two images are the same and zero if there is no shared information. So in general the correlation function will be maximized by an algorithm according to an operation like translation. For the translation τ the correlation integral R of the images p_1 and p_2 is:

$$R_{p_1,p_2}(\tau) = C \int_I p_1(t) \cdot p_2(t - \tau)dt \tag{2.31}$$

where I is the space of the image plane, C is the inverse of the area of an image and $t \in I$ a pixel position. This Function is called cross-correlation in signal processing

and corresponds to the convolution of the images at τ, which has to be maximized. This gives a very simple form for the optimization of Equation 2.31:

$$\max_{\tau \in I} R_{p_1,p_2}(\tau) = \max_{\tau \in I}(p_1 \circ p_2)(\tau) \tag{2.32}$$

Computing the convolution can be accelerated by using Fast Fourier transforms and the circular convolution theorem, which says for two continuous and integrable functions x and their Fourier transforms X, Y:

$$\mathcal{F}^{-1}(X \cdot Y)(\tau) = (x \circ y)(\tau) \tag{2.33}$$

With this transform the convolution can be calculated easily and the maximum can be searched in the convolution function.
Almost the same method can be used to do a rotational alignment. In that case the images are mapped to polar coordinates. Now the image is described buy a vector (r, ϕ), where r is the radius and ϕ is the rotational angle. A rotation around the center with an angle of θ is $(r, \phi + \theta)$ and a scaling of a factor of s is $(s \cdot r, \phi)$, so a trick is needed to get an addition, which can later be solved by the convolution:

$$\begin{aligned} s \cdot r &= \exp(\ln(s) + \ln(r)) \\ \phi + \theta &= \phi + \theta \end{aligned} \tag{2.34}$$

Using this trick the mapping of an image vector (x, y) will be:

$$\begin{aligned} r &= \frac{1}{2}\ln(x^2 + y^2) \\ \phi &= \operatorname{atan2}(y, x) \end{aligned} \tag{2.35}$$

The translation vector is $\tau = (\ln(s), \theta)$ and it will give the same optimization problem as Equation 2.31. This roughly described concept for image registration can be used to align images in the same view but can also further used to cluster images in which are not in the same view. Of course the problem is a little bit more complex and it is difficult to find the maxima so multiple maxima are recorded and the ratio is used to get the right on, because the signal ratio of the significant peak should be above the other ratios which are due to noise[26].

2.2.2 Radon Transform

The basis for all reconstructions is the radon transform and its inversion, which was introduced by Radon 1917[70]. The Radon transform can be defined in two or three dimensions, to simplify matters it will be discussed in two dimensions, so let $f(\mathbf{x}) = f(x, y)$ be a continuous function on a disc $D \subset \mathbb{R}^2$ and vanishing at the border ∂D. The Radon transform Rf is the function of line integrals though a centered plane:

$$\begin{aligned} x(r)_\alpha &= r\cos(\alpha), \\ y(r)_\alpha &= r\sin(\alpha) \end{aligned} \tag{2.36}$$

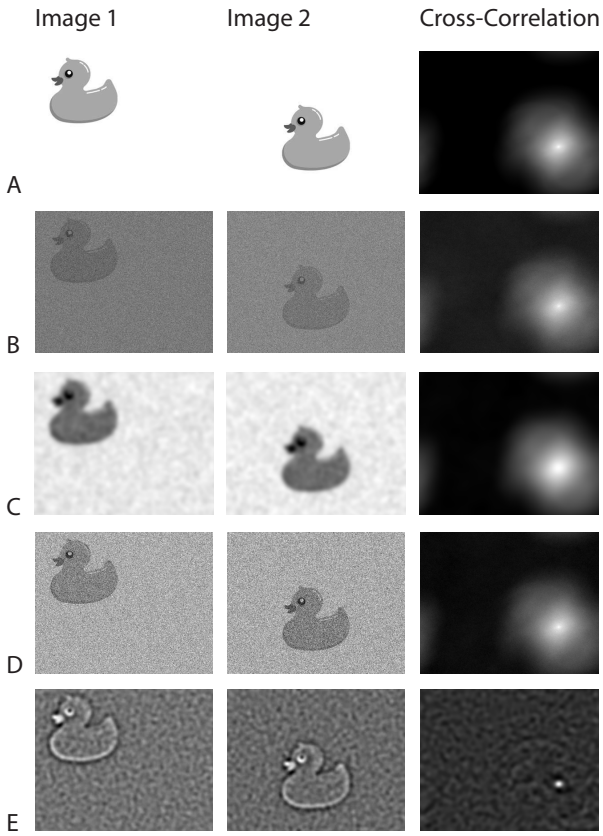

Figure 2.3: The columns show the sketch of a Rubber Duck in image 1. In image 2 a copy of image 1 shifted by a vector. In the last column the cross-correlation of both images is printed, the maximum is shifted from the center of the image by the same vector as image1 has to be shifted to become image 2. A special attribute of Fourier transforms and the cross-correlation are the periodic boundary conditions, that can be seen in the correlation that is reentering at the top and on the left side. Each pixel in the cross-correlation image corresponds to the correlation of image 2 and image 1 shifted by the vector equal to the pixel coordinates. In the first row (A) this is presented for the original image and the cross-correlation looks like a Gaussian, because the correlation for not perfect overlaps is not zero in this simple example with large unicolored areas. In the second row (B) a Gaussian pixel noise of $\sigma = 0.8$, 80% of the maximal density value, is added, which is a fine noise. In the correlation the low correlation values now point out some correlation but the shifting vector or maximum position is as good visible as in case A. In the third row (C) a Gaussian low pass filter is applied to B with a width of $\sigma = 10$ pixel, forming this coarse grained noise. Another effect is that the contours of the duck are smoothed. The cross-correlation looks like the correlation of A. At this Point it is obvious that the correlation is very powerful to compare images independent of noise. In the last two rows effects of filter are shown used for reconstruction form C. In row D a simple deconvolution is show by an inverse gauss filter. This is a very rough approximation of a wiener filter ignoring SNR. A lot of the information lost from B to C has returned in the images, The noise in D is not as fine as in B, because information was lost in the Gaussian filter by a multiplication with zero and those operations can not be inverted. The cross-correlation is not affected by this operation. In row E a Laplacian Filter is applied to C and now the edges of the duck become visible while the unicolored areas are still noisy. This transform has a large impact on the correlation, the maximum area is much better defined, due to image scalings the peak ares seems to be broaden but the maximum is still well defined. This technique is important for low resolution data with high SNR, because the peaks become more dominant.

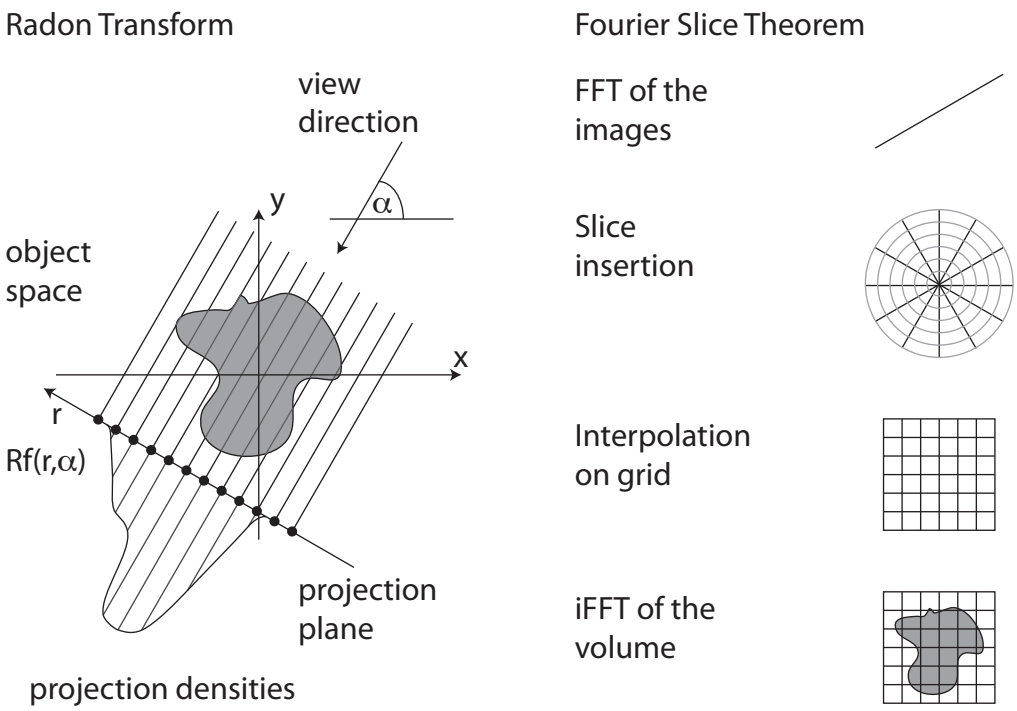

Figure 2.4: On the left side of the Figure is the Radon transform of a projection angle α. The line integrals through the density (grey) $f(\mathbf{x})$ in the object space. The resulting projection $Rf(r, \alpha)$ is in the projection space consisting of planes, which will be oriented by the angle α. On the right side is a diagram of the Fourier slice reconstruction. In the first step the images are Fourier transformed to get the frequency domain representation. In the second step the slices are oriented by α and combined in radial space. From radial space the volume is interpolated onto a grid to apply in the last step the inverse Fourier transform. This graph is a simplified model of the reconstruction and several enhancements have to be added to the process.

with its normals $(-\sin(\alpha), \cos(\alpha))$. The used path function for the integral will be $\gamma_{r,\alpha}(t) = (r\cos(\alpha) - t\sin(\alpha), r\sin(\alpha) + t\cos(\alpha))$, so that the Radon transform equates:

$$\begin{aligned} Rf(r,\alpha) &= \int_{\gamma_{r,\alpha}} f ds = \int_{-\infty}^{\infty} f(\gamma_{r,\alpha}(t)) \, \|\dot{\gamma}_{r,\alpha}(t)\|_2 \, dt \\ &= \int_{-\infty}^{\infty} f\left(\begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} r \\ t \end{pmatrix}\right) dt \end{aligned} \tag{2.37}$$

Hence the Radon transform equates to an orthogonal projection on the t-axis, it describes what happens in the transmission electron microscope (Eq. 2.8). The projection of the volume is determined by a rotation matrix $\mathbf{A}_\alpha$.
For a reconstruction the inverse Radon transform Rf^{-1} is needed, which can be constructed as the dual transform:

$$R^* g(\mathbf{x}) = \frac{1}{2\pi} \int_0^{2\pi} g(\alpha, \mathbf{n}_\alpha \cdot \mathbf{x}) d\alpha \tag{2.38}$$

where $\mathbf{n}_\alpha$ is the normal vector of the plane with angle α [37].
Solving the inverse transform is not easy but it will help to take a look at the Fourier transform of the volume function, that should be reconstructed (Eq. 2.8) in two dimensions $f(x, y) = f(\mathbf{x})$:

$$\mathcal{F}(f)(\mathbf{k}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{x})\ \exp\left(-2\pi i(\mathbf{x}\cdot\mathbf{k})\right) dxdy \tag{2.39}$$

where $\mathbf{k} = (k_x, k_y)^T$ is the reciprocal vector. If a slice through zero is selected from the Fourier transform $\mathcal{F}(f)(k_x, 0)$ for a slice orthogonal to the k_y-Axis or in a more general way $\mathcal{F}(f)(\mathbf{k}_\alpha)$:

$$\mathbf{k}_\alpha = \begin{pmatrix} k_x \\ k_y \end{pmatrix}_\alpha = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} k_r \\ 0 \end{pmatrix} = \mathbf{A}_\alpha \mathbf{k}_r \tag{2.40}$$

The slice is then:

$$\mathcal{F}(f)(\mathbf{k}_\alpha) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{x})\ \exp\left(-2\pi i\ \mathbf{x}\cdot(\mathbf{A}_\alpha\mathbf{k}_r)\right)\ dxdy \tag{2.41}$$

If the same rotation is applied to the volume function $\mathbf{A}_\alpha : \mathbf{r} \to \mathbf{x}$ and $\mathbf{r} = (r, t)^T$, the function will be:

$$\mathcal{F}(P)(\mathbf{k}_\alpha) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{A}_\alpha\mathbf{r})\ \exp\left(-2\pi i\ \mathbf{r}^T\mathbf{A}_\alpha^T\mathbf{A}_\alpha\mathbf{k}_r\right)\ drdt \tag{2.42}$$

Solving this Equation and using Equation 2.37 will give the projection slice theorem:

$$\begin{aligned}\mathcal{F}(f)(\mathbf{k}_\alpha) = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(\mathbf{A}_\alpha\mathbf{r})dt\right)\ \exp\left(-2\pi i\ (r\cdot k_r)\right)\ dr \\ = \mathcal{F}\left(Rf\right)(k_r, \alpha)\end{aligned} \tag{2.43}$$

Due to invariance of the integral to rotations of the coordinate system or in a mathematical way, because the Jacobian determinant of rotations is 1, the same proof will work for higher dimensions than two.

2.2.3 Back-Projection and Fourier Reconstruction

The easiest way to reconstruct 3D volumes from 2D images is to use the filtered back-projection, which tries to overlap the 2D informations in space and interpolates a volume[29][55][87]. This can be very complicated if it is done in real space; so often it is done in Fourier space by taking advantage of the projection slice theorem. This method has not less problems in the reconstruction, but most of them can be handled more easily.
Fourier space methods are inverting the projection slice theorem (Eq. 2.43) in a way that if one slice of the spatial frequency domain is a projection of the volume in reciprocal space, multiple slices can be added to invert the projection. The amount of information on inner shells around the center is larger than outer shells, it is obvious that the distance of points carrying information on different radii from the center is proportional to the radius. This means that the information

decreases by $1/k = 1/|\mathbf{k}|$ and the weighting has to be the inverse. This is an informal derivation of the contrast transfer function (CTF) of the back-projection. For real space methods this CTF has to be applied to the data at one point in the algorithm[42].
Mathematically this process can be described by a simple formula:

$$\mathcal{F}(V) = k \cdot \int_I \mathcal{F}(p) \tag{2.44}$$

where V is the Volume, $p \in I$ is an image and I is the space of images. Until this point everything seams to be easy in Fourier space but there are disadvantages to Fourier space reconstruction especially because the data sets are discrete and finite in space.
So called Phantoms appear after reconstructions at the borders of a volume box because discrete Fourier transforms assume a periodic signal outside the defined space. If there are undefined frequencies or especially if frequencies are cutoff, these phantoms will appear. To avoid these effects widow-functions are applied to the transform which try to fill missing frequencies[35]. On the other hand it is common to zero pad the images, this means adding areas with zero values on all sides which will be cutoff after back transform and phantoms will mostly appear in these areas and will have no effect on the volume [15].
Most of these algorithms are iterative today and using weighting techniques, which weight the amount of information in an image with its fit to the already reconstructed volume [87] [29] [38].
Real space reconstruction become today more and more important because of general-purpose computing on graphics processing units (GPGPU) which is aided by NVIDIA in science in the last six years. Nowadays a lot of groups started developing for these platforms. The problem of Fourier transforms is that they are not well scalable on massive parallel machines.

2.2.4 Detecting Projection Angles

The missing part is the angle refinement because the reconstruction can only be done if the orientation is known, on the other hand the angles can only be estimated if there is an idea of the three dimensional shape. For this reason the orientations are optimized during the reconstruction of the 3D density [29][87]. But the question is - what to start with? There are different answers some would say any starting model is acceptable others would say we can try to get a prototype by using pre-clustered images, only a subset of images or other fancy tricks.
The basic idea in the alignment process is to realign the images to the volume by generating a template volume and than realigning the images against the projections of this volume. This is connected to back propagation algorithms, which describe a learning model to optimize an error term. Many alignment procedures use angle classes in the alignment, starting with a small number of classes and increasing the number during the alignment. The advantage is that projections of the 3D volume have to be calculated for each class and not individually for each image. In this procedures it is helpful to have the images clustered by correlation, to have a better initial situation [87][2].

To create an initial volume several techniques are used, the most simple one is starting from a random volume. Any random asymmetric (or with the same symmetry as the specimen) model can be used to start with.
Another way is to use the common lines of the Fourier transformed images and optimize their fit. The idea is that Euclidean planes with at least on common point share at least one line or the entire area. This is the case for the centered slices described in the projection slice theorem. If the orientation of the planes vary, the angles can be estimated [65]. For small numbers of images this algorithm can be very fast and will give an initial model, that can be used for the angle alignment. Previously averaged image classes can be used as well as random chosen images from the entire set [29][53].

3 Chapter
Estimating the Conformational Variance of the Specimen

3.1 Estimating Errors of the Reconstruction

3.1.1 Statistical Values

The problem of estimating the variances in the reconstruction starting at the moment of freezing the specimen, the imaging in the electron microscope, till the density volume will be reconstructed, is not trivial. If all steps can be modeled by linear functions. Methods for calculating the variances has been presented by Liu and Frank [51][50] and Haley [31], which basically describe the same technique. The idea is to use the difference between the reconstructed model and the 2D projections. This was done by calculating the projections of the reconstruction and using the absolute value of the differences for reconstructing a variance map. In 2006 Penczek has shown a real space method for calculating the variance using bootstrapping (see 3.2.2)[64].
Starting from weak-phase approximation (Eq. 2.8) a discrete model can be assumed like:

$$\mathbf{p} = \mathbf{P}\mathbf{f} \tag{3.1}$$

where $\mathbf{f}$ is a vector containing a density grid of n voxels[1], $\mathbf{p}$ the m pixel image and $\mathbf{P}$ the $m \times n$ projection matrix. If we further assume, that the inverse transform $\mathbf{P}^{\dagger}$ exists with a smoothing function $\mathbf{S}$, the back transform can be written:

$$\hat{\mathbf{f}} = \mathbf{S}\mathbf{P}^{\dagger}\mathbf{p}. \tag{3.2}$$

In this case $\hat{\mathbf{f}}$ is the estimator for the reconstruction that will converge to the density $\mathbf{f}$. The variance of the volume is defined by

$$\sigma_{\mathbf{f}}^2 = \langle \mathbf{f}^2 \rangle - \langle \mathbf{f} \rangle^2. \tag{3.3}$$

The analog definition of the covariance can be simplified:

$$\mathbf{C}_{\mathbf{f}} = \langle (\mathbf{f} - \langle \mathbf{f} \rangle)(\mathbf{f} - \langle \mathbf{f} \rangle)^T \rangle = \langle \mathbf{f}\mathbf{f}^T \rangle - \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^T. \tag{3.4}$$

[1] voxel are volumetric pixels

Due to the fact, that the expected value is defined as a Lebesgue integral hence linear, it is invariant of the back projection matrix $\mathbf{P}^\dagger$ and the covariance matrix $\mathbf{C_f}$ can be written as a function of the covariance of the images $\mathbf{p}$:

$$\mathbf{C_f} = (\mathbf{SP}^\dagger)\mathbf{C_p}(\mathbf{SP}^\dagger)^T. \tag{3.5}$$

The estimator for $\hat{\mathbf{f}}$ of $\mathbf{f}$ can be calculated by solving the least square problem:

$$\hat{\mathbf{f}} = \mathbf{S}\left((\mathbf{P}^\dagger)^T(\mathbf{P}^\dagger)\right)^{-1}(\mathbf{P}^\dagger)^T\mathbf{p} = \mathbf{Rp} \tag{3.6}$$

where the pseudo-inverse of the reconstruction matrix is used. Now the covariance of the estimator $\hat{\mathbf{f}}$ is:

$$\mathbf{C}_{\hat{\mathbf{f}}} = \mathbf{RC_pR}^T. \tag{3.7}$$

Starting from this stochastic model it is possible to estimate further statistical values and to investigate the reconstruction process.

3.1.2 Sources of Variance

In the entire process (cf. Chapter 2) six major sources of variance can be constituted[64]:

1. *specimen* can be described by three different reasons for noise, (a) pre-experimental, for example impurities in the sample, and (b) during the experiment, damage by radiation and in particular conformational changes;

2. *medium surrounding* the proteins can cause irregularities in the amorphous ice and impurities can occur;

3. *specimen support film* used to stabilize the protein and can effect orientations, i.g. carbon;

4. *microscope* can induce a thermal drift, variances in the electron beam, electrostatic charges and scattering events of other particles can appear.

5. *data collection* will impact the image's graininess on the film and during digitalization or on CCD cameras, that is the reason additional noise, that depends on the dynamics of the image.

6. *image processing* can be split into three segments, (a) misalignment during the process of shifting and scaling the images, (b) reconstruction errors because of a non continuous space or rather missing information and (c) interpolation errors because of changing the grids.

At this point it seems to be important to discuss some aspects of the noise in detail. The noise of the particles itself is a reason for variance in the dataset, but on the other hand it is an additional information about the protein, that can be interpreted in terms of vibration or "harmonic" oscillations. This splits the noise into a background noise and a protein intrinsic component [60].
Another point seams to be important to be mentioned here (mathematical) convolutions introduce correlations, this has an effect on noise, that is affected by the CTF, which corresponds to a convolution of the images and on any interpolation,

which can be described as an convolution too. Further the pixel correlation will increase during the reconstruction, because the correlation is maximized to align the images.
It seams to be useful to describe the noise in three components [64]:

1. *Solvent variance* σ^2_{solv} can be easily estimated by just selecting a non protein area from the micrograph and calculating its variance. The variance contains irregularities and impureness of the ice. Both are effected by the CTF so the first estimation is not perfectly correct. Further the influence of a thin support film can be estimated by this variance. All together this can be used to estimate the noise uniformity, and the noise not affected by the specimen.

2. *Variance of volume* σ^2_{vol} is the variance of the reconstructed electron density, which will be affected by the image processing algorithms, nonuniform distributions of projections, conformational variance and a background noise.

3. *Variance of structures* σ^2_{struct} describes the small conformational changes of each particle in the specimen.

The first time Liu and Frank[51] mentioned two different types of noise based on intrinsics of the protein the variance of structures σ^2_{struct} and the solvent variance σ^2_{solv}.

3.2 Estimating the Conformational Variance

3.2.1 A stochastic model

Now we can set up a model to describe the noise of a reconstruction σ^2_{vol}. As described in the last Chapter the variance can be described in several ways and all should fit together to estimate the final variance:

$$\sigma^2_{vol} = \sigma^2_{Conf} + \sigma^2_{Ali} + \sigma^2_{Rec} + \sigma^2_{Back} \tag{3.8}$$

where σ^2_{Conf} is the conformational variance in the reconstruction, σ^2_{Ali} the error of the alignment and the variance of the projection distribution, σ^2_{Rec} the error of the reconstruction and σ^2_{Back} estimating the background noise. This Equation ignores correlation and especially the CTF, to keep this model simple.
Because of the reconstruction process and its averaging character it is easy to see that the conformational variance in the specimen has to be larger than the variance of the $\sigma^2_{struct} \geq \sigma^2_{vol}$ - equality is only possible if all molecules are equal [64].
In this case we will ignore the estimation of the variances except the conformational variance σ^2_{struct}.

3.2.2 Bootstrapping

The Method is used to identify variations within cryo-EM samples of 2D image data. The basic problem in single-Particle cryo-EM is that all images used for reconstruction of 3D volumes are taken from different particles and are assumed to be in almost the same conformation. If just a single reconstruction is done all

this slightly different information is condensed in one data set and the additional information is lost. To overcome this problem several resampling techniques have been developed i.e. Bootstrapping[22], which is a general purpose computer-based method for assigning measures of accuracy to sample estimates. The Particles of cryo-EM experiments can be assumed to be from an independent and identically distributed population and the entire set estimates the distribution. Random sampling with replacement can be used to obtain new set of equal size of the observed data set and be used for further reconstructions [64]. In a first step a variance map can be calculated from the 3D reconstruction of those bootstrapped data sets (Fig. 1). The Bootstrapping technique is already implemented in EMAN2 [87] and can be directly used in the reconstruction process of cryo-EM data.
It is important to note, that the ensemble of density maps does not represent single conformations of the protein but instead just represent the distribution of density values. Each of these maps is still an average of different conformations of the protein, but the distribution of conformations is conserved in the newly generated ensemble. The conformational variance is conserved in those bootstrapped maps and overlaid by several other sources for noise, like impurities, amorphous ice, microscope, data collection, image processing etc.
Penczek presents methods to calculate the structural variances from these volumes and shows how to eliminate the noise of the reconstruction and the background - containing all sources of noise in the micrograph - in one step by applying the bootstrapping technique to sections of empty space within the micrograph and estimating the noise distribution from averaging those noise bootstrapped maps [64]. In another step the variance of the alignment process can be estimated and separated from the conformational variance, which is not very stable [3].
The entire bootstrapping process will result in densities, describing only the structural variances of the specimen in a particle mesh representation.
The bootstrapped reconstruction is closely connected to the normal reconstruction:

1. do zero padding to avoid phantoms;
2. apply FFT;
3. loop over n volumes:
   - select with replacement m images from the set of all images;
   - use projection slice theorem to insert the image to the volume $\mathbf{f}_i$;
   - do filtering and inverse FFT of the volume.

From the bootstrapping it is possible to write the estimator for the electron density as:

$$\hat{\mathbf{f}} = \overline{\mathbf{f}} = \frac{1}{n}\sum_{i=0}^{n}\mathbf{f}_i = \frac{1}{n}\sum_{i=0}^{n}\mathbf{R}\mathbf{g}_i, \tag{3.9}$$

where $\mathbf{g}_i$ is the image set corresponding to the i-th volume.

3.2.3 Calculating the Conformational Variance

According to Equation 3.9 a variance can be calculated from the bootstrapping data:

$$\sigma^2_{\hat{\mathbf{f}}} = \frac{1}{n-1}\sum_{i=0}^{n}(\mathbf{f}_i - \bar{\mathbf{f}})^2, \tag{3.10}$$

It is obvious that the variance of the bootstrapped map $\sigma^2_{\hat{\mathbf{f}}}$ is equal to the variance of the volume σ^2_{vol} (cf. 3.8). Further Hansen et al.[33] has shown that for large numbers of projections the variance of the bootstrapping is linearly connected to the structural variance σ^2_{vol} by a linear factor of n:

$$\sigma^2_{\hat{\mathbf{f}}} = \sigma^2_{vol} = m\sigma^2_{struct} \tag{3.11}$$

Based on Equation 3.11 and 3.8 the conformational variance can be estimated as:

$$\sigma^2_{struct} = m(\sigma^2_{\hat{\mathbf{f}}} - \sigma^2_{Ali} - \sigma^2_{Rec} - \sigma^2_{Back}) \tag{3.12}$$

If we neglect the alignment error σ^2_{Ali}, the reconstruction error σ^2_{Rec} and the background noise σ^2_{Back} the conformational variance is very simple to resolve, but we have to assume that it is still overlaid by this noise.

3.2.4 Principal Component Analysis

To characterize the variance of multivariate ensembles a common mathematical procedure is the principal component analysis (PCA). The PCA transforms a set of possibly correlated observations into a set of linearly independent variables. Each of this new variables is called principal component. The basic idea is the algebraic concept of eigenvalues and eigenvectors applied on a covariance matrix (cf. 4.1).
By the PCA we can obtain vectors of density which can be used as an approximative basis for the phase space of the specimen. Large distinct conformational changes can be encoded in these eigenvectors. Due to this fact a PCA is often used to get a basis for clustering the dataset. In cryo-EM the problem is, that the data set is not part of the same space as the PCA, so it is used as a so called 'Codimensional PCA' by Penzcek and Spahn [82, 63]. They use the reconstructions to cluster the images by correlation into new groups describing well defined different conformations of the protein. Another effect is the reduction of noise because on the one hand the variety of distinct overlapping conformations in the images can be reduced, on the other hand it will help to avoid mixing up the different conformations into one unnatural mixed state. In this context it seems to be helpful to call the eigenvectors eigenvolumes.
The Process consists of 7 steps:

1. image alignment and reconstruction of a single density;
2. bootstrapping reconstruction to create an ensemble;
3. PCA to yield eigenvolumes as an basis for the volumes;
4. 2D projection of the eigenvolume basis to receive a basis in the image space;

5. determination of the factorial coordinates of the images on this basis;
6. clustering images by their factorial coordinates;
7. reconstruction of each cluster as a single volume.

This process will recover several details of the conformational space and can be improved by a modified sampling technique called hyper-geometric stratified re-sampling (HGSR)[63].
In this case the effect off non conformational noise in the bootstrapping variance can be neglected, because the correlation will help to limit the effect, but will not avoid over-fitting by to many iterations

Chapter 4
Implementation of an accurate Sparse PCA

4.1 Principal Component Analysis

The principal component analysis (PCA) is a singular value decomposition. The PCA is a multivariate statistical method, used to analyze large sets of data and to simplify it by a minor variety of statistical variables. The amount of parameters is reduced to linear uncorrelated "principal components". Even when the PCA was published by Karl Pearson in 1901 it gained importance with the availability of computer because of its algebraic complexity.
The underlying multivariate observed set is typically written as random vector $\mathbf{X} \in \mathbb{R}^n$, where n is the dimensionality of the observations. Mathematically the PCA is an eigenvalue problem, to find an orthogonal basis of the space or subspace of the data set. The resolution is a simple algebraic problem, that can be solved by several decompositions. The more interesting part of the PCA is the statistical interpretation:

1. the input matrix for the PCA is the covariance matrix,
2. the resulting eigenvectors form an orthogonal basis and
3. eigenvalues are the variance of uncorrelated components.

Of course the first interpretation is easy to understand, if a uncorrelated basis is needed, this means that all covariances are zero - not the variances on the diagonal of the covariance matrix, which is the intention of the eigenvalue problem. So in the first step the covariance matrix has to be computed:

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma} = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & \mathrm{Cov}(X_n, X_n) \end{pmatrix} \tag{4.1}$$

It is important to keep in mind that the covariance of the same random vector is the variance $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$. If the data set is uncorrelated, the covariance matrix would only have values on the diagonal:

$$\mathrm{Cov}(\mathbf{X}) = \mathrm{diag}(\mathrm{Var}(X_1), \ldots, \mathrm{Var}(X_n)). \tag{4.2}$$

If the covariance matrix is not diagonal it is at least a symmetrical matrix and positive semidefinite, because of the variances on the diagonal and the matrix is diagonalizable. In the second step the diagonalizing of the matrix gives a new orthogonal basis with no covariances and only variances. So finally the eigenvalues correspond to the variances on the corresponding eigenvectors. This is what the Karhunen-Loève theorem states for the PCA or more general for Fredholm integral Equation of the second kind [43, 52].

4.1.1 Degrees of Freedom

A very interesting point about covariance matrices is what happens if the matrix is rank deficient. We can think of basically two different cases the covariance matrix is rank deficient:

1. the data set is correlated in one or more dimensions;
2. the data set is not describing the entire phase space.

The first option can generally not be predicted at all and is in fact what the PCA is used for. In physics this would be the question, if a particle could move freely in space is limited to a (hyper) plane, which would eliminate at least one degree of freedom. This is similar to the definition in statistics; basically it is the number of dimension of the domain of a random vector or in simple words the number of elements till the vector is determinate.
The second problem is more interesting, because in this case only a subspace is described by the observation and the rank of the covariance matrix is only less or equal to the number of observations. So if the matrix of observables $\mathbf{X} = (\mathbf{X}_0 - \bar{\mathbf{X}}, \ldots, \mathbf{X}_m - \bar{\mathbf{X}},) \in \mathbb{R}^{n \times m}$, where $\mathbf{X}_i \in \mathbb{R}^n$ is the i-th observation vector, is used to build the covariance matrix by:

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T \tag{4.3}$$

its rank is described by a fundamental formula of linear algebra:

$$\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B)). \tag{4.4}$$

So if $m \ll n$ the rank of the covariance matrix is rank deficient with rank less equal the number of observations, so that only a subspace will be described by the covariance and several eigenvalues will be zero. In this case the amount of observation limits drastically the dimensionality of the problem.
This is typically the case for bootstrapping in cryo-EM, where the size of a map is always larger than $50 \times 50 \times 50 = 125000$ but only about 100 volumes generated. In this case a lot of dimensions of the phase space are not determinated, on the other hand we can assume that no reconstruction is in a plane of two other volumes, due to the noise in the entire system. All in all the dimension of the matrix is determined by the number of observations m.
Here the reduced degree of freedom by estimating the mean value will be ignored.

4.2 The Accurate Sparse PCA

4.2.1 Motivation

As described in the last Section the covariance matrix is of deficient rank, which will result in mostly zero eigenvalues. To diagonalize such a matrix in general more basic matrix operations have to be used for the diagonalization in the QR algorithm or any other algorithm and a lot more single calculation have to be done, which will affect the computational time. Another even bigger problem can be the size of the covariance matrix, which exponentially increases.
This technique can be applied to molecular dynamics simulations (MD) too, if one is interested in vibrations or a linear decomposition. So it seems to be useful to implement the algorithm for atomistic structures too. Later we will use this algorithm to analyze atomistic ensembles obtained from cryo-EM data.
If so much information of the covariance matrix is not well defined, there has to by a subspace in which the problem could be solved by less dimensions. Inspired by the concepts of the quasi inverse and the Gramian matrix there should be a way to calculate the eigenvectors by using the $\mathrm{Cov}(\mathbf{X}^T)$. This concept is easy to understand using the spectral theorem or with the following proof, which is focusing on the way the algorithm can be implemented.

4.2.2 Proof

Consider a stochastic process, that generates m points $\mathbf{X}_i$ in an Euclidean n-dimensional space, then the PCA is defined as a orthogonal linear decomposition of such a space, which tries to maximize the variance along its basis vectors. The dimension of this Euclidean subspace is $(m-1)$, if $m \leq n$, otherwise n, where m is the number of generated points. The average $\bar{\mathbf{X}}$ is the translation of the center which gives the new vectors $\mathbf{X}'_i = \mathbf{X}_i - \bar{\mathbf{X}}$.
Only if $m \ll n$, the covariance matrix will be rang deficient. So we will focus on that case. In general the next step is extending the subspace to a orthonormal basis and to transform the vectors, but due to the fact that the transform is linear invariant and not scaling invariant, the distances will be the same as in original space, so that the covariance matrix can be computed directly as a matrix product of:

$$\mathbf{X}' = (\mathbf{X}'_1, \ \ldots, \ \mathbf{X}'_m) \in \mathbb{R}^{n \times m} \tag{4.5}$$

And the covariance matrix is as follows:

$$\mathrm{Cov}(\mathbf{X}') = \mathbf{\Sigma} = \mathbf{X}'^T \mathbf{X}' \in \mathbb{R}^{m \times m} \tag{4.6}$$

To solve the eigenvector problem now a matrix of size $m \times m$ has to be stored and diagonalized instead of a matrix of size $n \times n$. The eigenvector problem can be written as a linear Equation:

$$\mathbf{\Sigma} \cdot \mathbf{V} = \lambda \cdot \mathbf{V} \tag{4.7}$$

where $\mathbf{V} = (\mathbf{v}_1, \ \ldots, \ \mathbf{v}_m) \in \mathbb{R}^{m \times m}$ is the matrix of eigenvectors and $\lambda = (\lambda_1, \ \ldots, \ \lambda_m)^T \in \mathbb{R}^m$ is the vector of corresponding eigenvalues. For the back-

transform to the phase space the calculation of a basis can be omitted by using the input data set as follows:

$$\begin{aligned} \mathbf{\Sigma} \cdot \mathbf{v}_i &= \lambda_i \cdot \mathbf{v}_i \\ \mathbf{X}' \cdot \mathbf{\Sigma} \cdot \mathbf{v}_i &= \lambda_i \mathbf{X}' \mathbf{v}_i \end{aligned} \tag{4.8}$$

with expansion of $\mathbf{\Sigma} = \mathbf{X}'^T \mathbf{X}'$:

$$\begin{aligned} \mathbf{X}' \mathbf{X}'^T \mathbf{X}' \mathbf{v}_i &= \lambda_i \cdot \mathbf{X}' \mathbf{v}_i \\ \mathrm{Cov}(\mathbf{X}) \cdot (\mathbf{X}' \cdot \mathbf{v}_i) &= \lambda_i (\mathbf{X}' \cdot \mathbf{v}_i) \end{aligned} \tag{4.9}$$

Now $\mathbf{X}'\mathbf{v}_i$ solves the eigenvector Equation for the covariance matrix in the phase space. In general it is necessary to normalize these eigenvectors to get a orthonormal basis for the PCA. Of course is the number of non zero eigenvectors less than m, because the number of varying values is less equal $(m-1)$, due to the fact that the mean is a varying value consuming an additional degree of freedom. To use this to minimize the size of the covariance matrix it would be necessary to compute the new basis and the transformation, which would increase the computation time to a greater extent than reasonable. This way of solving the PCA already needs more computation time for full rank matrices but with increasing nullity this effect will invert. The advantage of this calculation is, that it can decrease the size of the covariance matrix to be diagonalized from n^2 to m^2. We will discuss later, what this implies for the data size on some examples.

4.3 Comparison

The traditional and the new algorithm have been implemented using the LAPACK routine '?SYEV' for eigenvalue calculations via QR-decomposition, all other calculations are implemented in C. The entire program runs serial and no threading or other parallel techniques are used, even though the code can be executed in parallel too, which could improve the performance further.
All Tests have been executed on a desktop PC with Intel Core2Quad 2.66GHz CPU and 4 GB of memory and a Linux[1] system.
Several tests have been setup with random data to calculate average running times to examine the runtime behavior of both algorithms. Therefor a memory block was filled with random data and duplicated for each of the algorithms to take running times. This was done one hundred times to get a reasonable average for the computing times. In Figure (4.1) the execution time of both methods are plotted; in the first Graph (upper left) you can see that the traditional algorithm is more or less independent of the ensemble size especially for small system sizes. On the other hand the new method is independent of the system size and the running time is almost proportional to the ensemble size (Fig. 4.1,upper right).
The speed up is not that big (Fig. 4.1,bottom left) and only for systems with an ensemble size to system size ratio of less than 0.1 significant, what can be very well estimated (Fig. 4.1,bottom right) with the previous runtime function. The function is very rigorous an will in many cases select the normal PCA routine. If

[1] Debian Linux amd64

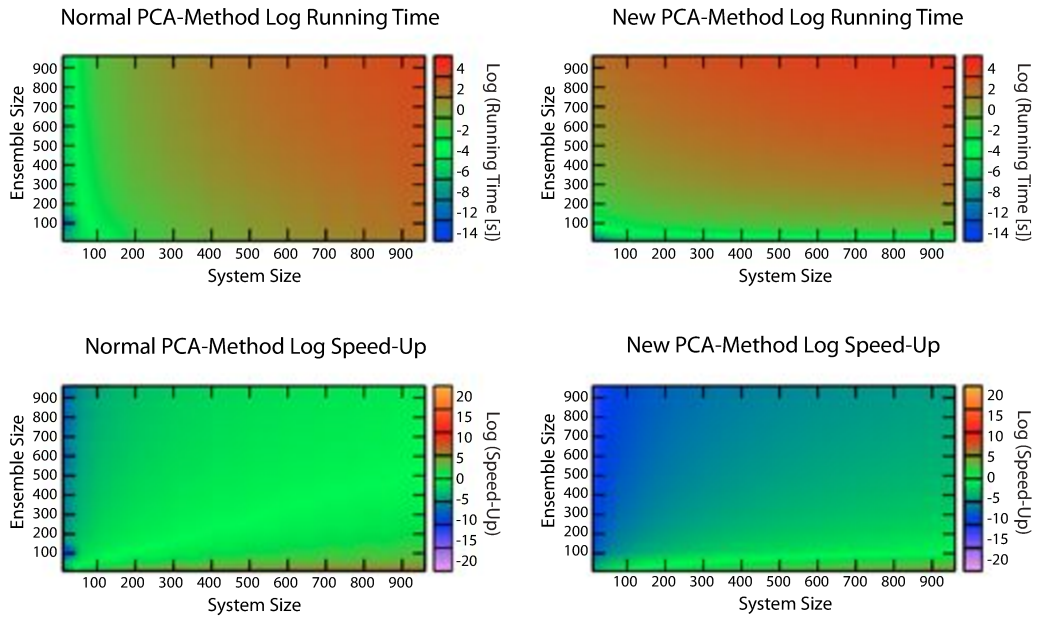

Figure 4.1: In the upper row the running times are plotted per system size on the x-axis and number of frames or ensemble size on the y-axis. The times are averaged over 100 test runs with random input sets. On the left side the traditional PCA algorithm is used and the best running times are obtained for small systems. This means that the algorithm is pretty much independent of the ensemble size while heavily dependent on the system size. The new algorithm on the right side is dependent on the ensemble size and not on the system size. In the lower row the logarithm of the speedup is plotted, on the left side is the measured speedup of the methods on the right side the theoretical speedup. The theoretical predicted advantages of one algorithm compared to the other is less important in the intermediate areas where both methods show the same performance. This can have two reasons either the eigenvector calculation is mores dominant or memory access becomes more important and limits the calculations.

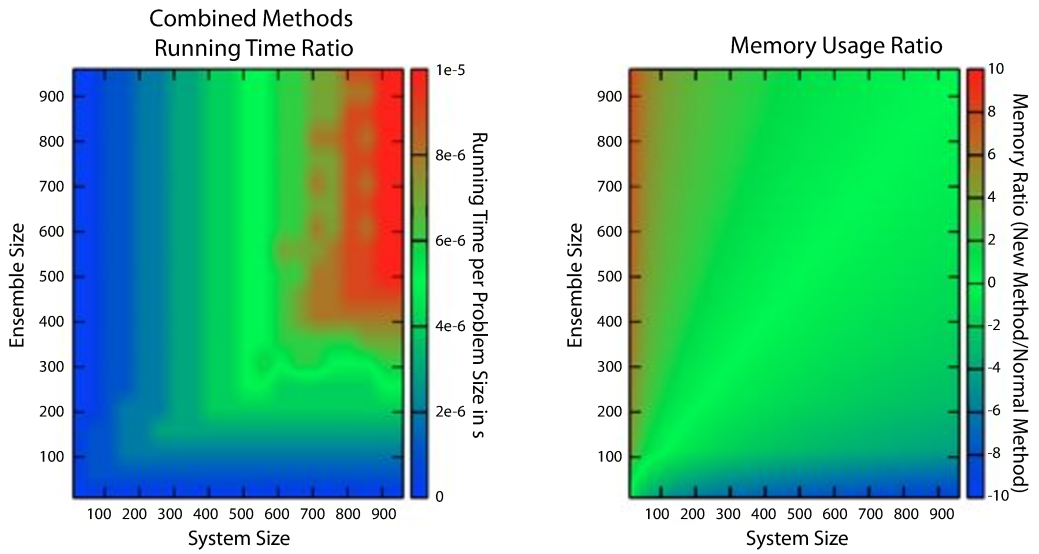

Figure 4.2: On the left a combination of both methods running time per problem size (system size time ensemble size) is plotted. This combined algorithm performs best on any tested problem. The running times along both axes are almost constant. The newly introduced method enables fast PCA calculation independent of the system size, which scale similar to the normal PCA, that is independent of the ensemble size. On the right hand side the memory ratio is plotted. The new method enables low memory usage PCAs for large system sizes as the traditional PCA performs on large ensembles.

the benefits of both PCA methods are combined and always the fastest is used, there will be a runtime like Figure (4.2 left). Only large ensembles of large systems will need a lot of computation time; large ensembles of small systems are fast to solve this the normal PCA and small ensemble of large systems can be solved by our method. This is highly correlated to the inverse memory consumption (Fig. 4.2 left), what means the fastest way to calculate the PCA is always the one with the lowest memory usage. The memory usage is proportional to the ensemble size divided by the system size.
As an example for a large biomolecule we took a 100 frames trajectory of the Ribosome with thousands of residues and more than 150,000 atoms. To calculate a PCA on the Ribosome it is necessary to store at least 42 MB for the input data and 9 kB for the covariance matrix. The QR-Algorithm used roundabout the double size of the covariance matrix as work memory. All in all the most memory is used by the input data but its is still so small that no in-place Algorithms have to be used. On the reference system it took 1 minute and 25 seconds to read in the data, align the structures, calculate the PCA and write all results to disk. The default algorithm would use more than 370 GB for the covariance matrix and it seems useless to run the program on a desktop PC or a server without swapping. The projections of the Ribosome trajectory on first and second eigenvectors describes a path through the used subspace, which can not be described as a Gaussian noise (Image).

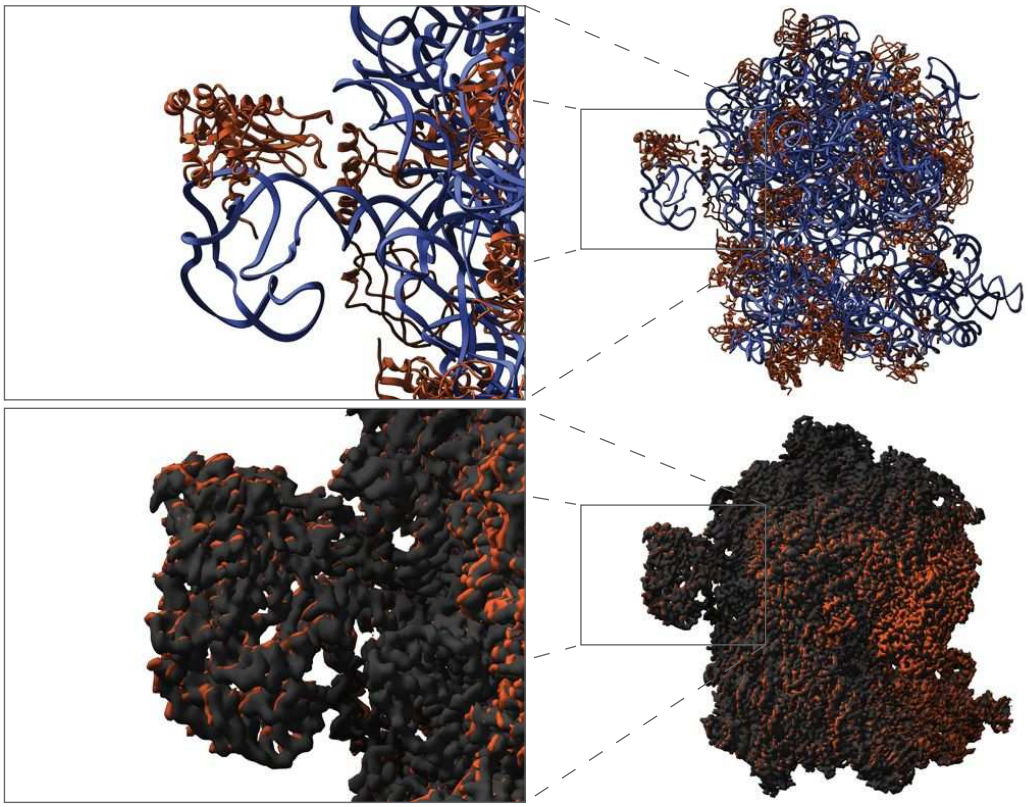

Figure 4.3: In the top row of the Figure is the structure of the ribosome drawn with proteins in orange color and nucleic acid in blue on the left side is a side domain zoomed in. The entire complex consists of 150341 atoms, which where used in the PCA. In the lower row the volume of the Ribosome is shown on a $348 \times 348 \times 348$ grid with a resolution of 1.3Å/pixel. The gray average map is overlayed with the largest eigenvector in orange, in the zoom on the left side the amount of information is presented, when comparing the detailed information of the density with the structure, more details than just secondary structure is presented.

As an example for a PCA on volumetric data the Ribosome data set is used again. The structures have been converted into 100 density maps with a symmetric box size of 348 voxel and a grid spacing of 1.3 A with a size of 160MB and more than 15 GB in total. This should not be calculated on a desktop PC anymore but can be computed on a recent server with 16GB free memory. The size of the work memory can be neglected, because it needs less than 1 MB. Again only our new algorithm is used to calculate the eigenvolumes of the Ribosome in 40 minutes.

5

Chapter

Atomistic Refinement

5.1 Basic Idea

Experiments on biomolecules often only yield low- resolution or sparse structural data for example cryo-EM. Because in contrast to many X-ray scattering experiments it is not possible to directly reconstruct the atomic structure from the densities reconstructed from an EM experiment. In a lot of cases this "missing" data can be provided by prior knowledge; in a further step this information has to be combined with the experimental information to achieve an atomistic structure. Today two different approaches are used to do the refinement of cryo-EM densities:

- force field driven refinements using the experimental information as an additional term;
- rigid body refinement estimating best placement of non-flexible domains.

On the other hand force field driven methods do not need a good high resolution model and can be used in example with homology models. The force field is used to restrain the bonds, bond angles and other elements of the local geometry. This approach needs a lot of computational power to refine a structure towards the experimental data. The differences in the conformations of the starting structure and the target have a big impact on the computation, because only small steps can be done by those MD simulations [41].
The rigid body fitting uses an existing structure, that is decomposed into its domains. The domains will be fitted into the density to get an estimation for the new structure. A positive effect of this techniques is that the secondary structure of protein is conserved and can be taken from high resolution data. A drawback is the missing of inner domain changes and the often unnatural bonds in the splitting zones of the decomposition [96].
Both methods have their pros and cons but can also be combined in a flexible refinement method, which is implemented i.g. in DireX.

5.1.1 MD Simulation

In MD simulations the phase space of a protein can be explored, based on an energy function, which grants a particular realistic environment. By the creation of a conformational space corresponding to a protein, the special conformation is search describing the density measured in the experiment better than any other. The MD simulations used in such a refinement are usually based on a hybrid energy function combining traditional MD force fields with and additional force on the target. The traditional force field is defined as a functions of all atoms $\mathbf{X} = (\mathbf{X}_0, \ldots, \mathbf{X}_N)$ with N the number of atoms:

$$\begin{aligned} V_{MD}(\mathbf{X}) = & \sum_{bonds} \frac{1}{2} c_b (d - d_0)^2 \\ & + \sum_{angles} \frac{1}{2} c_a (\theta - \theta_0)^2 \\ & + \sum_{torsions} \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)] \\ & + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \epsilon_{i,j} \left[\left(\frac{d_{0ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{d_{0ij}}{d_{ij}} \right)^{6} \right] \\ & + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \frac{q_i q_j}{4\pi\epsilon_0 d_{ij}}, \end{aligned} \tag{5.1}$$

where d is the distance of two atoms, θ the angle of the bindings regarding to the orbitals and ω the twisting angle of a bond due to other bonds. A lower 0 is indicating the reference value for that component[14]. The sums are representing:

1. the potential of the bonds modeled as a spring;
2. the orbital model of an atom with the deviation affecting a harmonic potential on the angle;
3. torsions of the bond model, which will be be expressed in a Fourier series;
4. the van der Waals forces approximated by Lennard-Jones potential;
5. the electro static potential.

This Potential will be modified by an extra term for the difference to the target T of the refinement, so that the potential becomes:

$$V_{refine}(\mathbf{X}) = V_{MD}(\mathbf{X}) + V_{diff}(\mathbf{X}, \mathbf{T}) \tag{5.2}$$

In cryo-EM the target is the electron density (cf. 2.1.1) so that the difference potential can be approximated by a harmonic potential:

$$V_{diff}(\mathbf{X}, \mathbf{T}) = c(f_{\mathbf{X}} - \mathbf{T})^2 \tag{5.3}$$

where $f_{\mathbf{X}}$ is the coulomb potential off the structure $\mathbf{X}$. This is just to get an idea of this kind of MD simulations and not very accurate, therefore you will have to expect several problems using this potential.
With such a modified force field it would be possible to do a refinement by MD.

5.1.2 Rigid Body Fitting

This is another completely different approach using an initial high resolution structure. The Idea is quite simple, just try to place the structure in the density in the way it fits best. The advantage is that over-fitting is no big problem because you will not change the structure itself. To get more precise descriptions the structure is often decomposed into domains which will be fitted. This domains are mostly larger than the resolution so the impact of over-fitting can still be neglected. A problem can arise with the decomposition of the molecule, because if it is split in the wrong position it is not flexible enough or the segments do not fit correctly to the density.
The method is always an optimization of a dimension describing the similarity of two densities, the target $\mathbf{T}$ and the structure $\mathbf{X}$. Due to the fact that the structure has a perfect resolution, its resolution is lowered by the convolution with a Gaussian $\mathbf{G}$ and the rasterizing to a grid. In most cases the correlation is used to compare the densities of the target an the structure:

$$\max \sum_i \mathbf{T}_i \cdot (\mathbf{G} \circ \mathbf{X})_i \tag{5.4}$$

an alternative method was presented by W. Wriggers, that uses the Lapalcian correlation and can be described as an contour fitting, which has a higher contrast [11]:

$$\max \sum_i \nabla^2 \mathbf{T}_i \cdot \nabla^2 (\mathbf{G} \circ \mathbf{X})_i \tag{5.5}$$

This method is very successful at resolution above 10 Å but lack on precision at higher resolutions. Depending on the domain size the data has a finer definition of the structure.

5.2 Approximation of a Forcefield

5.2.1 Forces of the Density Map

Both approaches are unsatisfying, one could be very accurate but slower than the other does not create reasonable structures and is not very accessible for small changes but fast. This brings in a third approach, which is inspired by elements of both methods and uses a simulation strategy. To use this approach, forces has to be derived from the two volumes, therefor the notion for the MD simulation refinement (cf. 5.3) is used, where a force can be calculated from the difference of those maps.
The forces introduced by the target depend on the correlation of the two densities; this forces are updated after each structure update. Due to the normalized character of the correlation function a pseudo energy could be defined as:

$$E_{corr} = \mathrm{Corr}(\mathbf{T}, (\mathbf{G} \circ \mathbf{X}) = \frac{\sum_i (\mathbf{T}_i \cdot (\mathbf{G} \circ \mathbf{X})_i)}{\sqrt{(\sum_i \mathbf{T}_i \cdot \mathbf{T}_i)^2 \cdot (\sum_i (\mathbf{G} \circ \mathbf{X})_i \cdot (\mathbf{G} \circ \mathbf{X})_i)^2}} \tag{5.6}$$

This energy should be minimized by the simulation, which is not very efficient and a stochastic approach is chosen to optimize the fit. Therefor random positions $\mathbf{r}_i$

are taken around the atom position $\mathbf{x}$ from a radial Gaussian distribution and the directional vectors are summed and weighted by the density difference. This is a stochastic differential operator, that should be more robust to noise than a same sized partial derivative operator:

$$\nabla f(\mathbf{x}) \sim \frac{1}{n} \sum_{i=1}^{n} \left((f(\mathbf{x}) - \lambda f(\mathbf{r}_i)) \frac{(\mathbf{r}_i - \mathbf{x})}{|\mathbf{r}_i - \mathbf{x}|} \right) \tag{5.7}$$

where λ is a scaling factor and should be chosen as 0.6. For this gradient the densities should be scaled to reasonable values or normalized to the normal distribution. This method gives a robust force towards the target map and is more efficient than solving the full partial derivative [78].

5.2.2 Sampling the Phase Space

The other component of a simulation based approach is the MD-forcefield, which is overloaded for the refinement of cryo-EM data. 1997 de Groot presented a method to predict conformational freedom from distance constraints [16]. This method uses an initial structure to generate a network of distance restrains as a constraint for probable conformations. These restraints include topological restrains to keep the correct stereo-chemistry and restraints to avoid overlapping of atoms [78].
In contrast to an MD simulation this algorithm perturbs the atom positions by a Gaussian. Then the atom positions are randomly changed, till all bond conditions are preserved. This is done iteratively by a random traversal through the list of restrains and moving those pairs along their common axis a bit towards its restrained distance.
By this it is possible to explore the allowed phase space very quickly and to obtain realistic structures. Using this sampling based method instead of a real potential for the refinement improves the speed of generating new structures.
Replacing the dynamics based exploration of the phase space by this sampling based method speeds up the entire process and can be used together with other forces. For example the force derived from the volumes can be applied to the atoms at any iteration to converge the sampling.

5.2.3 Deformable Elastic Network

At this point the refinement can be done very quickly, the problem is that during the refinement the simulation will not converge. This is caused by the low information of local structure features in the cryo-EM data and is still visible at 3Å densities. Another problem is over-fitting, because the structure is still to flexible on local areas can be refined to noise in the density and the structure will be distorted, this could even happen to secondary structure elements before a good fit is reached [78].
Due to the Problems in the refinement an additional deformable elastic network(DEN) is used to stabilize local structure elements and to reduce over-fitting and to improve the rate of convergence. The DEN potential is designed to stabilize the local structure but not to avid flexibility, so it is implemented to be time dependent. This dependency on time dependency is modeled into the concept of a harmonic

restraint. Due to the fact that no realistic forces are computed the time step is a virtual time step and can be better described as a refinement iteration step. The resulting energy term for the network at a recent iteration step n can be written as:

$$E_{DEN}(n) = k \sum_{pairs\ i,j} \left(d_{ij}(n) - d_{ij}^{(0)}(n) \right)^2, \tag{5.8}$$

where $d_{ij}(n)$ is the distance of the restraint pair i and j and $d_{ij}^{(0)}(n)$ is the corresponding equilibrium distance. The force constant k is used as a constant scaling factor while the equilibrium distance depends on the step number. The interesting part of the network is the deformation which allows the network to follow slowly the structure and still resists random fast vibrations. The update of the network is done after each refinement step and defined by recursion:

$$\begin{aligned} d_{ij}^{(0)}(n+1) = d_{ij}^{(0)}(n) \\ + \kappa \cdot \gamma \left(d_{ij}(n) - d_{ij}^{(0)}(n) \right) \\ + \kappa \cdot (1-\gamma) \left(d_{ij}^{(0)}(n) - d_{ij}^{(0)}(0) \right), \end{aligned} \tag{5.9}$$

where κ determines the speed of adapting the new position and γ is the balance between the adaptation of the new state and preserving of the initial state. This model can be enhanced by using another reference than the initial structure and written in a simpler way:

$$\begin{aligned} d_{ij}^{(0)}(n+1) = (1-\kappa) d_{ij}^{(0)}(n) \\ + \kappa \left(\gamma d_{ij}(n) + (1-\gamma) d_{ij}^{ref} \right), \end{aligned} \tag{5.10}$$

where d_{ij}^{ref} can be any reference; for $d_{ij}^{ref} = d_{ij}^{0}(0)$ this Equation is equivalent to (5.9) [80].

With this additional network it is possible to refine structures, the process converges against an optimal value for the correlation, which corresponds in general to a good RMSD. A deeper view on the refinement and its parameters will be done in the next Chapter.

All together the DEN can be described as a harmonic elastic network with the equilibrium state coupled to another weighted harmonic potential which enables the deformability of the DEN. Putting this together with the conformational sampling and the stochastic gradient calculation, it is possible to set up a fast and robust refinement system for resolutions from 16Å down to 3Å, which includes the typically obtained resolutions in cryo-EM.

Chapter 6

Validation of the Refinement Process

6.1 Cross-validation

6.1.1 Choice of the Test set for Cyro-EM Data

For cross-validation the data set needs to be splited into two independent parts. For this it is convenient to represent the data set by structure factors, which are the Fourier components of the density map, since each of these components contain global information on the entire system. However, several factors lead to correlations between these structure factors in cryo-EM derived density maps: cryo-EM images are usually taken at defocus to improve the image contrast. The corresponding contrast transfer function, which describes the spatial frequency dependency of the transmitted signal, depends on this defocus and is an oscillating function that contains multiple zero crossings. In Fourier space the CTF, thus, imposes correlations on structure factors between neighboring Fourier shells. In addition, the alignment of the images during the density reconstruction procedure introduces further correlations of the noise in these images [84]. In cryo-EM the structure factors are therefore too strongly correlated such that a random choice of the structure factors for the test set, as is done in crystallography, is not optimal. Furthermore, the signal-to-noise ratio (SNR) for cryo-EM density maps decreases for higher spatial frequencies. To visualize this, the Fourier shell correlation (FSC) [34] can be computed which is a measure of the signal-to-noise ratio in the individual Fourier shells and is shown in Fig. (6.1) for three model systems described below at two different resolutions of 5 Å and 10 Å. The reconstructed density maps are usually filtered to remove the noise originating from the higher spatial frequency range, i.e., information from this range is often neglected in the interpretation of the density. However, the signal in this high-frequency band might still be strong enough to be useful for validation, as is shown below.
We therefore propose to define as test set for the cross-validation a continuous band (the 'free band') from this high-frequency region. The wider this band, the less crosstalk occurs between structure factors within and outside the band and the less correlated is the free band with the work band. More specifically we choose the

free band in the range where the FSC is between 0.2 and 0.6, which includes the point where FSC = 0.5, which is a common definition for the resolution of a density map ($FSC_{0.5}$ criterion). The red bars at the top of Fig.1 indicate the regions that are used here for the free set: for the 10 Å data we use the range 7−11 Å and for the 5 Å data, we use the range 4−6 Å. The choice of resolution shells for the selection of the test set has been described before for X-ray crystallography [24][45][1] to reduce correlations between the test and work set in the case of high non-crystallographic symmetry. It should be noted that the common application of additional filters, such as Gaussian low-pass filters, introduces additional correlations and should therefore not be used when preparing a density map for the refinement of atomic models.

6.1.2 Implementation

The approach has been implemented into the program DireX. DireX performs real-space refinement of atomic models against density maps using an efficient geometry-based conformational sampling algorithm [17][79]. It optimizes the overlap of a density map computed from the model with the target (experimental) density map. For the cross-validated refinement, we compute the model density map using only Fourier components from the work band and also filter the target density map with a rectangular filter as defined by the work band.
During the refinement of the atomic coordinates, restraints are applied to maintain local stereo-chemistry and prevent atom overlaps. In addition, DireX uses deformable elastic network (DEN) restraints [80][79] to account for the low observation-to-parameter ratio at low resolution. These harmonic restraints are defined between randomly chosen atom pairs that are within a distance range of typically 3 to 15 Å. The deformability is achieved by allowing the equilibrium distances to change, which effectively moves the minimum of the network potential. This minimum adapts itself to balance the influence of the density map and a set of reference coordinates, which in the cases presented here are equal to the coordinates of the starting models. The strength of these restraints relative to other forces is determined by the weight factor w_{DEN} and the deformability of the network is controlled by the parameter γ, where $\gamma = 0$ means no deformability and $\gamma = 1$ means maximum deformability, i.e. no information about the reference model is used. These two parameters, γ and w_{DEN}, need to be optimized and it is demonstrated here how this can be done using cross-validation. Other refinement programs that use different types of restraints will need to optimize different parameters, which we expect to be possible analogously with the cross-validation approach presented here.

6.1.3 Measure of Fit

The traditional measure of the fit of a model to diffraction data is the R-value:

$$R = \frac{\sum_{h,k,l} ||F_{obs}(h,k,l)| - |F_{calc}(h,k,l)||}{\sum_{h,k,l} |F_{obs}(h,k,l)|} \tag{6.1}$$

which compares the amplitudes of the structure factors as this is the most accurate information obtained by crystallography while the phase information is either

missing or usually more inaccurate. The free R-value is then defined by summing over structure factors from the test set T:

$$R_{free} = \frac{\sum_{h,k,l \in T} ||F_{obs}(h,k,l)| - |F_{calc}(h,k,l)||}{\sum_{h,k,l} |F_{obs}(h,k,l)|} \tag{6.2}$$

We denote the free R-value as R^{rnd}_{free} when the structure factors from the test set T are selected randomly, and R^{int}_{free} when they are selected from an interval.
Electron microscopy measures both amplitudes and phases, with usually even higher phase than amplitude accuracy. In this case, a more natural choice for the measure of fit is the correlation of the density map computed from the model, ρ_{calc}, with the experimental density map, ρ_{obs}. The map correlation includes the phases and amplitudes and is scale independent. Here we consider two different correlations: 1) the free map correlation, C_{free}, where only structure factors from the free band were used to compute both density maps:

$$C_{free} = \frac{\sum_{i,j,k} \left(\left(\rho^{free}_{calc}(i,j,k) - \bar{\rho}^{free}_{calc} \right) \left(\rho^{free}_{obs}(i,j,k) - \bar{\rho}^{free}_{obs} \right) \right)}{\sqrt{\sum_{i,j,k} \left(\rho^{free}_{calc}(i,j,k) - \bar{\rho}^{free}_{calc} \right)^2} \sqrt{\sum_{i,j,k} \left(\rho^{free}_{obs}(i,j,k) - \bar{\rho}^{free}_{obs} \right)^2}} \tag{6.3}$$

and 2) the work map correlation, C_{work}, which is analogously defined for ρ^{work} for which only structure factors from the work band were used. It should be noted that the absolute values of $Cfree$ and C_{work} cannot be compared directly as they are computed on different frequency ranges, unlike R^{rnd}_{free} and R^{rnd}_{work}, which are drawn from the same distribution of R-values. For higher spatial frequencies smaller changes in the atomic coordinates lead to larger changes of the correlation; map correlations computed from maps with higher frequency components are therefore more sensitive to structural differences.

6.2 Testing the Method

6.2.1 Tests with simulated data

We tested the approach on three different proteins with simulated cryo-EM density maps at 5 and 10 Å resolution. The starting models are homology models taken from the benchmark set of Topf et al. [88], where we chose an easy (1ake, single-domain), an intermediate (1ikn, two-domain), and a hard case (1hrd, two-domain). The sequence identity of 1ake, 1ikn, and 1hrd is 46%, 46%, and 28%, respectively, and the corresponding initial root-mean square deviation (RMSD) of the starting from the target structure is 3.6 Å, 7.7 Å, and 6.0 Å, respectively. An overview of all cases is given in Table (6.2.1).
To receive realistic cryo-EM density maps we first generated 1 Å density maps from the atomic target structures. These high-resolution maps were then used to compute 900 projection images with the project3d command of EMAN [54]. Gaussian noise was added to these images where the standard deviation was chosen such as to yield a resolution of the final reconstruction of 5 and 10 Å, respectively. The images were split into three equally sized groups. A contrast transfer function (CTF) and an envelope function were applied to the images in each of

		opt. restraints			no restraints		
PDB ID	Initial RMSD	RMSD Å	C_{Work}	C_{Free}	$RMSD$ Å	C_{Work}	C_{Free}
		5 Å					
1ake	3.60	1.40	87.4	21.1	1.86	88.0	18.9
1ikn	7.73	1.80	82.2	35.5	7.94	85.3	17.5
1hrd	5.96	3.88	83.2	23.5	3.99	85.0	16.7
		10 Å					
1ake	3.60	1.47	90.7	52.4	2.67	91.1	42.5
1ikn	7.73	2.14	85.9	44.1	8.30	87.1	35.4
1hrd	5.96	4.30	85.4	36.5	4.95	87.4	31.6

Table 6.1: Summary of refinement results for three models with synthetic density maps. The three models were taken from the homology model benchmark set of Topf et al. [88] and represent an easy (1ake), an intermediate (1ikn), and a hard case (1hrd), in terms of structural similarity between starting and target model. Refinements were done with and without DEN restraints for two resolutions, 5 and 10 Å. Results for the optimum restraints correspond to the DEN parameters that lead to the highest free density map correlation, C_{free}. The root mean square deviation (RMSD) of the refined to the target structure is always lower when using optimal restraints. The work map correlation, C_{work}, is always higher without restraints, compared to using optimum restraints, since the density map is closer fitted by the model. However, without restraints, the RMSD of the refined to the target structure is always higher, indicating that the density is over-fitted. C_{free} is always higher when refining with optimum restraints compared to refinements without restraints and higher C_{free} values always correspond to better structures with lower RMSD.

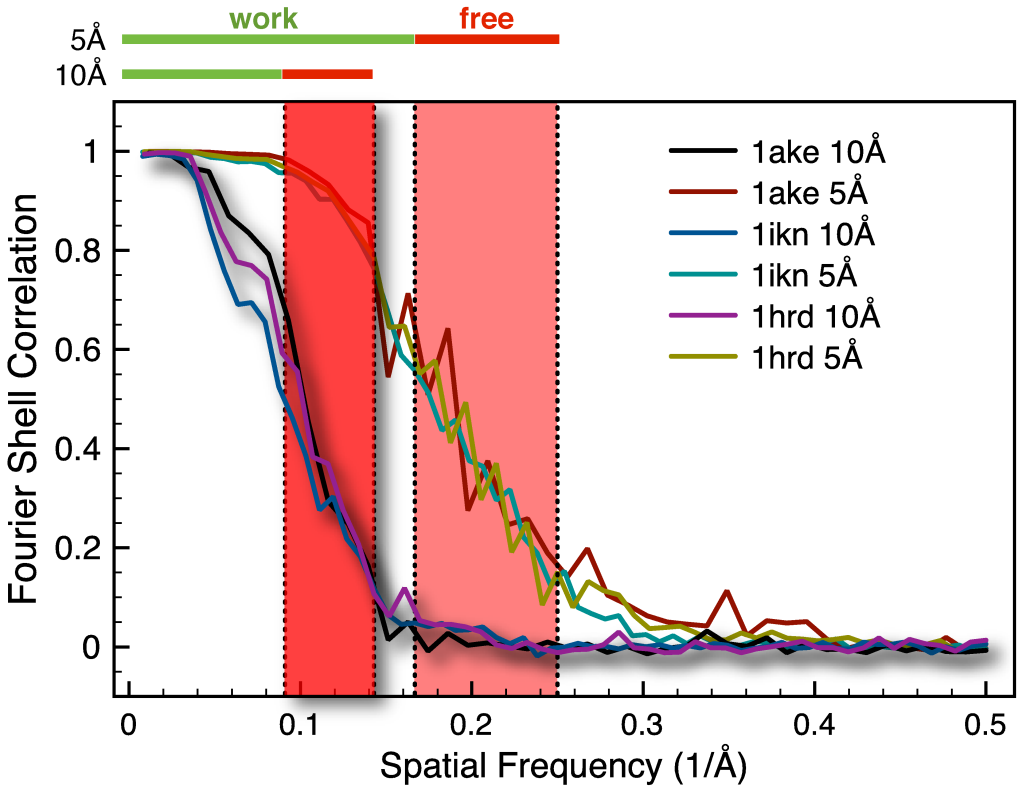

Figure 6.1: Fourier shell correlation curves of simulated cryo-EM density maps are shown for all three models (1ake, 1ikn, 1hrd) at resolutions of 5 and 10 Å. The resolution is defined here as the spatial frequency where the FSC is equal to 0.5. Each reconstruction was generated from 900 projection images. Gaussian noise was added to the images to adjust the resolution of the final reconstruction to be 5 or 10 Å. Red (green) bars on top indicate the spatial frequency range that is used as the free (work) band. The signal-to-noise ratio in the free band is significantly reduced but is strong enough to be useful for model validation.

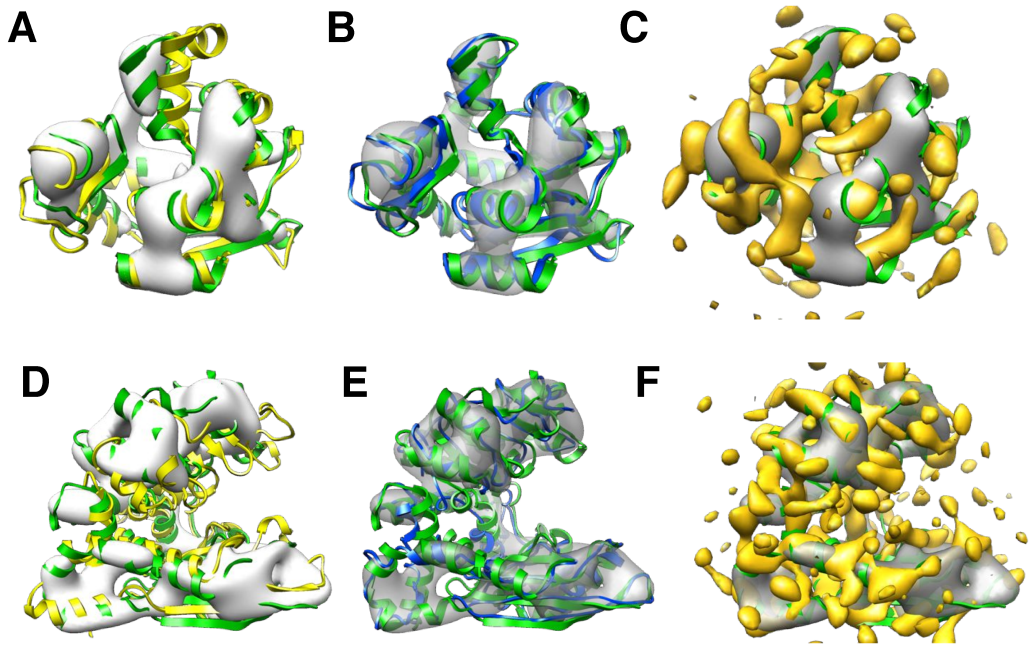

Figure 6.2: Test cases with simulated data. The density maps were simulated based on known high-resolution crystal structures, while the refinements were started from homology models. Two of the three proteins that were used for testing are shown together with density maps at a resolution of 10 Å: The easiest case, 1ake (A-C), has high sequence identity (46%) and a low initial root-mean square deviation (3.6Å). The most difficult case, 1hrd (D-F), has a low sequence identity (28%) and a high initial root-mean square deviation of 6.0 Å. The target, starting and refined models are colored green, yellow, and blue, respectively. The density maps that are used for refining the atomic models are computed using frequency components from the work band and are shown in gray. The free density maps (yellow) are used for validation only and contain frequency components from the free band (7-11 Å).

these groups corresponding to a defocus of 1.3, 2.0, and 2.6 μm, respectively. The three-dimensional density reconstruction was performed with CTF correction. No additional filters were applied to the final reconstructed density maps.
To simulate realistic density maps, we computed these 900 projection images from the target structures with the program EMAN [54] and applied noise to them (see Methods). The noise level was chosen such as to obtain a resolution of 5 or 10 Å ($FSC_{0.5}$ criterion) for the final reconstructed density. The images were split randomly into three groups, in each group a contrast transfer function (CTF) was applied to the images corresponding to a defocus of 1.3, 2.0, and 2.6 μm, respectively. This was done to simulate the commonly used collection of a defocus series [66]. These images were then used to reconstruct a density map with EMAN. Figure (6.1) shows the FSC curves for all six cases (three models/maps at resolutions of 5 Å and 10 Å).
The initial placement of the starting models was done using the rigid-body fit feature of the program Chimera. The number of DEN restraints was chosen as two times the number of atoms. DEN restraints were selected between randomly

chosen atoms that are within a distance range of 3 to 10 Å in the starting model. The refinement was done in 200 steps for 1ake and in 400 steps for 1hrd and 1ikn. The computer runtime needed was relatively short with, e.g., 6 min for 200 steps of 1ake at 10 Å, and 17 min for 400 steps of 1hrd at 5 Å.
Map correlations, C_{free} and C_{work}, were calculated for each DEN parameter combination and averaged over the last 5 structures from the refinement trajectories for ten independent refinement runs started with different random number seeds. Figure (6.2) shows two of these three cases, 1ake (Fig. 6.2, A-C) and 1hrd (Fig. 6.2, D-F). The starting homology model (yellow), the target structure (green), and the refined model (blue) are superimposed on the work density map, ρ_{obs}^{work}, (gray) corresponding to the 10 Å data sets. Figures (6.2) C and F show in addition the free density map,ρ_{obs}^{free} (orange), computed with the spatial frequency components in the range of $7 - 11$ Å. The free maps show little resemblance with the protein structures, as they are composed of only a narrow band of high frequency components and, in addition, these components contain a significant level of noise (cf. Fig. 6.1). However, the signal in this free map is sufficient to be useful for validation as is shown below.
For each case we performed 300 refinements in total with 5 different w_{DEN} - and 6 different γ-parameters in the ranges $0.0 - 0.4$ and $0.0 - 1.0$, respectively. For each of these 30 DEN parameters combination 10 independent refinement runs were performed with different random number seeds. For the first case (1ake) at 10 Å, contour plots (see Fig. 6.3) show the dependency of the root-mean square deviation (RMSD) of the refined structure to the target structure, C_{free}, R_{free}^{int}, and C_{work} values on the w_{DEN} and γ parameters. The best structure, which corresponds to the lowest RMSD value of 1.45 Å, is obtained for $w_{DEN} = 0.1$ and $\gamma = 0.2$ (cf. Fig. 6.3, A). This parameter combination yields the third highest C_{free} value. Whereas, the highest C_{free} is obtained for $w_{DEN} = 0.2$ and $\gamma = 0.6$, which in turn yields a structure that has an RMSD of 1.46 Å to the correct structure which is very similar to the RMSD of the best structure (1.45 Å). This means picking the best C_{free} yields a model that is very close to the best solution.
High γ-values and low w_{DEN} values correspond to weak restraints and lead to over-fitted structures and hence to a large RMSD. The work map correlation (C_{work}) is highest for these over-fitted high RMSD structures, indicating that C_{work} is not a good measure of the quality of the structure. In contrast, the contour plots of the RMSD- and C_{free}-values have a very similar shape, in particular the largest free correlation is found in the same region where the RMSD is lowest. The corresponding contour plots for the third case (1hrd) at 10 Å resolution are shown in Figure (6.4) . While for the easy case (1ake) many different choices of DEN parameters yield low RMSD values, for this difficult case the optimal DEN parameters are confined to a small region. This region of low RMSD values clearly overlaps with high C_{free} values. The corresponding contour plots for all other cases are shown in Figure (6.6, 1-4).
The correlation between C_{free} and RMSD is very strong for all systems we studied with -0.90 averaged over all six cases, suggesting that C_{free} is in fact a good measure to detect the optimum structure. The R_{free}^{int}-value shows a good agreement with the RMSD as well; the correlation between R_{free}^{int} and RMSD averaged over

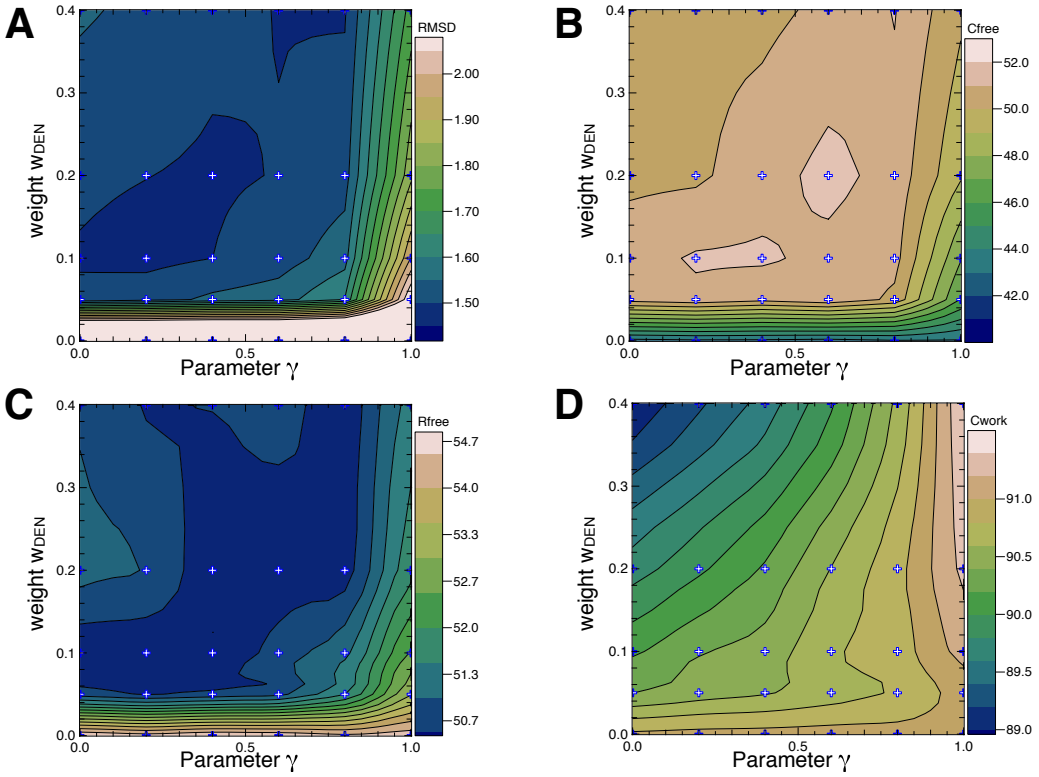

Figure 6.3: Refinement results for the easiest case 1ake at 10 Å. Contour plots showing (A) the root-mean square deviation (RMSD) between refined model and correct crystal structure, (B) the free correlation, C_{free}, the free R-value, R^{int}_{free}, and the correlation of the work maps, C_{work}, as a function of the strength, w_{DEN}, and the deformability, γ, of the elastic network restraints. The highest C_{free} values fall into the same region of parameters w_{DEN} and γ, for which the RMSD is lowest. The C_{work} value instead increases constantly for weaker restraints (smaller w_{DEN} values) and higher deformability (larger γ-values).

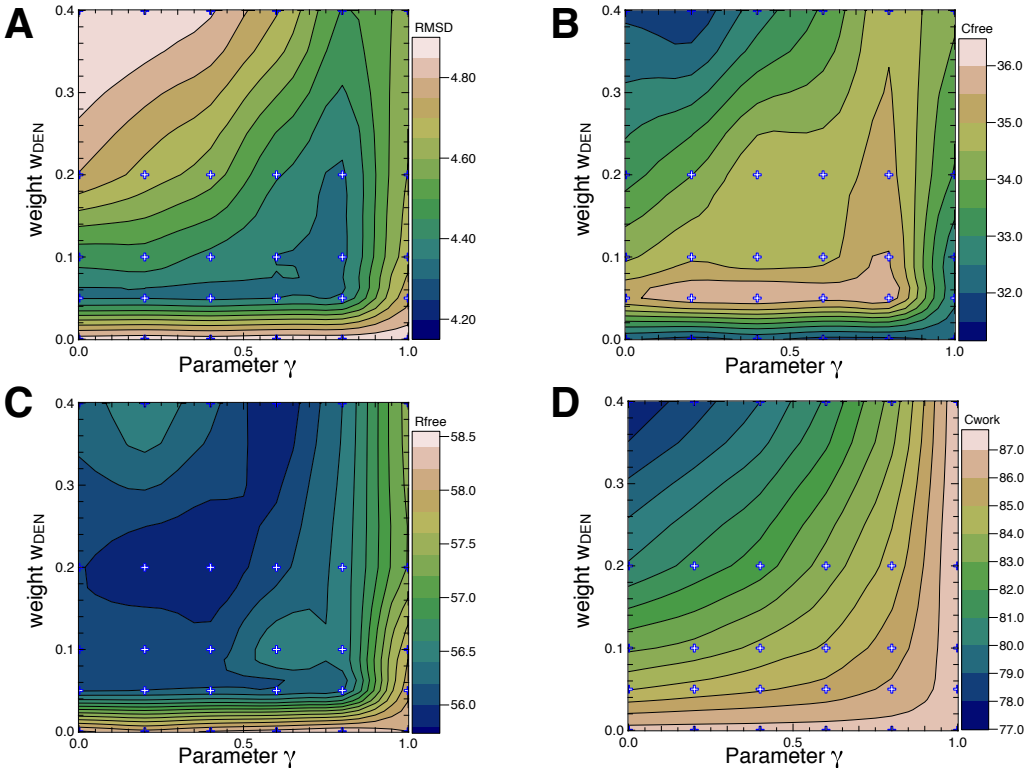

Figure 6.4: Refinement results for the most difficult case 1hrd at 10Å. The contour plots show the same quantities as in Figure (6.3). The optimal region is significantly smaller than for the easy case 1ake. However, high C_{free}-values correlate well with low RMSD values, even though the refined structure is still far away from the correct solution; the best RMSD value is 4.3 Å.

all six cases is 0.84. However, R_{free}^{rnd} is not correlated with the RMSD, with an average correlation of -0.18.
It should be noted that the R-values are here typically significantly larger than what is observed for diffraction data. The typically upper limit for the R-value of 59% as obtained for random atom positions [93] does not hold here, since the power spectrum of the target density map is significantly different from the power spectrum of the model density map. The reason for this is that a CTF function has been applied to the images but not to the model density map that is computed by DireX during the refinement. We plan to correct the model density calculation for this effect in a future version of DireX.
An overview of the results of all refinements is shown in Table 1, where refinements using optimum DEN restraints are compared to refinements without DEN restraints. For all cases the optimum DEN parameters are determined by the maximum $Cfree$ value. Without DEN restraints most models are strongly over-fit, leading to a lower RMSD as compared to when using optimal restraints. As expected, C_{work} is higher for these models, since without restraints the model can be refined further to fit the density better. The C_{free} value is instead always higher for the optimally restrained model and is therefore in all cases able to detect the better model. C_{free} can detect the better model even when the model is far from the correct structure as is demonstrated by the 1hrd case. The template used for building the model for 1hrd has a relatively low sequence identity (28%), the initial homology model has therefore several regions with wrong secondary structure, loops, etc., which cannot be corrected by refinement alone but instead would need extensive remodeling. However, even though the RMSD values of the refined structures lie in the range of $4.3 - 5.0$ Å, the low-RMSD structures still yield the higher C_{free} values (cf. Fig. 6.4 and Fig. 6.6, 4A-B).

6.2.2 Model Quality versus Spatial Frequency Cutoff

For cross-validation data need to be left out which necessarily impacts the quality of the refined structure. The information content of cryo-EM density maps varies for different Fourier shells. The lower frequency shells obviously contain little information on high-resolution details, but for increasing spatial frequency the shells also contain increasing amounts of noise. One can therefore expect that there is an optimum choice of the cutoff of the spatial frequency, ν_{max}, which best trades off resolution and noise. Choosing a low value for ν_{max} ignores high-resolution signal in the data and prevents the refinement to improve structural details in the model. On the other hand, including high frequency components can be expected to be detrimental in the refinement, as these high frequency components will contribute excessive noise to the density map.
To test how the result depends on this cutoff value, ν_{max}, we performed refinements for all three synthetic cases with different cutoff values. As a quality indicator for the refined structure we computed the RMSD to the correct structure. Figure 6.5 shows these RMSD values for different cutoff values, ν_{max}, and for the 5 Å (solid line) and 10 Å (dashed line) data sets of all three starting models. It should be noted that ν_{max} does not correspond to the resolution but just determines which

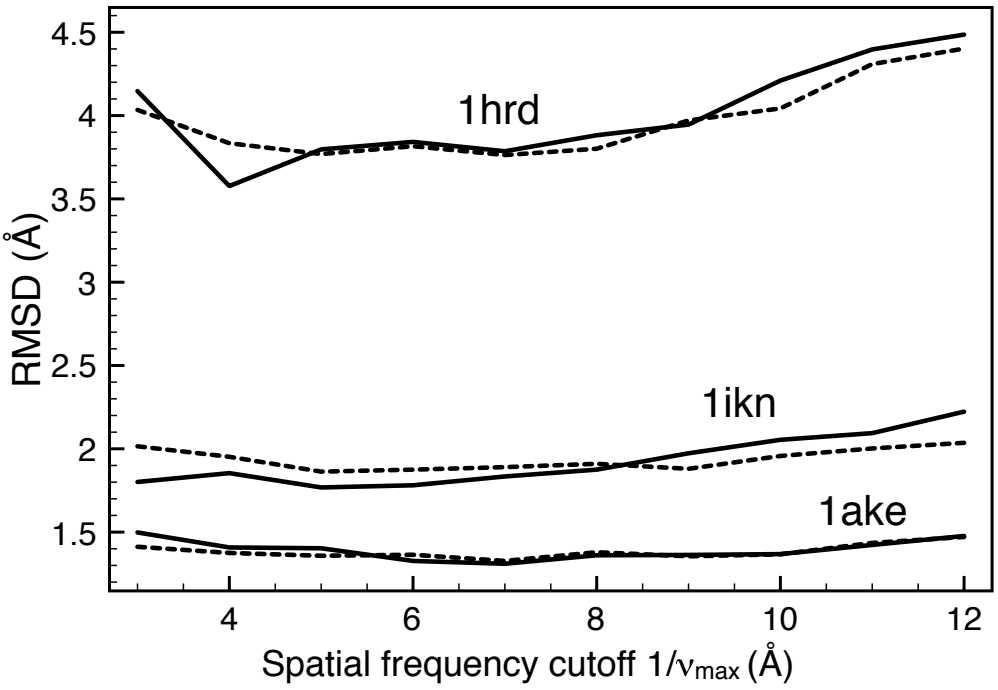

Figure 6.5: Model quality versus spatial frequency cutoff. The root-mean square deviation (RMSD) is shown for all three cases (1ake, 1ikn, and 1hrd) at both resolutions, 5 Å (solid) and 10 Å (dashed) as a function of the higher frequency limit of the work band, which is identical to the the lower limit of the free band. Only the work density map, composed of Fourier components from the work band, is used for the refinement. Overall the RMSD does not depend strongly on this frequency cutoff. The RMSD noticeable tends to increase for higher frequency cutoffs because of the lower signal-to-noise ratio in this frequency range. Note that the data have a resolution of 5 or 10 Å ($FSC_{0.5}$ criterion), which means that with a cutoff at, e.g., 3 Å, the density maps contains large amounts of noise. However, DireX is not very sensitive to this noise since it uses a robust method to compute forces on the atoms.

Fourier components were used to compute the density maps from the 5 Å and 10 Å data sets.
Overall the quality of the models does not depend strongly on the frequency cutoff. One reason for this is that the main conformational change between the homology model and the correct structure is captured well already by the lower frequency components. More serious errors such as register shifts or regions with wrong secondary structure cannot be corrected by refinement alone but instead would need extensive remodeling. Some of the gross errors in the 1hrd model might for example be correctable with the 5 Å data set and a large frequency cutoff. Such automatic or manual model rebuilding is, however, beyond the scope of this work The detrimental effects of the noise at higher frequency cutoff are reduced by the particular algorithm, which DireX employs to refine atomic models. Rather than computing an analytical gradient to optimize the atomic coordinates, a stochastic

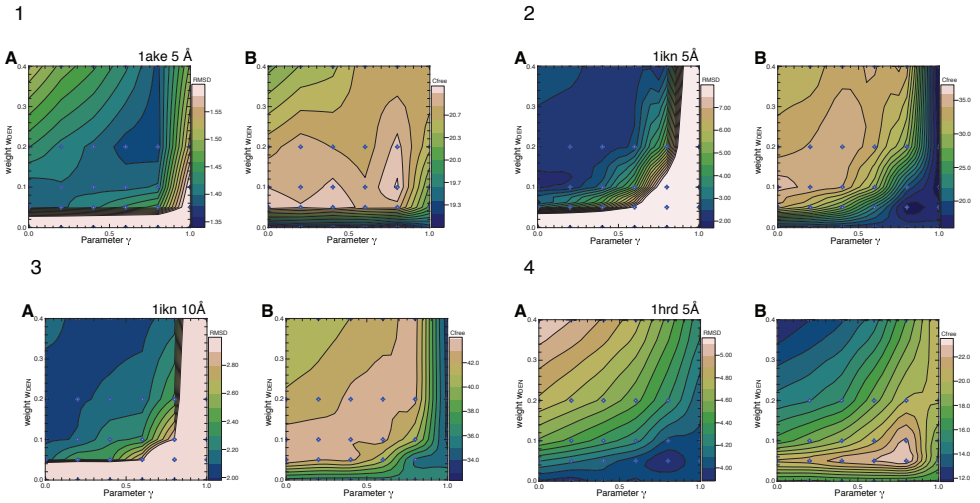

Figure 6.6: Refinement results for 1ake at 5 Å(1), 1ikn at 5 Å(2), 1ikn at 10 Å(3) and 1hrd at 5 Å(4). Contour plots showing (A) the root-mean square deviation (RMSD) between refined model and correct crystal structure and (B) the free correlation, C_{free}, the free R-value, as a function of w_{DEN} , and the deformability, γ, of the elastic network restraints. While in (2) and (4) the dependency is obvious, there is no fit of both functions in (1) and (3). In both cases the C_{free} value does not match the best RMSD but is still describing areas with smaller RMSD values. This can be due to the small overall RMSD values (< 3Å) compared to the refinements of (2) and (4).

gradient is computed by scanning the local environment of each atom, which makes it relatively robust against even very noisy density maps [78].

6.2.3 Application to Real Data of GroEL

One complication with testing a new method on real data is that the correct structure is not precisely known. We chose GroEL as a test case, since it has been studied extensively by both X-ray crystallography and cryo-EM. The crystal structure by Braig et al. [8] (PDB ID 1OEL) fits relatively well to the cryo-EM density map described by Stagg et al [83] (EMD-1457). The cryo-EM has a resolution of 5.4 Å measured by the $FSC_{0.5}$ criterion.
Since the exact high-resolution structure corresponding to the cryo-EM density map is not known, we chose to compare our refinement to a conservative rigid-body fit instead. The generation of this conservative model is motivated by the observation that, when comparing the conformations of individual subunits in different GroEL crystal structures, the conformational differences can be captured to a large extent by breaking each GroEL subunit into three rigid domains: an equatorial domain, an intermediate domain, and an apical domain. These three domains taken from the crystal structure (1OEL) were docked individually as rigid bodies into the density map using the program Chimera [68]. The obtained atomic model fits the density very well and can be assumed to not to be overfitted to the density as only 18 degrees of freedom (three domains with translational

and rotational freedom) were used per subunit. This model serves as our target structure for comparison with the cross-validated DEN refinement with DireX.
The structure refinements were started from the complete crystal structure including all 14 subunits. Figure (6.7, A) shows a superposition of the starting model (yellow), the refined structure with optimum DEN restraints (blue), the conservative three-domain rigid-body fit (green) for one subunit, and the density map (gray).
The density map of GroEL was obtained from the EMDatabank (http://www.emdatabank.org, EMDB ID 1457). The map was not filtered [83]. As a starting model we chose the crystal structure (PDB ID: 1OEL). As a conservative rigid-body fit, we split the subunit of the crystal structure into three domains: the equatorial domain (residue ranges 2 - 136 and 410 - 525), the intermediate domain (residue ranges 137 - 191 and 374 - 409), and the apical domain (residue range 192 - 373), and fitted each domain as a rigid body into the density map using the program Chimera. This models serves as the target for comparison with the DEN refined structure.
The contour plots (Fig. 6.7, B-F) show the results of the DEN parameter grid searches, which were done similar to the three synthetic cases described above. The only difference is here that the weight w_{DEN} of the DEN restraints is kept constant, instead only those DEN restraints that involve loop regions are weighted with the factor $w_{Loop-DEN}$ (see Methods), which accounts for the fact that α-helices or β-sheets are usually structurally more conserved than loop elements.
For comparison we performed two complete sets of DEN parameter optimizations with two different choices of the free band: a narrow band of $5-6$ Å and wider band of $5-9$ Å extending to lower frequencies. The wider band results in a lower resolution of the work density map used for the actual refinement. For both choices a grid search for optimal DEN parameters was performed. All 14 subunits (53858 atoms in total) were refined simultaneously into the entire density map in each DireX run, which consisted of 200 steps. The runtime of each run was about 90 min. The number of DEN restraints was chosen as three times the number of atoms. The strength of the DEN restraints (DireX parameter *den_strength*) was kept constant at the value of 0.5. Those DEN restraints that involved loop residues were scaled by a factor w_{Loop}-DEN (corresponds to DireX parameter *den_secstr_loop*) which was changed in steps of 0.2 between 0.0 and 1.0.
For the narrow free band, the best RMSD to the rigid-body fit is 1.13 Å (red circle in Fig. 6.7, B). The highest C_{free} value yields a RMSD value of 1.17 Å , which is very close to that of the optimal structure. The highest C_{free} values are obtained for larger γ-values than the lowest RMSD values. This means that the cross-validation suggests that the structure is allowed to be deformed more than the three-domain rigid-body fit without being overfitted, which seems reasonable given the relatively high resolution of the work density map (high-frequency cutoff 6 Å). The C_{work}-contour plot (Fig. 6.7, D) shows the highest values for large γ-values corresponding to easily deformable structures, which are significantly over-fitted. The largest C_{work} yields a structure with an RMSD of 1.48 Å to the rigid-body fitted structure.
The wide free band results in a lower resolution of the work density map (high-frequency cutoff at 9 Å) than the narrow band. For the wide free band the best

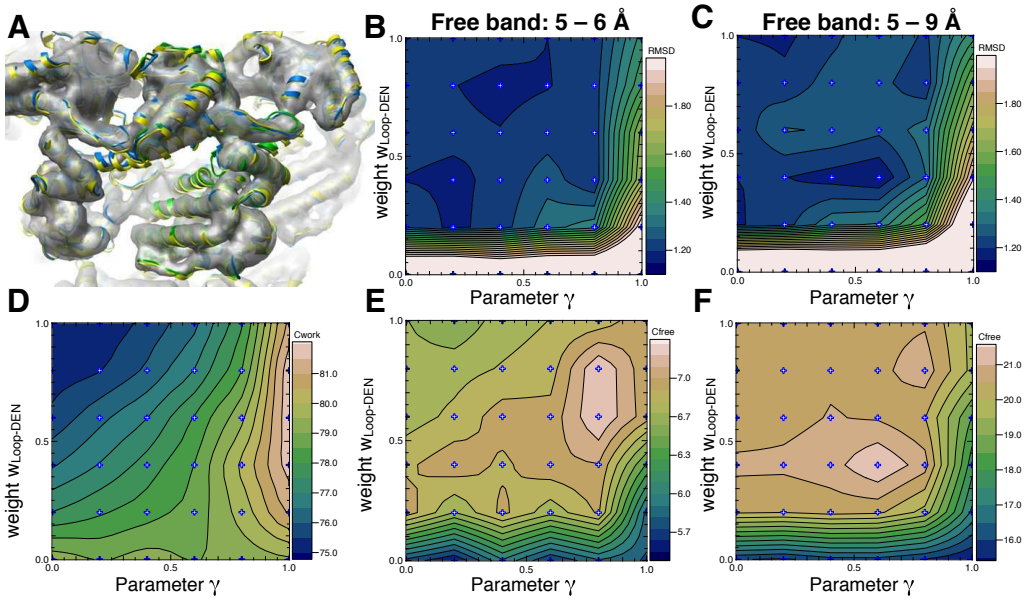

Figure 6.7: Refinement of a GroEL crystal structure (PDB ID 1OEL) against an experimental cryo-EM density map (EMD-1457) at a resolution of 5.4 Å ($FSC_0.5$ criterion). The structure refined with optimum DEN parameters is compared to a model that was obtained by docking the equatorial, intermediate, and apical domain of one GroEL subunit individually as rigid-bodies into the density map. (A) Shown is a superposition of the starting model (yellow), the DireX/DEN refined structure (blue), the three-domain rigid-body fit (green), and the density map (gray). Results of the refinements with different choices for the DEN parameters γ and $w_{Loop-DEN}$ are shown as contour plots for two different choices of the free band 5-6 Å ('narrow band', B,D,E) and 5-9 Å ('wide band', C,F). (B) Shows the RMSD of the DEN refined model to the three-domain rigid-body fit for the narrow band. (D) Cwork is largest for the highest deformability of the elastic restraints, which corresponds to high RMSD values. (E) The optimal C_{free} value yields a relatively low RMSD value but corresponds to a larger deformability of the elastic restraints than the lowest RMSD values, indicating that with a work map resolution of 6 Å (the upper limit of the free band), it is justified to allow flexibility during the refinement instead of fitting the individual domains as rigid-bodies. However, for the wide band with a lower resolution cutoff of the work map (9 Å), the lowest RMSD (C) is obtained exactly for those DEN parameters for which C_{free} (F) is highest, which means that the optimal solution determined by cross-validation is most similar to the rigid-body fit.

RMSD to the rigid-body fit is 1.10 Å (red circle in Fig. 6.7, C) and the highest C_{free} values now coincide exactly with the lowest RMSD regions. That means, at this lower resolution, the cross-validation identifies as optimum the solution that is most similar to the rigid-domain fit with the advantage that there is no need to know in advance where to break the protein into rigid domains. This example demonstrates that the cross-validation approach is able to determine depending on the resolution how strongly the structure needs to be restrained to prevent over-fitting. At lower resolution the optimum structure convergences to the structure obtained by rigid-body fitting.
The focus in this example is on the rather small structural differences to demonstrate the sensitivity of this cross-validation approach. Larger deviations from the optimal structure, due to either under- or over-fitting, are usually even easier to identify.

6.3 Results

The refinement of large biomolecular structures against low-resolution density maps obtained from single-particle cryo-EM is highly susceptible to over fitting, as the number of parameters, i.e. the atomic coordinates, is typically much larger than the number of experimental observables. We present an approach for the cross-validation of structure refinement against such cryo-EM density maps that is able to detect over fitting. The structure factors that are omitted from the work set and used for validation are taken from a spatial frequency range with a relatively low signal-to-noise ratio. These structure factors are typically not reliable for direct interpretation and are usually ignored. Their inclusion in the density map calculation would lead to an increased level of noise in the density. However, the signal in this frequency range is still strong enough for validation: a significant increase in the free map correlation, even if the absolute value of the correlation is low, can be assumed to be most likely due to an improvement of the model, since information from this free frequency range has not been used for the refinement.
The broader this test frequency range is and the more it extends towards low frequencies with larger signal-to-noise ratio, the more robust is the validation measure. However, the more signal is omitted and not used for the refinement, the lower is the quality of the refined structure. We think the frequency ranges proposed here provide a good trade-off for most cases, but it is possible that in other situations a larger range could be necessary or a smaller range might be sufficient.
We proposed a measure, the free map correlation C_{free}, for which we have shown that it correlates well with the overall correctness of the model. Refined structures with a large C_{free} value also have a low RMSD to the correct structure for three test proteins with simulated data. This means that C_{free} can be used to optimize the choice of restraints and their strengths used during the refinement. Depending on which optimum parameters are chosen by C_{free}, DEN refinement can cover the entire range from completely unrestrained positional refinement to (almost) rigid-body fitting.
In X-ray crystallographic refinement with high non-crystallographic symmetry (NCS), which is for example the case for icosahedral viruses, cross-validation with

a random choice of test set reflections cannot be used due to strong correlations between structure factors imposed by the high symmetry. In analogy to the approach proposed here, it is conceivable that reflections in the non-complete high-resolution Fourier shells, which are usually neglected, could be used as test set reflections.
The cross-validation approach itself is independent of the particular choice of restraints, so we expect that our approach is of general applicability and can be used to optimize very different types of restraints as used by all other flexible fitting or refinement tools. For example, C_{free} could be used in elastic normal mode based fitting to determine the optimum number of eigenmodes to be included in the fitting. It should also help to decide whether, in the case of very low-resolution data (> 10 Å), flexible refinement can be justified at all, or whether rigid-body fitting should instead be pursued. Finally, we expect that this cross-validation approach increases the reliability of refined structures and reduces mis- or over-interpretation of noisy and low-resolution density maps obtained from cryo-EM experiments.

Chapter 7

Determination the Principal Motions of the Cryo-EM Data

7.1 Bootstrapping the Density Reconstruction

A single 3D reconstruction can be determined from a stack of single-particle images obtained from cryo-EM micrographs. In addition, as described in Section 3.2.3, the variance of a dataset can be explored via bootstrapping by generating an ensemble of density maps. This ensemble is typically used to calculate a variance map, which describes the density fluctuations at each individual grid point. However, this density ensemble further includes the dependencies of density fluctuations between different grid points. We here develop an approach to determine correlated fluctuations in the density. This approach is applied to two data sets of chaperonin molecules, for which large scale conformational motions have been suggested.

7.2 Chaperonins as Test Systems

Chaperonins are protein complexes involved in assisting the folding of newly synthesized proteins. The typical architecture of chaperonins involve a barrel-like structure with a central folding chamber. The unfolded substrate enters the chamber, which closes upon ATP hydrolysis to initiate folding of the substrate. Finally the folded substrate is released.
Two classes of chaperonins are distinguished, group I chaperonins like GroEL are found in prokaryotes, have a cofactor GroES to close the folding chamber, while group II chaperonins close the chamber with a built-in lid, requiring a large conformational rearrangement. Group II chaperonins are found in eukaryotes (TRiC) and archaea. We studied in detail the chaperonins Mm-CPN from the archaea Methanococcus maripaludis and GroEL/ES from the bacterium E.coli (Fig. 7.1). GroEL consists of two homo-heptameric rings stacked together back-to-back, where each ring forms a reaction chamber. The cofactor GroES is a homo-heptamer, which binds to one side of GroEL thereby closing one reaction chamber. The ring to which GroES binds is referred to as the *cis*-ring, while the opposite ring is called the *trans*-ring.

The Mm-CPN is a homo-hexadecamer consisting of two rings with eight subunits each. Both rings form a reaction chamber, which is closed by an iris-like motion of the helical protrusions of the apical domains. Both GroEL and Mm-CPN monomers are typically segmented into three rather rigid domains, the apical, intermediate, and equatorial domain (cf. Fig. 7.1).
Both data sets for Mm-CPN and GroEL/ES have been published as single reconstructions by Zhang et al. [102] and Chen et al. [12], respectively. Bootstrapped density maps were computed for both data sets using the the *calculateMapVariance.py* program of the EMAN toolkit with the same set of parameters as were used for the originally published reconstructions. The resolution of the obtained densities is about 8Å. For GroEL/ES and Mm-CPN, 100 and 99 densities were generated, respectively.
The Mm-CPN wild-type showed a strong orientational preference in the experiment, which limited the resolution of the reconstruction. Therefore the helical protrusion in the apical domain was truncated by 22 residues, which resulted in an increased number of side-views and a consequently increased resolution.

7.3 Analysis of Eigenvolumes

Recently it has been suggested [82, 63] to calculate eigenvolumes from bootstrapped density maps and to interpret them in terms of the underlying conformational changes of the protein. These eigenvolumes are calculated by applying a PCA to an ensemble of bootstrapped densities.
The PCA on the densities maximizes the density fluctuations which means that the largest eigenvalues correspond to the largest correlated volumetric changes.
We are ultimately interested in the motions of the protein structure itself, and, thus, need to translate the density fluctuations into atomistic fluctuations. It is however important to realize that the largest volumetric change is not necessarily caused by the largest atomistic motion. Vice versa, large protein motions do not necessarily cause large density changes.
To examine this effect, we consider the density map of a helix at a resolution of 8 Å (see Fig. 7.2). The axis of the helix is aligned along the y-axis. We will discuss the effects of translations of the density, exemplified by 4 Å shifts along the three coordinate axes. The corresponding RMSD values between the initial and shifted helix positions is therefore 4 Å. These simple translations will always lower the cross-correlation between the initial and the translated density map.
The cross-correlation for the maps translated along the x-, y-, and z-axis is .87, .93, and .69, respectively (Table. 7.3). The smallest change of the cross-correlation compared to the perfect overlap is obtained for a shift along the helical axis (y-axis). The same 4 Å shift along the perpendicular x-, and z-axis leads to a larger decrease. Obviously, the change in correlation depends on the shape of the molecule.
In the same way, a rotation of 180 degree around the helical axis yields a high cross-correlation of .91. So, even dramatic conformational changes might yield very small correlation differences. It is therefore not possible to deduce the extent of the atomic motion from the change of the cross-correlation.

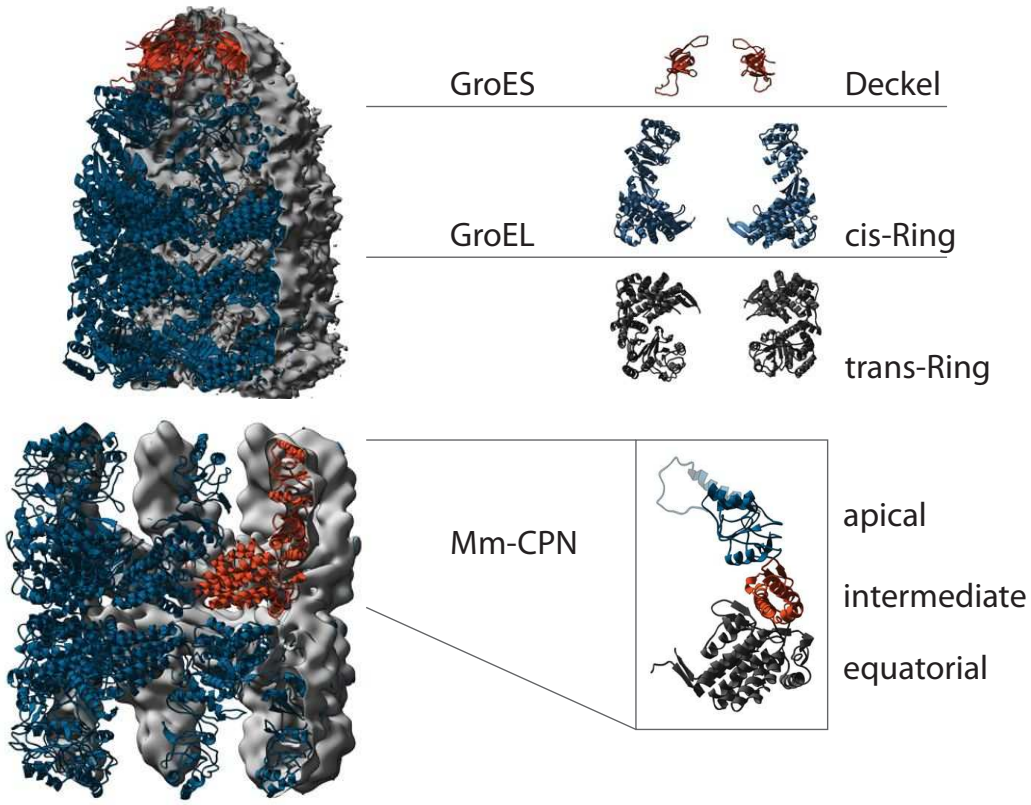

Figure 7.1: Shown at the top is the GroEL/GroES complex (PDB ID: 1AON) superimposed to the average density of the bootstrapped ensemble. The front half of the density map was removed for clarity. The GroES heptamer (orange) closes the *cis*-ring (upper reaction chamber) of GroEL (blue). The lower *trans*-ring is in a more compact conformation. At the bottom the Mm-CPN (PDB ID: 3IYF) is shown with its average density superimposed. Again, the front half of the density was removed for clarity. The density is weaker in the apical regions and the secondary structure is entirely outside the density surface. We studied a genetically engineered version of Mm-CPN where the helical protrusion was truncated by 22 residues. On the right side a single subunit of the truncated Mm-CPN is overlaid with the wild-type (transparent) subunit in the apical domain. The individual subunits of both Mm-CPN and GroEL are usually segmented into three rather rigid regions: apical, intermediate, and equatorial domain.

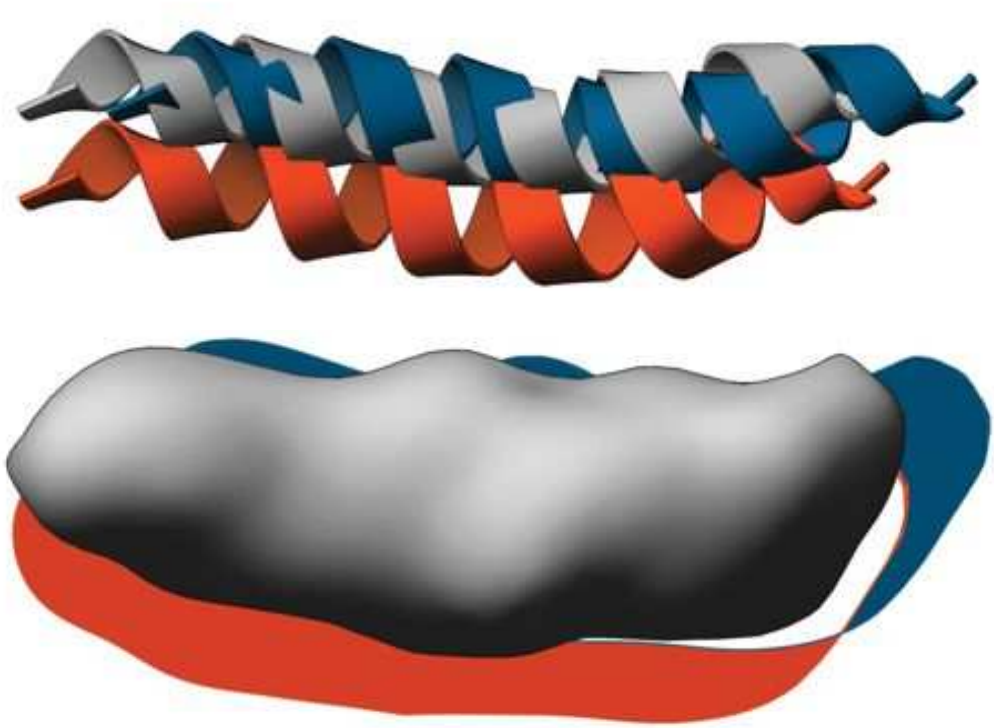

Figure 7.2: On top the helix is shown forming the lid in the native Mm-CPN structure (3LOS). In grey the initial structure is drawn, which was used as the reference for calculating the RMSD in Table 7.3. In blue the helix is shifted by 4 Å to the right, which corresponds to the y-axis. In orange the helix is translated in the perpendicular plane along the z-axis. The RMSD of both translated helices is 4Å calculated with the initial grey structure as reference. Just by the structures the correlation in the densities can be estimated to be very different. At the bottom the reference density is shown in grey. In blue areas are marked, which are occupied only by the helix shifted to the right, in orange the areas for the helix shifted orthogonal to the helix axis are marked. It is easy to see that both in RMSD identical shifts have a different overlap with the initial structure, that will imply large effects on the covariances.

As a result, the PCA of a given density ensemble will yield eigenvolumes that describe the uncorrelated fluctuations with the largest variance, which not necessarily correspond to the largest atomic fluctuations. In the synthetic test case of the helix any 4 Å translation along the coordinate axes had the same impact on the structural variance, while the shift along the z-axis will dominate the variance of the densities and would yield the largest eigenvalue.
Finally, the extent of conformational changes obtained from the density PCA is different from the actual atomistic conformational changes. The shape determines which components of the atomistic fluctuations cause the largest change in density. The components of the fluctuations are therefore weighted differently in the density PCA, due to their impact on the density variation.
It is of course still interesting to perform a density PCA to analyze bootstrapped density ensemble, as it helps to reveal dominant conformational fluctuations.

Translation	RMSD (Å)	Correlation
x-axis	4.0	.87
y-axis	4.0	.93
z-axis	4.0	.69

Table 7.1: A α-helix aligned on the y-axis is translated by 4 Å along the axes of the coordinate system (Fig. 7.2). The RMSD of each shift is 4 Å but the correlations of the densities with the initial state varies depending on the direction. The shift along the axis of the helix (y-axis) causes very small loss of correlation, while a shift perpendicular to the helix axis can cause large changes in the correlation. While the translation along the x-axis induces just a little bit smaller correlation a shift along the z-axis reduces the correlation by 0.31.

However, our goal is to interpret the bootstrapped density ensemble in terms of atomistic motions. Because of the described effects, a density PCA is not optimal for this interpretation and it becomes evident that an ensemble of atomistic models is needed that represents the information from the density ensemble. For this, we refined atomistic protein structures against each of the density maps to obtain an ensemble of protein structures.

7.4 Refinement of Atomic Models against Bootstrapped Densities

Due to the fact that in cryo-EM multiple proteins of same type are on the micrograph, there is at least a small conformational variance and this will be part of the reconstruction. The previous explained technique of bootstrapping can be used to get volumetric representations of the conformational space. As shown previously, it is inadequate to perform a PCA on the densities to obtain dominant atomic fluctuations of a system, in contrast the bootstrapped densities have to be translated into an atomistic representation of the data. The variance of the bootstrapped densities can then be expected to be represented to a large extent by this atomistic ensemble.
The atomistic ensemble is obtained by refining a starting model against each of the bootstrapped densities individually. For GroEL/ES the refinement was started from a crystal structure (PDB ID: 1AON) and for Mm-CPN the refinement was started from a previously determined model for the open state that was based on a homology model built from the crystal structure of the thermosome [102, 20].As the fitted models should capture fluctuations around the average conformation it is in general advisable to start the refinements from a model that either has been fitted to, or, as in our cases, is close to the average density.
For the refinement the flexible fitting program DireX was used. DireX allows to use only a small parameter set to control the refinement as described in Section 5. A density map is computed from each fitted model that is compared to the target density. This comparison was then used to find the optimal parameters in DireX, as described in the next Section below.

The resolution cutoff (rectangular low-pass filter) was chosen according to the Fourier Shell Correlation (FSC) [72] of the original reconstructions. For GroEL/ES the cutoff was set to 7 Å and to 8 Å for Mm-CPN. This implies that higher frequencies have not been used at all in the refinements and the model densities have been calculated with the same cutoff.

7.4.1 Choice of the Resolution Cutoff

The first important parameter is the resolution cutoff, which is necessary to avoid over-fitting and interpreting noise as fluctuations of the structure. The basic method to calculate the resolution is to use the Fourier Shell Correlation (FSC) [72]. This methods assume that the FSC will converge to 0.0 for high frequencies and there will be only uncorrelated noise that spectra.
This results can not be used for bootstrapped density ensemble, probably the resolution of each density is still in that area, the focus is on the differences between multiple densities. According to the bootstrapping method, the density ensemble has some unique characteristics. The FSC of two bootstrapped densities converges to all value of about 0.45 for Mm-CPN and for GroEL/ES it is about 0.2 (c.f. Fig. 7.3).
The reason for this is that the dataset used in the reconstruction is not independent from the others anymore, because each density shares a certain amount of images as basis for the reconstruction. The Result is correlated noise. In principal this should not be a problem, but in each of the reconstruction of the bootstrapped densities single images have been used multiple times. So at the same resolution the FSC has converged, the information in one of the bootstrapped maps can not be separated from those artifacts.
To avoid any influence of those effects a resolution cutoff was chosen according to the FSC of bootstrapped maps. For GroEL/ES the resolution was set to 7 Å in the refinement and to 8 Å for Mm-CPN. This implies that higher frequencies have not been used at all in the Refinement and the model densities have been calculated with the same cutoff.
After the refinement a ramp filter was used on the bootstrapped ensembles to reduce the remaining noise and to allow an optical comparison of the noisy bootstrapped maps and the smooth model maps.

7.4.2 Optimization of the Refinement

The challenge is to find parameters for which the fitted models describe the differences between the densities in a significant way. It is essential that enough of the characteristics of each bootstrapped density is transferred to the atomistic model. Otherwise, the information on the conformational variance would not be encoded in the ensemble of atomistic models.
The optimal parameters that control the restraints in DireX were determined in an iterative way: starting with very strong restraints, which makes the structure very stable and rigid, the stiffness was incrementally decreased. At each iteration the fit of the model was analyzed by calculating the cross-correlation between the model densities and each target density (Fig. 7.4). As the criterion for the best parameters we required that the cross-correlation between the model densities and

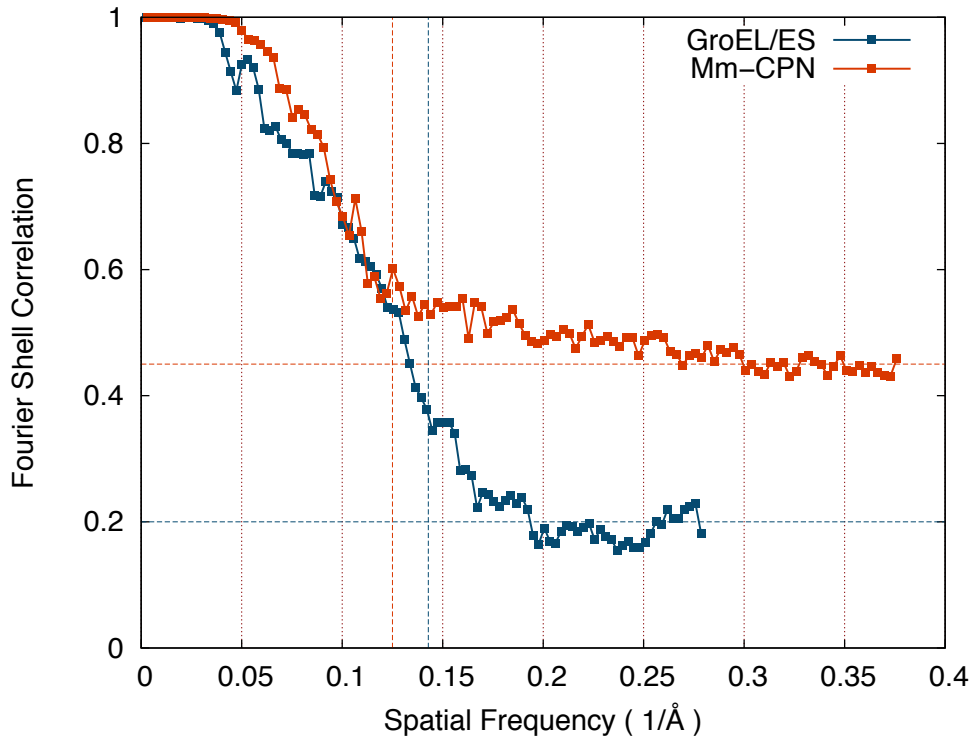

Figure 7.3: The FSC of two bootstrapped maps plotted for GroEL/ES (blue) and MM-CPN (orange) are plotted. For Mm-CPN the FSC drops till 0.45 and for GroEL/EL the FSC converges to 0.2. The resolution choose in the refinement are marked by horizontal lines. The Resolution is always chosen at a position the FSC has not reached the convergence level. The vertical lines mark the used resolutions in the refinement procedure. For Mm-CPN 8 Å were chosen and for the GROEL/ES complex a little bit higher resolution of 7 Å .

their corresponding target densities was higher than the correlation between the model densities and all other bootstrapped densities. That means each model was required to yield the highest correlation to the density it has been fitted to (Fig. 7.4). The iteration was stopped when this criterion was reached. In other words, those restraints were chosen that were as strong as possible, but still allowed to fit the models such that they are closer to their respective target maps than to any other bootstrapped map.
In Figure 7.4 the cross-correlation values are plotted as projections onto the target and model ensemble of density maps for GroEL/ES and Mm-CPN. The target density maps are numbered from 1 to 100 and model i denotes the model that was refined against target density i. Figures 7.4 A and B show the correlations of the refined model densities plotted versus their target densities, which are the bootstrapped densities calculated from the experimental data set. Each value on the x-axis represents one of the target maps and each dot is the correlation of one of the 100 model maps with this target map. The correlation between model density i and target density i is shown in blue. Obviously, the blue dots are either the best or at least among the best correlation values. This means that the restraints fulfill the criterion for the best parameters and that the ensemble of fitted models captures most of the conformational variance.
Comparing one target density to all model densities yields a small range of correlation values (Fig. 7.4 A and B). However, comparing one model density to all target densities yields a large range of correlation values (Fig. 7.4 C and D). Surprisingly, while the blue dots in Fig. 7.4 A and B almost always yield the highest correlation, this is not the case in Figure 7.4 C and D. That means if a model fits better to its target density than any other model, there could be another density that fits even better to this particular model than the density the model has been refined to. Some densities therefore appear to be more difficult to fit with a single model than others.
These observations can be understood by considering that the model fitting employs restraints to maintain a reasonable atomic structure (stereochemistry, secondary structure, side-chain packing, etc.). Clearly, a density that is unphysical cannot be fitted without violating at least some of those restraints. Two effects influence the physicality of a bootstrapped density map: 1) noise and 2) the particular mixture of conformations that was used to generate the density map. Noise obviously has a random effect of how close a density is to the true structure and always makes the density maps more unphysical. Furthermore, each reconstructed density map is composed of a mixture of different conformations. These conformations are weighted differently in the bootstrapping. The different weighting can lead to combinations of conformations that are more unphysical than others, which means some densities can be fitted better with a single model than others.
This problem is not that significant if the overall conformational variance is small, the average over similar conformations can still be represented well by a single model. However, if the conformational variance is large, the average over these very different conformations can in fact be far from a reasonable single conformation.
In the ideal case where each target map of the ensemble represents a perfectly reasonable protein conformation, the each fitted model fits best to its corresponding target density map, and vice versa, each target map fits best to the model that

was used to fit this target map, which means the blue dots in Figure 7.4 A - D should always be on top.
The results for GroEL/ES (Fig. 7.4 A and C) and Mm-CPN (Fig. 7.4 B and D) show a significant difference. The fitted model fits always best to its corresponding target density (blue dots) and worse to all other density maps. Vice versa, the target density map fits better to the corresponding model than to most other models. The picture is much less clear for Mm-CPN. The reason for this is that the open conformation of Mm-CPN has a larger flexibility such that the reconstructed density is averaged over more dissimilar conformational states than in the GroEL/ES case. As a result the density of the apical domains is quite unphysical and cannot be interpreted well by a single model (see Fig. 7.4).
The optimal parameter set for the refinement with DireX have been chosen according to the correlations of the target densities and the calculated model densities of the resulting structures. The goal was that every model density fits its target density as one of the best of all model densities. The main problem is that there is a danger of over-fitting. So it is necessary to restrain the models as much as possible while still permitting the models to fit best to their targets. This is not always possible as seen in the Mm-CPN ensemble, when it becomes important to balance between fitting real information of the density and maintaining the quality of the model to avoid fitting to noise. The quality of the model was monitored by the secondary structure content of the models using the program Molprobity. The percent residues within the allowed region of the Ramachandran plot was as average 60.1% for GroEL/ES and 61.4% for Mm-CPN.
The parameters that have been chosen by this iterative approach are listed in Table 7.4.2 In particular, the γ-parameter was set to zero, i.e. the harmonic distance restraints were not deformable to fix the center of the DEN ensemble onto the input structure and to allow only small fluctuations around the equilibrium state between the density and the initial structure. The number of DEN restraints was set to three times the number of atoms, such that the entire structure was well restrained.
Optimizing the parameters to obtain significant fits on the one hand, and to avoid over-fitting on the other hand, is always a problem for refinement. Since the noise is isotropic and normally distributed, the effect of over-fitting can be assumed to be isotropic and normally distributed as well (Eq. 3.12)[64]. Over-fitting therefore biases the models and perturbs the models in an isotropic way which contributes only little to the largest components of the molecular dynamics. As we are interested mostly in those principal motions of the molecules, the model ensembles are analyzed by a principal component analysis, as discussed below. This statistical analysis helps to filter out the isotropic noise by averaging and linear decomposition[78].

7.5 Calculation of Positional Variance and B-factors

Before the biological implications of the protein dynamics can be discussed, the variance of the fitted model ensemble is quantified and analyzed by statistical methods.

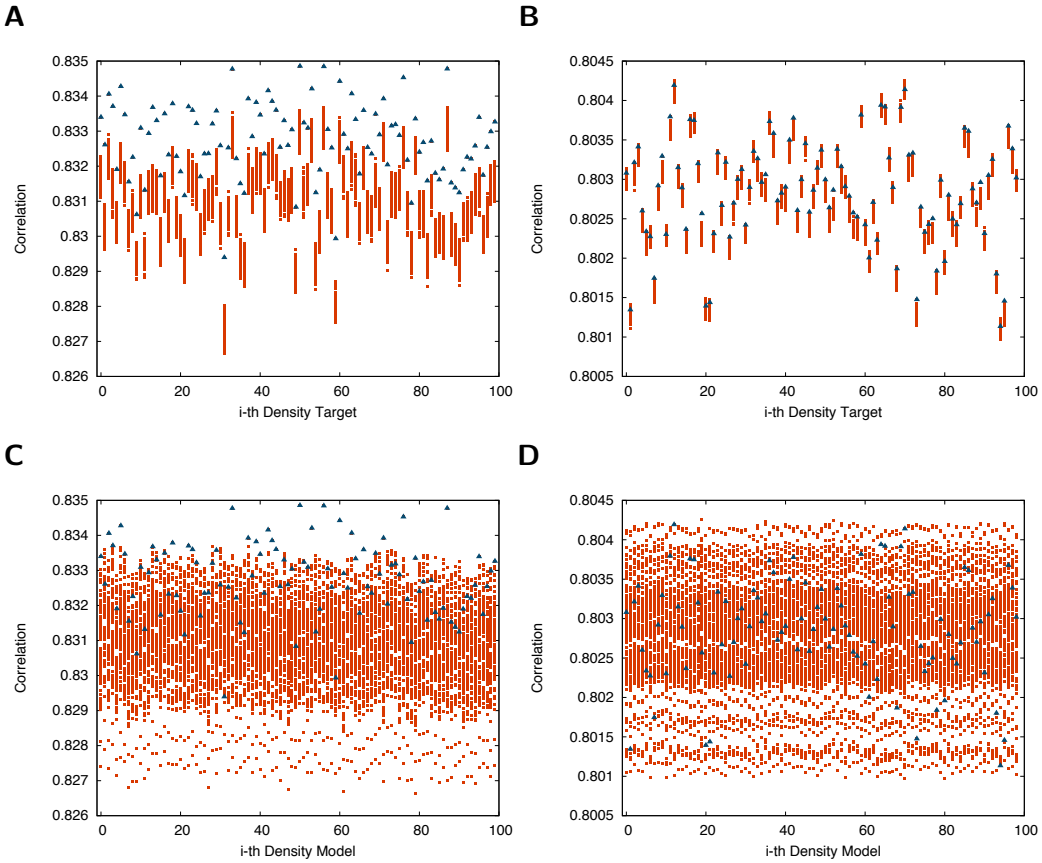

Figure 7.4: The cross-correlation matrices between model and target density maps of GroEL/ES (left side) and Mm-CPN (right side) are shown. A and B show the correlations of each of the 100 model densities (computed from the fitted models) to each of the 100 bootstrapped target densities plotted versus the target density number. The correlation values for which the model number is identical to the target density number is shown by blue triangles, which corresponds to the correlation of the model density with the target density to which the model was fitted to. For GroEL/ES these blue triangles have always the highest correlation compared to all other models (red dots). This is less pronounced for Mm-CPN where the blue triangles are not always on top, but are, however, among the top values. C and D show the same correlation values as in A and B, instead plotted versus the model number. It is obvious that the spread of correlation values is much larger for a given model (A and B) than for a given target density (C and D). In addition, the blue triangles which means that for a given model density there could be target densities that yield a higher correlation than the target that was used to fit this particular model. The reason lies in the fact that the bootstrapped densities are averages over many differently weighted conformations, which determines how well they can be represented by a single model (see text for details).

Parameter	GroEL/ES	Mm-CPN
nsteps	200	200
sampling	concoord	concoord
perturbation	0.0	0.0
DEN_{ratio}	3.0	3.0
$DEN_{strength}$	3.0	3.0
DEN_{lower}	3.0 Å	3.0 Å
DEN_{upper}	15.0 Å	15.0 Å
DEN_{γ}	0.0	0.0
DEN_{κ}	0.2	0.2
$MAP_{strength}$	0.04	0.04
$MAP_{damping}$	30	30
map kernel	gaussian	gaussian
map resolution	7.0 Å	8.0 Å

Table 7.2: List of all parameters that have been modified from the default settings in DireX.

As for the 2D projections of the particles in the experiment, we can assume the particles to be independent and identically distributed. Further it is obvious that the expectation value and variance are well defined and we can use the Lindeberg-Lévy central limit theorem, which states that such a distributed set will converge to a normal distribution, by which we can assume the underlying distribution to be normal if we have used enough samples n:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\right)-\mu\right)\rightarrow N(\mathbf{0},\mathbf{\Sigma}) \tag{7.1}$$

where μ is the expectation value, $\mathbf{\Sigma}$ the covariance matrix and $\mathbf{X}_i$ the i-th observed structure. Here we approximate the distribution by a Gaussian distribution, which is one of the most basic assumptions in statistics. If the motions are uncorrelated the covariance matrix is a diagonal matrix with the variances on its diagonal. Further the variance can be considered as an isotropic attribute of each atom yielding a measure of the positional precision, which is directly related to the B-factor in crystallography. The variance can be expressed as a crystallographic B-factor by

$$B_i = 8\pi^2\sigma_i^2. \tag{7.2}$$

The B-factor from the bootstrapped refined ensemble B_i^{boot} is typically smaller than the real B-factor of the structures, which is explained by Bienaymè's formula:

$$\mathrm{Var}\left(\overline{\mathbf{X}}\right)=\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\mathbf{X}_i\right)=\frac{\sigma^2}{n} \tag{7.3}$$

with $\mathrm{Var}\left(\sum_{i=1}^{n}\mathbf{X}_i\right)=\sum_{i=1}^{n}\mathrm{Var}(\mathbf{X}_i)$. This formula requires the variables to be uncorrelated, which is fulfilled in a cryo-EM experiment as the single particle

images are even independent. We can therefore assume that the real B-factor can be estimated by:

$$B_i^{struct} = n \cdot B_i^{boot}. \tag{7.4}$$

where n is the number of experimental observations that were used in the averaging. We use a simple approximation to estimate the factor n. As the atomic variance is determined by the density variance we assume that this factor n is identical to the factor by which the density variance is reduced upon averaging. Each 2D particle image contributes N^2 data points, a slice of the volume with N^3 data points (voxels). If there are k particle images then each voxel is averaged over k/N data points. With these assumptions B-factors can be estimated for atomic models refined to cryo-EM derived density maps.

As a first step to analyze the conformational variance of GroEL/ES and Mm-CPN we calculated the B-factors for each atom using Eq. 7.3. The atomic variances were calculated from the atomic positions in the model ensemble. Figure 7.5 shows one GroES subunit and two subunits of GroEL from the cis- and trans-ring (Fig. 7.5 A) as well as one subunit of Mm-CPN (Fig. 7.5 B). The left half of each panel shows the atomic model color-coded by the B-factor from white (low) to red (high); for both structures the B-factor values were capped at 10.0Å^2 to get a useful scaling for most of the atoms, since some atoms seem to be too weakly restrained and are fluctuating in the input ensemble. This problem is eliminated by this choice of threshold for the color-coding.

Figure 7.5 shows the B-factors of the refined and RMSD-aligned structures and the coefficients of variations of the densities (CV-map) (Eq. 7.5).

The coefficient of variation c_v is defined as the ratio of the standard deviation σ to the mean μ :

$$c_v = \frac{\sigma}{\mu}. \tag{7.5}$$

The positional atomic variances observed in the fitted ensemble for GroEL/ES and Mm-CPN are 0.17 Ą and 0.08 Ą, respectively,

$$B_i^{\text{struct}} = \frac{k\, n_{\text{sym}}}{N} \cdot B_i^{\text{boot}}. \tag{7.6}$$

where k is the number of total images used for the reconstruction, n_{sym} is the symmetry factor (7 for GroEL/ES and 16 for Mm-CPN), N is the number of grid points along one axis of the density map. The corrected average B-factors are then 600 Å^2 for GroEL/ES and about 900 Å^2 for Mm-CPN, which yields an estimate for the true positional variance for GroEL/ES and Mm-CPN of 3.4 Å and 2.8 Å, respectively.

It should be noted that the absolute values of the positional variances and therefore the B-factors might be underestimated, since the models were strongly restrained during the refinement to not over-interpret the density ensemble.

The largest B-factors of GroEL/ES (Fig. 7.5 A) are located at the bottom, the apical region of the seven chains forming the trans-ring. This fits to the expectation that the open arms in the trans-ring are more flexible until being fixed upon binding GroES. This region can also be identified in the CV-map as the area with the largest deviation on the right side of the Figure. Furthermore, the GroES region

shows large fluctuations, which results in high B-factors for GroES. Clearly, the high B-factor regions strongly overlap with high CV-map values.
Figure 7.5 B shows that Mm-CPN has high B-factors in the apical domains and in between the intermediate and equatorial domains, which corresponds to the region of the nucleotide binding pocket (see. Fig. 7.5). These regions are interesting in the context of the function of Mm-CPN. The apical domains need to undergo large motions during the closing of the entire complex and the observed large fluctuations in the apical domains could be connected to this closing motion. The nucleotide binding pocket shows strong density variation as indicated by large CV-map values, however the average B-factors in this region are not as high. It seems the atoms involved in this motion are rather on the outside of the subunit as indicated by slightly higher B-factors, which is typical for a rotation or shearing motion. A further interesting area seems to be the beta-hairpin which forms an extended beta-sheet with the neighboring subunit and which also have high B-factors.

7.6 Disentangling Significant Motions from Noise

In cryo-EM the data are affected by conformational variance and this variance has a big impact on the resolution of the 3D reconstructions. The conformational variance is dominated by large scale collective motions of the protein which are typically tightly connected to its function. The more local conformational motions are often smaller in size and do not contribute as much to the variance observed in cryo-EM images. If overall the motions are rather small they can be assumed to be linear. The approximate model for an observation $\mathbf{X}_i$ would be:

$$\mathbf{X}_i = \mu + s\mathbf{D} + \varepsilon \tag{7.7}$$

where s is a scaling value, $\mathbf{D}$ a displacement vector and ε the error term. If we assume Gaussian distributions along all these components, the fluctuations can be expressed by a covariance matrix, so that the created ensemble is normal distributed like:

$$\Phi(\mathbf{x};\mu,\mathbf{\Sigma})_n = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right). \tag{7.8}$$

Also, the covariance matrix can be estimated from the bootstrapped ensemble and the components of the conformational changes can be estimated by the eigenvectors of the covariance matrix. This is the well known technique of Principal Component Analysis (PCA), which is based on diagonalizing the covariance matrix to solve the eigenvalue problem. As long as the number of observables is smaller than the dimensionality the covariance has to be assumed being underestimated. So the resulting distribution is a weak estimator for the conformational local phase space of the specimen (cf. Chapter 4).
But if we take into account, what was mentioned above about the collective motions and its linear expression (Eq. 7.7), it is obvious that the covariance matrix encodes linear components *and* isotropic vibrations. If the terms of l linear components are

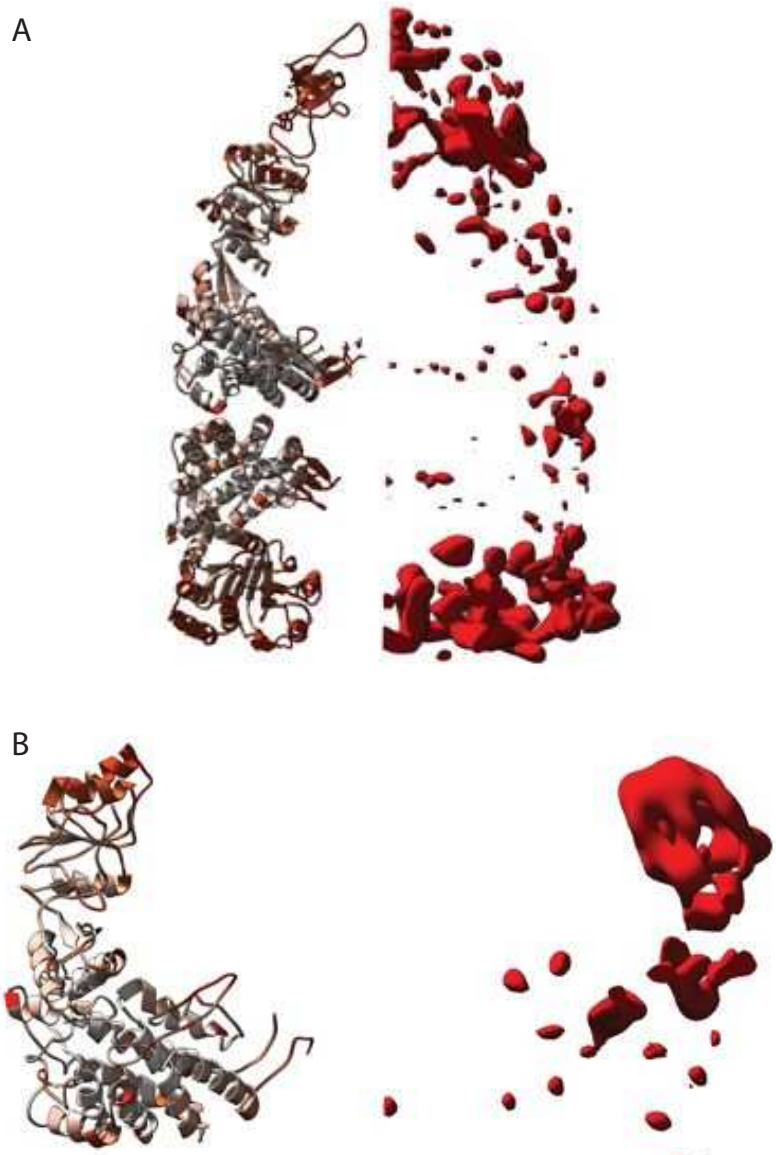

Figure 7.5: On the left side of (A) the averaged B-factors per residue are shown for GroEL/ES for a subunit from the open trans-ring, another from the closed cis-ring and a subunit of the lid, GroES. The right side shows the cross section of the CV-map from the other side of the GroEL rings. The largest B-factors are located in the GroES and at the apical domains of the trans-ring. In between there are only a few spots with high variance. Based on this image large motions can be expected in the trans-ring and at the lid. The same regions shows significant density variance as illustrated by the corresponding CV-map on the right side. (B) shows the same plot as in (A) for a single subunit of Mm-CPN. This corresponds to the asymmetric unit because of the D8 symmetry. Mm-CPN has dominant B-factors in the apical domain and in between the intermediate and equatorial domain. Furthermore the hairpin loops show high B-factors. This is similar to the CV-map where the main variance is located in the apical domain and at the lower end of the intermediate domain. Briefly, in both cases the regions of large B-factors overlap well with regions of major relative fluctuations of the densities.

larger than the vibrations, they will dominate the covariances and the matrix can be split into its components:

$$\mathbf{\Sigma} = \left(\sum_{i=1}^{l} s_i^2 \mathbf{A}_i \right) + \mathbf{\Sigma}_\varepsilon, \tag{7.9}$$

where s_i^2 is the variance along the linear component i, $\mathbf{A}_i$ its direction $\mathbf{a}_i$ written as a matrix and Σ_ε the remaining covariance matrix. At this point a corrected B-Factor B_i^{corr} can be defined from the underestimated covariance matrix Σ_ε. By this the variance can be split into a linear and an isotropic component with variance σ_{iso}^2 which are independently distributed. If the fluctuations are dominated by inherent conformational motions, the isotropic term is negligible in such a direction ($s_i^2 + \sigma_{iso}^2 \approx s_i^2$). The distribution function for the j-th single atom can then be written simply as:

$$\begin{aligned} \Phi(\mathbf{x}_j; \mu_j, \sigma_j) = \frac{1}{\sqrt{(2\pi)^3 \sigma_j^2}} \cdot \exp\left(-\frac{1}{2} \sum_{i=1}^{l} \frac{\mathbf{a}_{ij}(\mathbf{x}_j - \mu_j)}{s_i^2} \right) \\ \cdot \exp\left(-\frac{(\mathbf{x}_j - \mu)^2}{2\sigma_{iso}^2} \right). \end{aligned} \tag{7.10}$$

This leads to a linear approximation of the system around a center μ for which the best estimator of the bootstrapped set is the mean value: $\hat{\mu} = \bar{\mathbf{X}}$. To calculate linear fluctuations for such a system the PCA is a good choice as it describes the ensemble by uncorrelated components.
Finally this can be used to interpret the PCA of the bootstrapped refined ensemble, because dominant linear changes can be estimated independently from the isotropic components and the amount of input data is large enough for valid and robust results. Especially the direction of the eigenvectors can be assumed to be well determined. Briefly, the eigenvectors corresponding to the largest eigenvalues are in general good estimators for the global conformational changes.

7.6.1 Symmetry

A special problem is the symmetry of the specimen or more precisely the symmetry used in the density reconstruction. Both density ensembles have been created using symmetry constraints *C7* for GroEL/ES and *D8* for Mm-CPN. In DireX it is not possible to use symmetry constraints during the refinement. DireX can only use similarity restraints between the subunits, which keeps the subunits similar to each other but does not restrain or constrain the relative position of the subunits according to the corresponding symmetry.
As a solution to this problem we tried to symmetrize the slightly asymmetric ensemble after the fitting the models to the individual densities. The symmetry operation is a rotation around the center of geometry for already aligned entire structures. In general, the symmetric ensemble X has m symmetric subunits with n atoms. The ensemble matrix X consists of n column vectors X_i, which can be split into its subunits $X_{i,j}$. Let R'_j be the corresponding rotation matrix for a single

subunit j depending on the angles ϕ_j, ψ_j and θ_j to rotate onto the first subunit that is used as reference:

$$R'_j = \begin{bmatrix} \cos\theta_j \cos\psi_j & -\cos\phi_j \sin\psi_j + \sin\phi_j \sin\theta_j \cos\psi_j & \sin\phi_j \sin\psi_j + \cos\phi_j \sin\theta_j \cos\psi_j \\ \cos\theta_j \sin\psi_j & \cos\phi_j \cos\psi_j + \sin\phi_j \sin\theta_j \sin\psi_j & -\sin\phi_j \cos\psi_j + \cos\phi_j \sin\theta_j \sin\psi_j \\ -\sin\theta_j & \sin\phi_j \cos\theta_j & \cos\phi_j \cos\theta_j \end{bmatrix} \quad (7.11)$$

This can be used for the multi-diagonal matrix of the rotation of a subunit $R_j = \mathrm{diag}(R'_j, \ldots, R'_j) \in \mathbb{R}^{3n \times 3n}$. The symmetric ensemble will be $X^*_{i+m \cdot j} = R_j X_{i,j}$ and the PCA can be calculated for this set. Without loss of generality we assume that $\mathbf{v_k}$ is the k-th eigenvector of the entire system and it can be split into its components of a subunit $v_{k,j}$ and rotated by the symmetry operator, that for all $j \in 1, \ldots, m$ $v_{k,1} = R_j v_{k,j}$. It is obvious that $v_{k,1}$ is an eigenvector of the ensemble with applied symmetry. The proof is very similar to the one presented for the sparse PCA (cf. Chapter 4) and will not be reiterated here.

This works well for symmetric structures, but the results from DireX are not perfectly symmetric and this asymmetry will affect the results for the symmetrized structures. The applied symmetry will decrease the number of degrees of freedom of the system and slightly different motions on different symmetric subunits will have to be composed into new or split into new eigenvectors. The problem is that the statistical number of degrees of freedom is still lower than the physical number of degree of freedom of the system, for which reason the system will be still underestimated in phase space. As the space is not well defined by the data, the synergetic effect related to the central limit theorem, does not apply for the eigenvectors. So diagonalizing the covariance matrix will in general gather almost parallel motions in different eigenvectors. This is a well known problem for high dimensional vectors. The increased statistical number of degrees of freedom by applying the symmetry will favor such a splitting. As a result less eigenvalues stand out as significantly larger than others and the eigenvectors describe motions that are less global and clear.

Since this effect makes the interpretation of the resulting eigenvectors vague and less clear no symmetry was used in or after the refinement.

7.7 PCA and the Significance of Eigenvalues

Because of the relatively large conformational fluctuations it is not appropriate to assume isotropic variances. The variance therefore needs to be computed in three dimensions for each atom to obtain useful results. This variance can be separated by a PCA into uncorrelated components and variances that could be correlated. By taking advantage of the sparse PCA algorithm(Section 4.2) it is possible to compute the PCA not only for the C_α trace but also for all atoms.

It is however not clear how to decide which eigenvalues are significant. A first idea was to compute confidence intervals for a normal distribution using the χ^2-distribution. This yields an estimate which eigenvalues and corresponding eigenvectors describe directed motions that are significantly different from just noise. Figure 7.6 shows the 99% confidence interval in addition to the eigenvalues for GroEL/ES and Mm-CPN.

The confidence intervals are relatively low declaring the largest 30% of the eigenvalues as significant, which seems unrealistic. The reason for this is that the den-

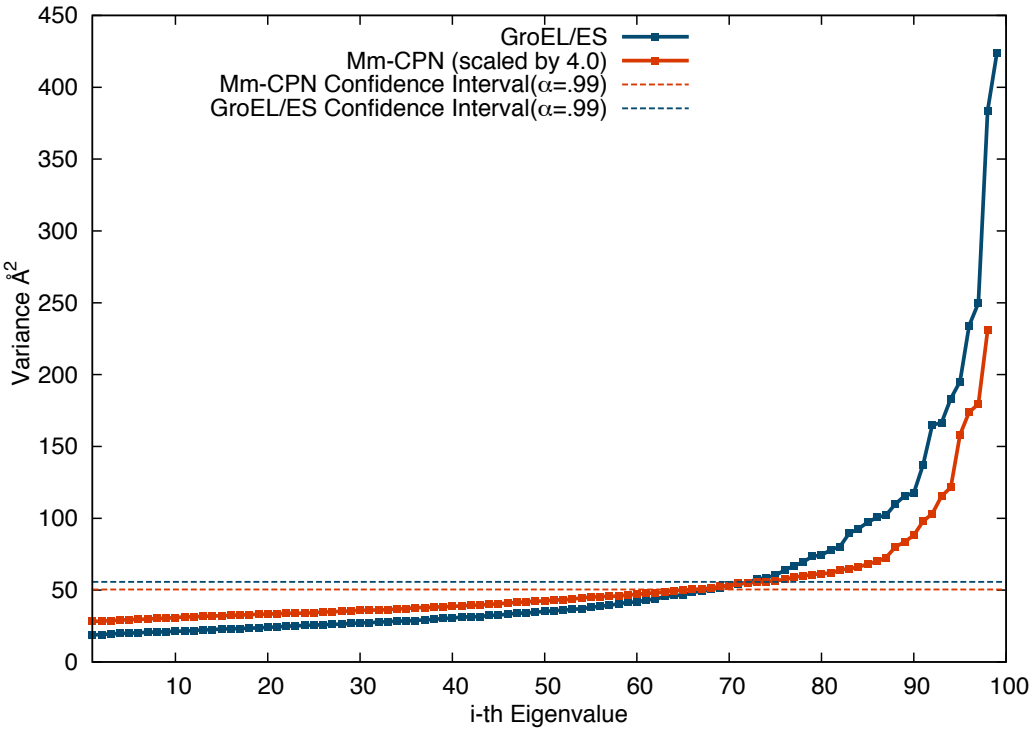

Figure 7.6: The eigenvalues of the protein structure ensembles sorted by amplitude (solid lines) and the 99% confidence interval (dashed lines) are shown in blue for Mm-CPN and orange for GroEL/ES. For GroEL/ES the two largest eigenvalues are significantly larger than the others with a large gap to the next lower eigenvalues. For Mm-CPN only the one largest eigenvalue stands out although not as significant as in the GroEL case.

sity ensemble contains information mostly for global conformational changes, but not for smaller fluctuations. Strong restraints were therefore used in the structure refinement to avoid fitting noise. These strong restraints used in the refinement make the structural ensemble narrow and allow only global collective conformational changes. The result is a small average atomistic variance. In particular most of the smaller eigenvalues will be strongly underestimated. The confidence intervals computed for these distributions are therefore too low and cannot be reliably used to define which eigenvectors describe significant motions beyond isotropic fluctuations.

Figure 7.6 shows the eigenvalues of the protein structural ensemble sorted by amplitude . The eigenvalues for GroEL/ES are much larger than the values for Mm-CPN. We calculated the 99% confidence interval for both χ^2-distributions as a guideline for significant eigenvalues, which is 55.7 $Å^2$ for GroEL/ES and 12.6 $Å^2$ for Mm-CPN. In both cases there would be more than 20 significant eigenvectors, which seems to be a lot for global principal motions.

Computing the variance using only the C_α atoms instead of all atoms yields smaller values for both the eigenvalues and the confidence interval, however the number of eigenvalues above the confidence interval is very similar to the full atom case. The reason is that the side chains were strongly restrained during the fitting and so the

sidechain and C_α motions are highly correlated. The average RMSD of the C_α was 0.15Å for Mm-CPN and 0.37Å for the complex of GroEL/ES, as also illustrated by the larger eigenvalues of GroEL/ES.
As we do not have a strict criterion for the significance of the eigenvectors we decided for conservative choice and analyzed only the eigenvectors that correspond to the two largest eigenvalues.
The eigenvectors corresponding to the largest eigenvalues are used to create trajectories to visualize the motions of the proteins and to study conformational fluctuations. The largest eigenvectors typically contain mostly collective global motions, such that the motions of entire protein domains can be analyzed. Due to the limitations of the correlation coefficient and therefore the PCA, only linear correlations can be analyzed; higher order correlations could still be in the data but are not revealed by a PCA. However, more complex correlations between of any two structural quantities such as distances, orientations, positions, etc. calculated from the ensemble can be obtained by analyzing the structural ensemble directly instead of using the PCA.

7.8 Principal Motions of GroEL/ES and Mm-CPN

GroEL/ES

The eigenvalue spectrum of GroEL/ES (Fig. 7.6, blue line) shows that the last two values are much larger than all others. We therefore chose the corresponding two eigenvectors for further analysis. In Figures 7.7 and 7.9 the eigenvectors are represented by arrows pointing from the C_α positions of the average structure into the direction of the eigenvectors.
The motion described by the first eigenvector (Fig. 7.7 D) predominantly involves the trans-ring. The apical domains of the trans-ring subunits undergo large rotational motions with the rotation axis parallel to the long axis of GroEL. Interestingly, the trans-ring apical domains also need to rotate (and finally to lift up) to bind GroES and eventually become a cis-ring in the following cycle of the chaperonin machinery. An onset of this motion seems to be already encoded in the equilibrium fluctuations observed here. Furthermore, the increased flexibility could facilitate binding of the unfolded substrate to the trans-ring. The differently rotated subunit conformations expose different epitopes which might contribute to a 'conformational selection' type of binding mode. It should be noted that the subunits show nearly identical motions as a result of the C7 symmetry that has been applied to the density map and that potential deviations of the actual protein motion from this symmetry cannot be studied here.
Another interesting part is GroES which shows a rotational motion (see Fig. 7.7 B) that is coupled to an upward shift. This motion seems to fluctuate between a tighter and weaker binding of GroES to GroEL, resembling a screw cap on a bottle.
To get more information on the principal motions it is helpful to take a look at a single chain of the subunits of GroEL. We expect to observe internal subunit motions which are not that clearly visible in the analysis of the entire complex. In Figure 7.8 (A) the first eigenvector of the cis-ring of GroEL contains a drifting

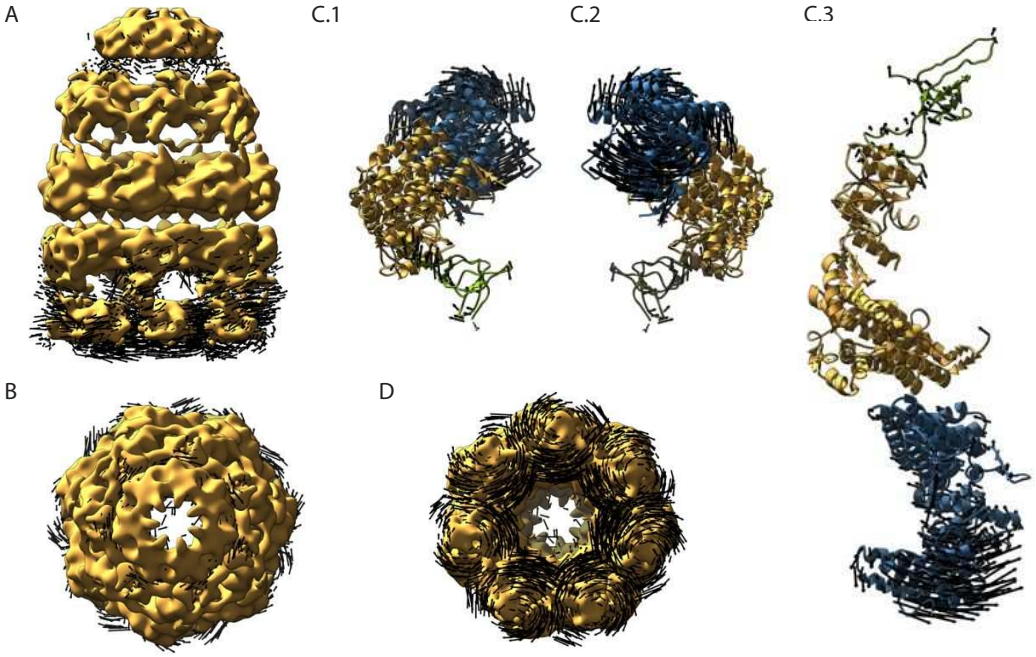

Figure 7.7: The largest eigenvector of GroEL/ES is shown as vectors superimposed on the average density of the bootstrapped ensemble (A,B,D) as well as on the atomic structure (C.1, C.2, C.3). (D) The first eigenvector shows large rotations of the individual trans-ring subunits, which dominate the entire eigenvector. This rotation is dominantly on the apical domains of the trans-ring of GroEL. Another area of interest is the GroES which performs a rotation inverse to the rotation in the trans-ring. In (C.1) (C.2) and (C.3) one asymmetric unit with a trans-ring subunit (blue), a cis-ring subunit (orange), and a GroES subunit (green) is presented in different orientations. Again, the main motion is a rotation of the apical domain of the trans-ring subunits.

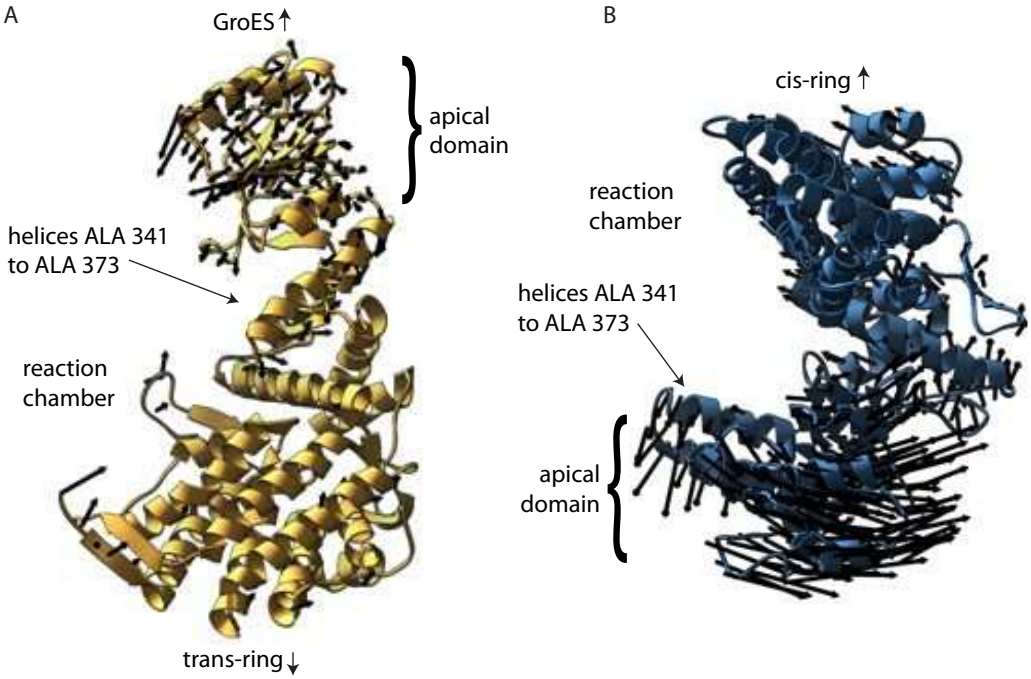

Figure 7.8: (A) presents the 1st eigenvector of a single subunit of the cis-ring of GroEL, where the major motion is a lowering of the apical domain and a small shift of the two helices from ALA 341 to ALA 373 toward the outside of the reaction chamber. (B) shows the subunit of the trans-ring. The dominant motion is in the apical domain (here at the lower end). The middle part of the apical domain flips into an more open state, while the two helices from ALA 341 to ALA 373 are rotating inward. The rotation discussed in the text is not visible in this perspective.

of the apical domain to the inside of the reaction chamber, while the GroES is performing a motion in the opposite direction (Fig. 7.7). They seem to slide along each other without losing contact.
Furthermore, the two helices from ALA 341 to ALA 373 have to undergo a large conformational change from the closed (trans) to the open (cis) state. Here these two helices flip outward, away from the reaction chamber. Together with the motion of the apical domain this fluctuation could facilitate the conformational change from the open to the closed state.
In the trans-ring the already discussed rotation of the apical domains can be amplified by the global opening of the reaction chamber. Because this motion can also be associated with the conformational change between the open and closed state, we will have to investigate further if there is a correlation between the motion of the two different rings of GroEL.
The second eigenvector shows again dominant motions in the trans-ring apical domains as well as in GroES (Fig. 7.9), as in the first eigenvector. However, in contrast to the first eigenvector the relative amplitude of the GroES rotation is much stronger than the trans-ring motion. In addition the sense of the trans-ring apical rotation is reversed with respect to the rotation of GroES. These observa-

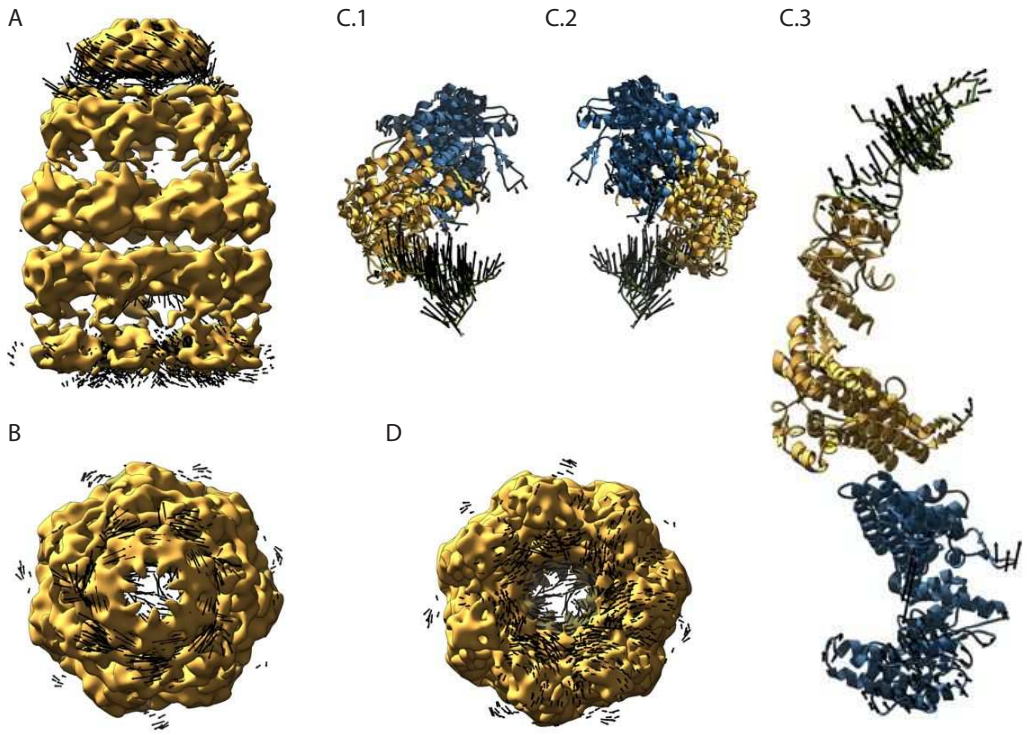

Figure 7.9: The second eigenvector shows the rotation of GroES as the dominant motion, with a smaller contribution of the trans-ring apical domain rotations (D). This is in agreement with a potential correlation of the motion of GroES with the motion of the trans-ring. In the trans-ring structure are motions on the inner side of the apical domain, which can clearly be seen in the atomistic structures of a subunit (C). These motions are directed to the center of geometry of GroEL

tions suggest that both motions are slightly correlated which hints at a potential coupling of GroES binding to a motion in the trans-ring. Such allosteric coupling has been suggested previously for GroEL/ES. However, as is discussed below, the correlation determined from the model ensemble is not significant and any potential coupling is hidden behind noise.
To further investigate potential couplings especially between the cis- and trans-ring, the correlation of specific structural quantities was calculated directly from the structural ensemble. To detect a correlation between the motions of the intermediate domain orientations between the trans- and cis-ring the orientation of the helix from GLY 344 to ILE 353 in the trans- and cis-ring was tested for correlation without any significant result. Furthermore the centers of geometry of the apical domains from GLY 192 to VAL 336 and the β-sheet of GroES from ARG 9 to ILE 11 and from LEU 84 to SER 87 were tested for correlations. In the results the correlation to any motion of the GroEL was below $f = 0.12$ and only the correlation of the comparison of the apical domains of the cis- and trans-ring gave a correlation of $r = 0.34$ that the apical domain of the trans-ring moves towards the central z-axis and the apical domain of the cis-ring moves along the z-axis.
Another approach to quantify the correlation between the different segments of GroEL/ES For this a PCA is computed separately for GroES, one of the cis-ring subunits, and one of the trans-ring subunits. The model ensemble is then projected onto the first eigenvector from each of the three PCAs. The correlation between any pair of projections was always smaller than 0.1, which means there is no detectable correlation between either ES and the cis-ring, ES and the trans-ring, or the cis-ring and the trans-ring. Comparing the correlation coefficients with the eigenvectors suggests that the motion of GroES is not strongly coupled to the motion of GroEL in the bound state.
The internal motions of a single subunit are presented in detail in Figure 7.10 for the second eigenvector. For the cis-ring subunits this second eigenvector is most dominantly a collective rotation of the apical and intermediate domains. For the trans-ring subunits the second eigenvector describes an inward tilting of the apical domain toward the reaction chamber. This trans-ring tilts seems to be a response to the stretching of the cis-ring. This is another hint for a connection between motion on the cis- and trans-ring, which cannot be detected by just using the PCA.
The next smaller eigenvectors do not seem to contain any global information and the motions are widely randomly spread over the structure.

Mm-CPN

The first eigenvector of Mm-CPN (see Fig. 7.11) is mainly one large motion in the apical domain (orange), where those domains fluctuate toward (and away from) the reaction chamber. This motion seems to be the beginning of the conformational changes that Mm-CPN needs to undergo to close the reaction chamber during uptake of the substrate and before the substrate is folded. Interestingly, an onset of this closing motion seems to be encoded already in the equilibrium fluctuations that we observe here. The full closing motion will lead to an almost spherical shape

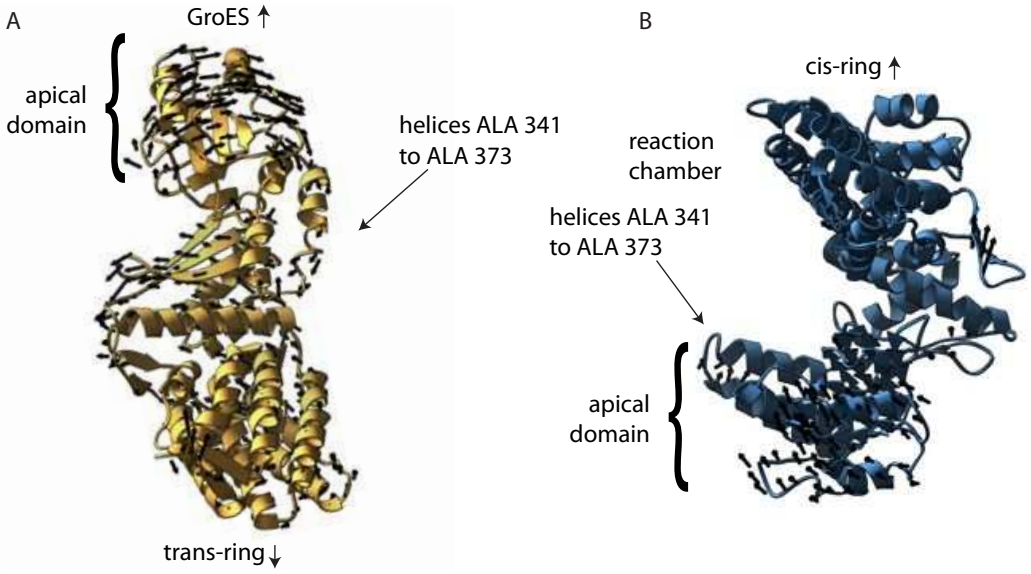

Figure 7.10: In (A) the cis-ring subunit is shown from inside the reaction chamber to more clearly demonstrate the motion corresponding to the second eigenvector. The apical domain is shifting to the right and stretches out to the neighboring subunit. The two helices from ALA 341 to ALA 373 are pivoting closer to the backside of the subunit to fill the hole opened by the rotation of the neighboring apical domain. (B) In the trans-ring the apical domain tilts into the direction of the reaction chamber.

of the entire protein, if both sides are closed. This is the only large global motion in the first eigenvector, all other motions are rather small.
Some individual residues in the stem-loop and the N- and C-termini which form a β-sheet show relatively large components in the first eigenvector. However, those motions most likely arise from the fact that the density in those regions is poorly defined which leads to larger fluctuations. These motions are not discussed here in more detail to avoid over-interpretation of these mostly random effects.
In the second eigenvector several different components can be seen in Figure 7.12: first of all there is again a dominant motion in the apical domain, it is almost the same motion as in the first eigenvector. From this observation it can be concluded that this wiggling motion of the apical domain is independent of all other motions in this eigenvector. One potential interpretation for this could be that this is a safety mechanism to make it more difficult to close the chamber only by ATP hydrolysis without a bound substrate. Closing the chamber without a substrate would unnecessarily waste ATP. Furthermore, binding of the substrate would then facilitate to close the lid.
This second eigenvector also contains a motion toward the equatorial plane, which could be part of the closing process, if the arms twist around each other to form a sphere (see Figure 7.12 A). This rotation of the outer residues of the intermediate and equatorial domain includes the stem-loop (D). From this finding we assume that the stem-loop, which is tightly connected to the neighboring subunit through an extended beta-sheet, transmits this motion to the entire ring.

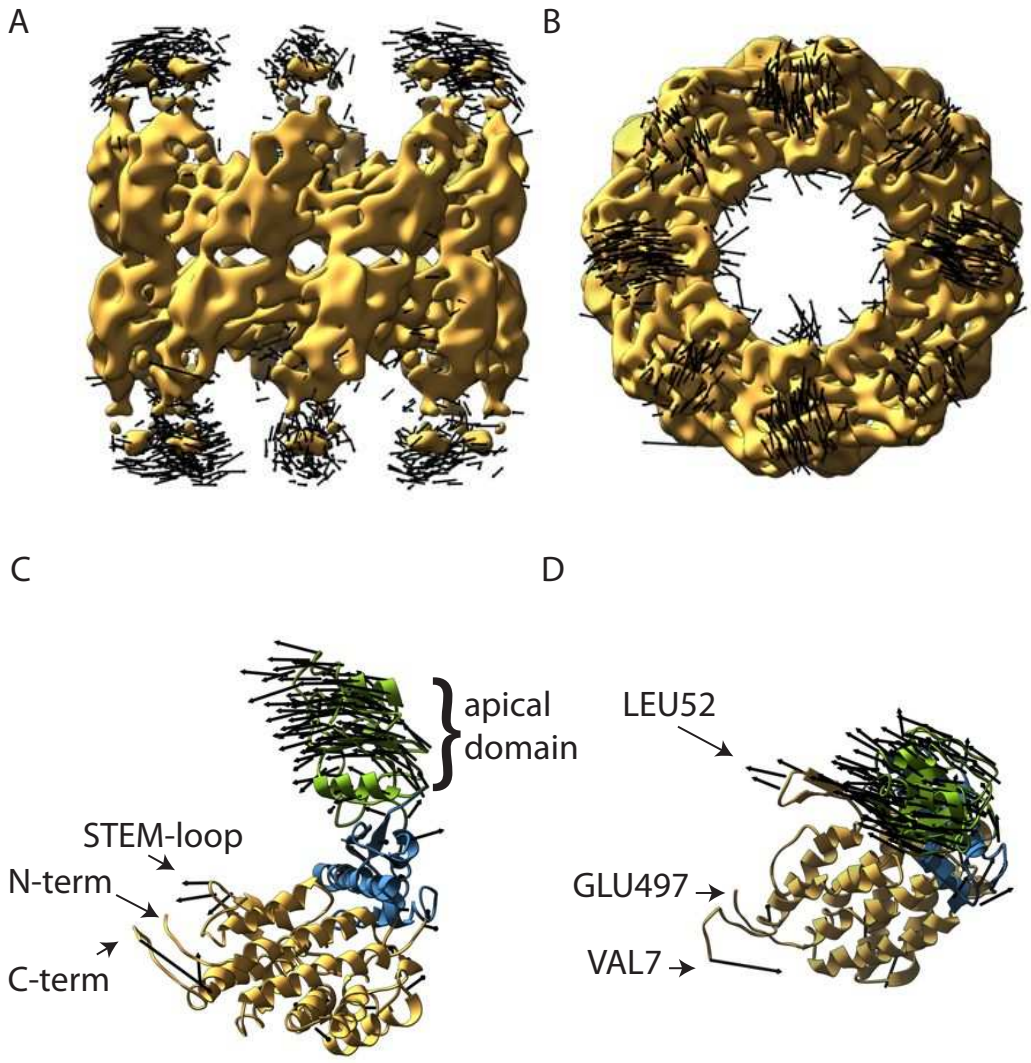

Figure 7.11: The first eigenvector on a single subunit of Mm-CPN contains mostly a motion of the apical domain toward inside of the reaction chamber, this is part of the closing of a reaction chamber of Mm-CPN. This is very clear in the projection of the vectors onto the densities of Mm-CPN in (A) side view and (B) top view. In (C) and (D) the apical domain is colored in green and the entire closing motion is performed by domain. All other motions are less dominant and seem to be randomly directed. Especially the stem-loop interaction should not be considered, because the corresponding section in the beginning and at the end of the next chain are fluctuating randomly in this eigenvector (C). There is even a large motion on the LEU52 at the middle of the loop section.

In the side view (C) an interesting area is the nucleotide binding pocket (blue) in the middle of the chain located between ASP 368 and ASP 60. The large upper helix containing ASP 368 is rotating and shifting inward in the horizontal plane, while the ASP 60 is shifting upwards. This closes the binding pocket. Because the stem-loop interaction with the neighboring subunit, the closing of the binding pocket, and the motion of the apical domain are happening within the same eigenvector one might assume that they are correlated; all these motions would be all necessary to get into the closed state. To get further information about correlated motions of separated domains on Mm-CPN the position of ALA525 (see Fig. 7.14) projected on its eigenvector in every bootstrapped density was tested for correlation with the distance between ASP 368 and ASP 60 in the ensemble. The resulting correlation is about $r = .11$ which means that a coupling between the opening or closing of the reaction chamber with the closing of the nucleotide binding pocket cannot be observed. Furthermore the motion of the apical domain is in the wrong direction: in this eigenvector the apical domains are opening while the nucleotide binding pocket is closing.
In summary the interpretation of the second eigenvector is difficult without prior knowledge about the detailed mechanism of Mm-CPN and is further complicated by the fact that many motions are already rather local conformational changes. The next smaller eigenvectors contain further information about the conformational changes of the protein but are not as dominant compared to other random changes. Exactly how many eigenvectors are significant and should analyzed in detail is still an open question that needs to be address in more detail in future studies.

7.9 Validation

In this paragraph different approaches are presented and discussed to validate the model ensemble that was obtained by fitting a crystal structure against a series of bootstrapped density maps. We discuss whether the model ensemble is in fact a valid interpretation of the variance in the experimental data set.

7.9.1 Comparison of Volumetric Variances

Due to the fact the refinement is based on the optimization of differences between the reconstructed densities and densities calculated based on atomic models, the density ensembles can be compared by calculating the correlation coefficient of corresponding density maps. At first the correlation coefficient of the average density map of the ensemble for GroEL/EL the mean map The correlation between the average density of the bootstrapped ensemble an the calculated model density ensemble after the refinement was $r = .847$ using the same spatial frequency cutoff of 9.0 Å that was to filter the density maps used during refinement process. For the ensembles of Mm-CPN a correlation of $r = .855$ was obtained with the same spatial frequency cutoff. In both cases the correlation coefficient of the average maps is markedly larger than all individual correlation coefficients reached in the refinement, which is an indicator for the statistical stability of the ensembles and justifies the assumption of a Gaussian distribution for the underlying distribution.

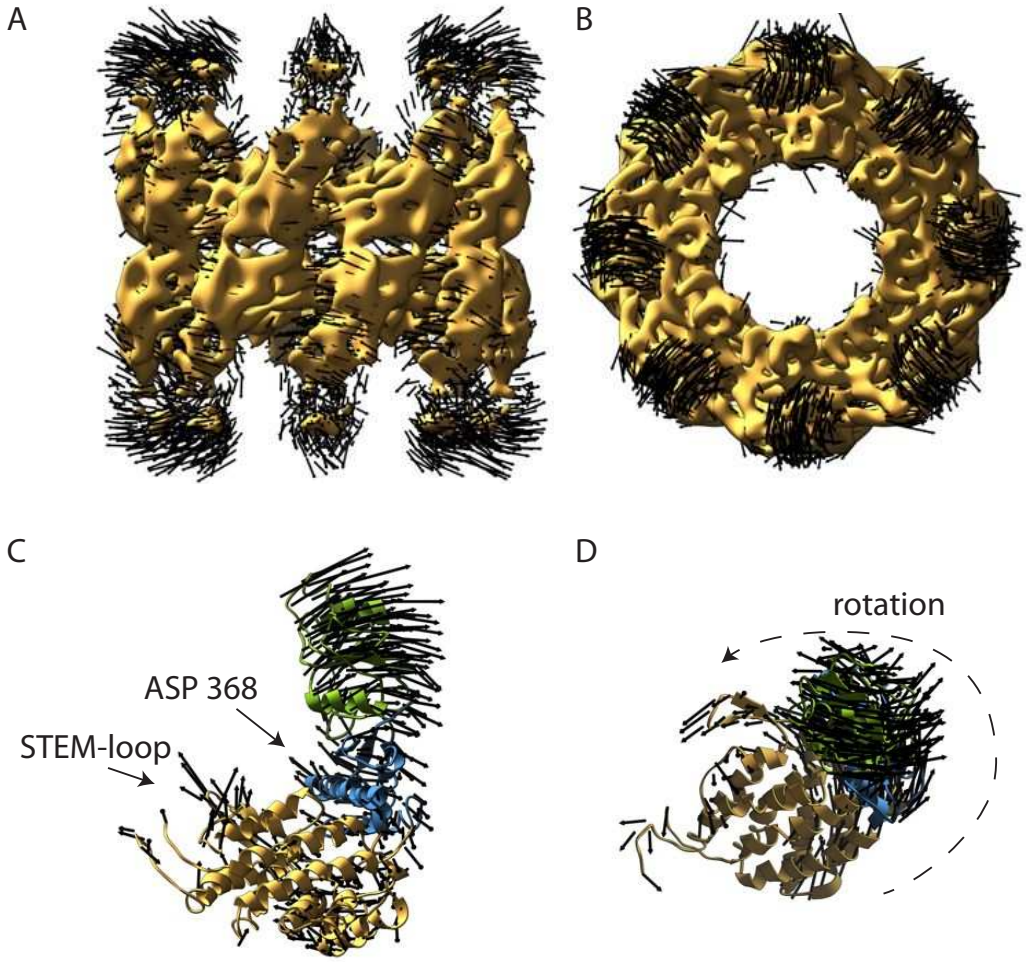

Figure 7.12: The second eigenvector of Mm-CPN shows an opening motion of the apical domain (green) and a rotation of the intermediate (blue) and equatorial (yellow) domains as dominant components. The motion on the apical domain with respect to the intermediate and equatorial domains seems to be reversed compared to the closing motion in the first eigenvector. This suggests that the motion of the apical domain is rather independent of all other motions in those eigenvectors. In the top the atomic eigenvector is superimposed onto the density of Mm-CPN. In (A) and (B) is an opening motion of the apical domains visible and dominant, in (C) and (D) it can be observed that this opening is again only on the apical domain. (A) shows the rotating motion in the outer intermediate and equatorial domains. This can be also observed in the atomic plots of a subunit (C,D). Especially in (D) this motion is visible and extends to the stem-loop as indicated by the dashed arrow. The third interesting part in (C) is the ATP binding pocket in the middle part, between the intermediate (blue) and equatorial domain (yellow), which performs a breathing motion, when the upper helix containing ASP 368 moves forward and the lower part lifts up the ASP 60 amino acid.

Measure	GroEL/ES	Mm-CPN
Mean-Map Correlation	.847	.855
CV-Map Correlation	.752	.816
Best-Fit Correlation	.835	.804
Worst-Fit Correlation	.830	.801

Table 7.3: Correlation coefficients calculated from the density maps. The average maps yield larger correlations than the best correlations from the individual refinement, which indicates that the ensemble is distributed around one conformation. The CV-map correlations are lower but still significant and it can be assumed that a large part of the conformational variance is represented by the ensemble of atomic models, while ideally random noise and other sources of variance are missing.

In the next step it is investigated how the variance of the model density ensemble compares with the bootstrapped density ensemble. Ideally, if the model ensemble is a perfect description both variance values should be very similar. In practice, a number of effects make this comparison difficult: the densities were reconstructed with symmetry which results in a radial variance map around the center of mass and large variances along the rotation axis. Furthermore, the large fluctuations of the apical domains (in particular in the Mm-CPN case) reduce the density in these regions which is not accurately accounted for in the model density maps.
To get a better feeling for the variances and to compare the results of the refinements process with the bootstrapped ensemble, we calculated the coefficient of variation (CV) maps to handle low densities in the apical domains and symmetry based artifacts which will show high variances in absence of any protein structure[12]. The CV-map is calculated from the normalized standard deviation per average density, to obtain an equal weighting of the standard deviation. This basic statistical measure of dispersion of a probability distribution handles the problem of varying contribution of atoms to the density, which can not be estimated accurately for the refinement and the resulting bias will be removed in the obtained relative standard deviation.
The CV-maps appear to be a good measure to compare the conformational variances of the ensembles. The CV maps are again filtered with a spatial frequency cutoff of 9.0Å and the correlation of both maps has been calculated. For GroEL/ES the correlation is $r = .75$, which is 0.1 lower than the correlation of the average maps but still significant, and we would expect a lower correlation because we cannot project any arbitrary conformational information on the structures. The correlation is limited by noise, resolution and the fact, that we try to avoid overfitting of the density, all this reduces the ability to fully capture the conformational variance. For the CV-densities of Mm-CPN we reached a correlation coefficient of $r = .82$, which is much closer to the correlation between the average maps. We assume that we can describe the ensembles by Gaussian distributions and estimate how well the structural ensembles correspond to the experimental sets. For both proteins the distributions are well refined and a large part of the variance is projected onto the structures.

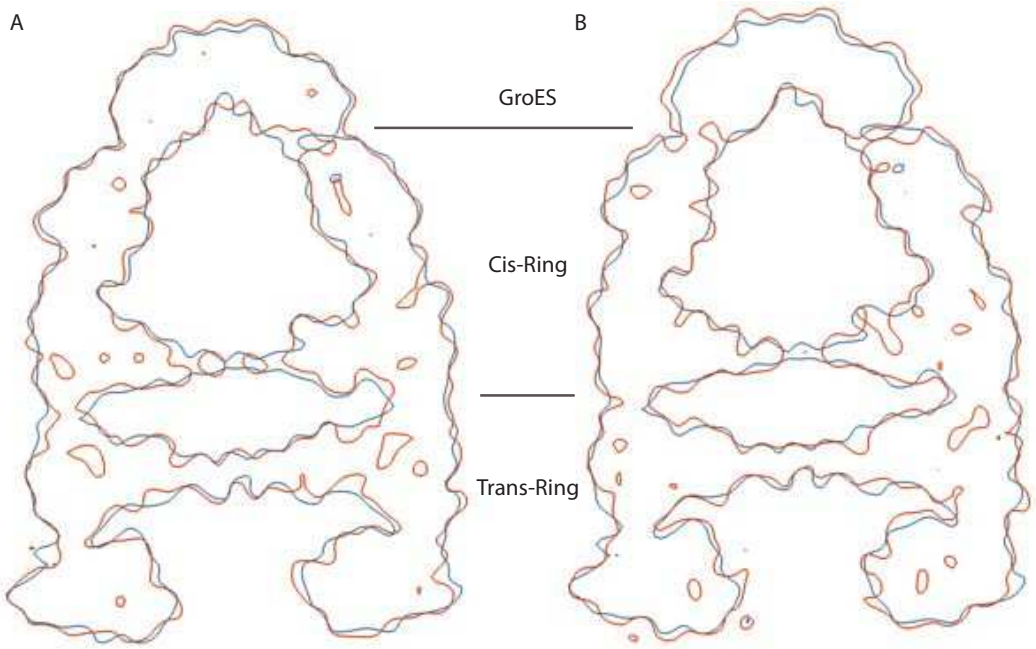

Figure 7.13: On the left side (A) a slice through GroEL average density (blue) and a shift along the first eigenvector (red). No trend can be directly extracted from this eigenvolume. Knowing the principal motions of the atomistic structures of GroEL/ES, a lifting of the GroES can be guessed, where the top region is moved upwards and the density decreases at the interface between GroES and GroEL. On the right (B) is a similar plot for the second eigenvector, where a similar motion can be seen. Here the separation of GroES and GroEL is more dominant, on the right side the volumes are separated and on the left side is still a small connection.

7.9.2 Comparison to Eigenvolumes

Analogous to the variance map it seems to be helpful to calculate the CV-eigenvolumes from the eigenvectors, by taking the square root of each eigenvector and dividing it component-wise by the average map. This helps to remove noise remaining from the reconstruction of the densities and help to reweight the variance.
Figure 7.13 A and B show slices through two states of the first and second eigenvector of GroEL/ES, respectively. The volumes (red) are calculated by adding the scaled eigenvector to the mean map. To be able to compare the data more easily the average density (blue) is superimposed. The changes in the density are rather small and not easy to be interpret as atomistic motions. Another problem is the fact that most atomistic motions are dominated by rotations, which are less dominant in the eigenvolumes. For the both ring structures of GroEL no density difference can be connected to the difference in the fitted models. Especially in the trans-ring no change comparable to the first atomistic eigenvector could be seen. The only motion in GroEL/ES that is observable is the lifting on GroES in Figure 7.13. At the top the red slice is slightly bit higher than the average structure and in the areas of contact with GroEL the density is reduced. This can be seen in both eigenvolume shifts from the average structure similar to the atomistic structure.

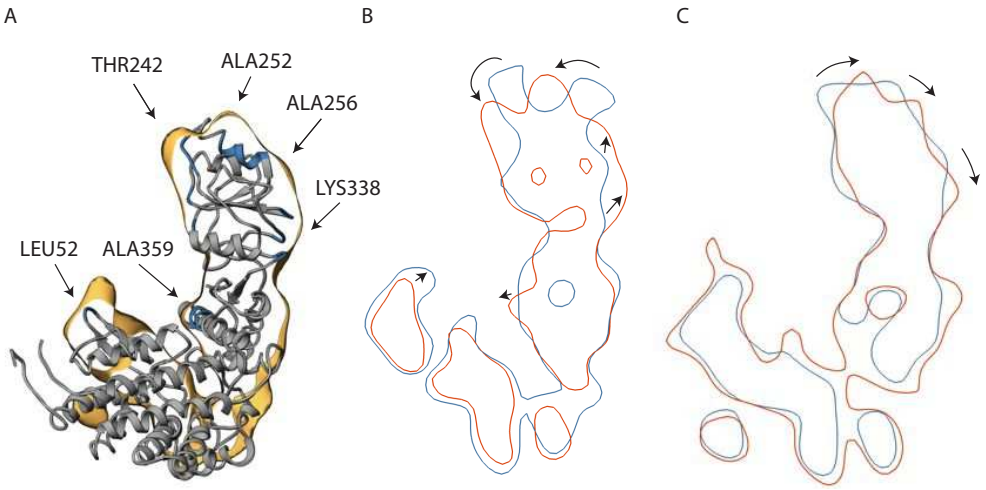

Figure 7.14: (A) shows a single assigns significant sections of the density volumes to the atomistic structure of a subunit of Mm-CPN. The orientation apical domain can be characterized by THR242, ALA252, ALA256 and LYS338. The stem-loop is represented by LEU52 and the upper section of the nucleotide binding pocket by ALA359. (B) shows slices of two volumes of the 1st eigenvector plotted. The most dominant conformational change is the bending of the apical domain toward the inner cavity of the Mm-CPN ring. Further a lift of the stem-loop can be identified at LEU52. The last interesting change is connected to a breathing of the nucleotide binding pocket below ALA359. In (C) is the similar plot of the second CV-eigenvolume, where especially the opening of the apical domain can be observed.

The atomistic and volumetric eigenvectors represent different kinds of information, which mostly depend on the type of motion. The volumetric variance is not very sensitive to rotations or translations along long extended structural elements, which in contrast contribute strongly to atomistic eigenvectors as is explained in Section 7.3. Here, the motion of GroES is large enough to be visible also in the eigenvolumes. The fact that a large component of the first CV-eigenvolume scales the density in the outer regions where no protein is present further complicates its interpretation.
For Mm-CPN it is much easier to compare the eigenvectors because the atomistic eigenvectors have the closing motion in the apical domain which also has a significant impact on the volumes. For Mm-CPN it is possible to see a very significant similarity in a slice through the subunit in the plane of the closing motion. The position of significant residues, THR242 and ALA252, are marked in the volume slice, to illustrate the eigenvector trajectories. The position of residues ALA256 and LYS338 on the outside of each subunit are also shown in this slice.
The first CV-eigenvolume is visualized by adding (blue) and subtracting (red) the scaled CV-eigenvolume to/from the average volume. Figure 7.14 B shows a slice through these two density maps representing the first eigenvolume. In this first eigenvolume the closing of the apical domain is very well defined. For the areas of all indicator residues the closing can be seen in the volumes. This is in the same direction in the volumetric and atomistic eigenvector.

Next we focus on the LEU52 at the stem-loop and ALA359 in the helix on top of the nucleotide binding pocket (Fig. 7.14 A). The stem-loop at LEU52 shows an upward motion which has been suggested to trigger the closing motion in neighboring subunits. Another conformational transition is the closing of the nucleotide binding pocket. At the position of ALA359 a conformational change is seen in the first eigenvector, where the densities move to a lower position closer to the inside. In conclusion, several motions of the atomistic ensemble can be are in agreement with the density changes along the first CV-eigenvolume.
The second CV-eigenvolume (cf. Fig. 7.14 C) also contains a motion in the apical domain. This opening motion can also be found in the second eigenvector of the atomistic second eigenvector. This can identified by using the significant positions of THR242 and ALA252. In this case the correspondence is not as clear as for the first CV-eigenvolume and all other possible motions can not assigned of motions of the atomistic eigenvectors.
In summary, the information in the CV-eigenvolumes is similar to the eigenvectors on the atomic structures but can not be interpreted as easily. The direct comparison of atomic eigenvectors and eigenvolumes is difficult as the individual motional components are mixed differently in the CV-eigenvolumes.

7.9.3 Random Ensembles

As an alternative approach to estimate the significance of the eigenvalues we generated an ensemble of models fitted against an ensemble of random density maps that were created to have the same average values and point variances as the original bootstrapped density ensemble. The ensemble of these random density maps is therefore very similar to the original bootstrapped density ensemble except for missing correlations between the density values at different grid points. The idea was that if the density variations actually report on true conformational fluctuations, these fluctuations should be encoded in the correlations. By comparing the ensembles with and without these correlations we expected the eigenvalues of the model ensemble fitted to the randomized density maps to be smaller than those of the original bootstrapped maps. The goal was to determine to what extent the eigenvectors are determined by the correlations.
To calculate the random ensembles we used the average map as the initial position and added point wise scaled normal distributed random values. This random values have been scaled by the variance map at the position to achieve the same variance map. In a next step the symmetry was applied again onto the density maps to get a similar situation compared to the original bootstrapped maps. This changes the variance of the entire ensemble and we applied a point wise correction factor based on the ratio of variances on the grid positions. Now we got correlations of more than $r = .99$ .for the mean and variance maps with identical minimum and maximum values, which proofs we generated the same mean and variance density. The intention is that global, collective and thus correlated, motions are more dominant and so all eigenvalues should be much smaller for such an ensemble, which was in fact the case for GroEL/ES. Figure 7.15 shows the eigenvalues of the random atomistic GroEL/ES structures (orange) and the eigenvalues from the refinement of the bootstrapped maps (dashed line). The largest eigenvalues of the

	GroEL/ES		Mm-CPN	
Measure	bootstrapped	random	bootstrapped	random
RMSD Å	0.44	0.31	0.21	0.34
total variance $Å^2$	5927	2786	1294	3191
largest eigenvalue $Å^2$	424	138	58	271

Table 7.4: The RMSD, the total variance and the largest eigenvalue are are shown in this Table. For all of them we can see a similar trend for the ensembles. The variation of the bootstrapped GroEL/ES is larger than the variations of the random ensemble. This is different for Mm-CPN where the fluctuations in the random ensemble are much larger than motions in the bootstrapped ensemble.

bootstrapped data are far above the larger eigenvalues of the random example. So the eigenvectors calculated from the bootstrapped ensemble vary more than the random ones.
The eigenvectors of the random GroEL/ES ensemble (see Fig. 7.16) are similar to ones obtained from the bootstrapped ensemble. The only significant difference is that the motions of the individual subunits are less symmetric . This can be explained by the way the random maps are generated: the variance map itself (without the correlation between the density grid points) encodes already a large portion of the conformational variance. Furthermore, the models were refined using strong restraints, which means only the subspace of global and collective conformational motions is accessible which leads to a significant overlap with the eigenvectors obtained from the bootstrapped data.
For Mm-CPN it is necessary to understand why the eigenvalues of the random ensemble are larger than the eigenvalues of the bootstrapped ensemble and why the eigenvectors represent a rather unlikely motion. The second question is in this case very simple to answer with the help of Fig. 7.14. The volumetric eigenvectors are in good agreement with the eigenvectors of the bootstrapped ensemble and at the same time are very different from those in the random ensemble. Since the only difference between the bootstrapped and random density maps are the correlations between density grid points, those correlations give rise to the difference between the bootstrapped and random eigenvectors. This is a good indicator that the motions determined from the bootstrapping are in fact reasonable.
To explain the large eigenvalues it is helpful to understand that the random ensemble basically consists of smeared average densities. Since there are no correlations present in the random density maps each of the random density maps is just a randomly perturbed average density. In these weakly defined tubes the rotation is always one of the most probable motions (see Section 7.3 on page 65).
The random ensemble based atomistic eigenvectors are in fact mostly rotations of each subunit as shown in Fig. 7.17. The first eigenvector (Fig. 7.17 A and B) describes a small rotation out of the reaction chamber which decreases through the outer side of the entire subsection toward the ring-ring interface. The axis for this rotation is far outside of the density. An unusual motion with a focus on the opening and closing motion of the reaction chamber.

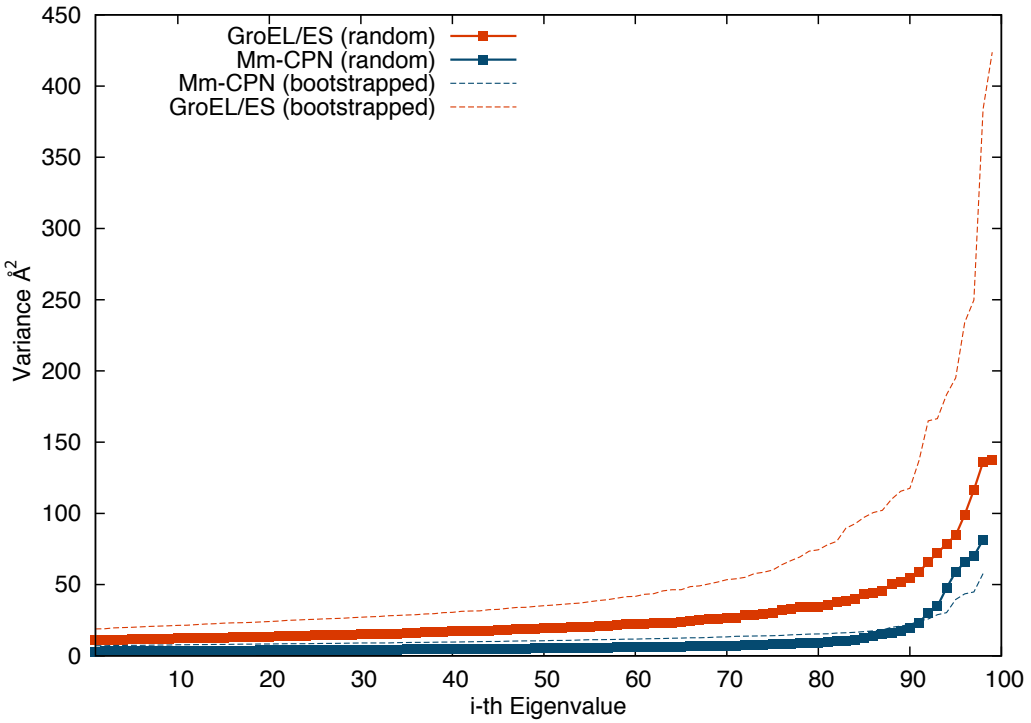

Figure 7.15: Showing the eigenvalues of structural ensembles obtained from random density ensembles (solid lines) and from the bootstrapped density ensemble (dashed lines). For GroEL/ES (red) the eigenvalues for the bootstrapped ensemble are much larger than the eigenvalues of the random ensemble. In such a case it is obvious that the largest eigenvectors of GroEL are dominant. Mm-CPN (blue) shows an opposite behavior, the random ensemble has much larger eigenvalues than the bootstrapped ensemble.

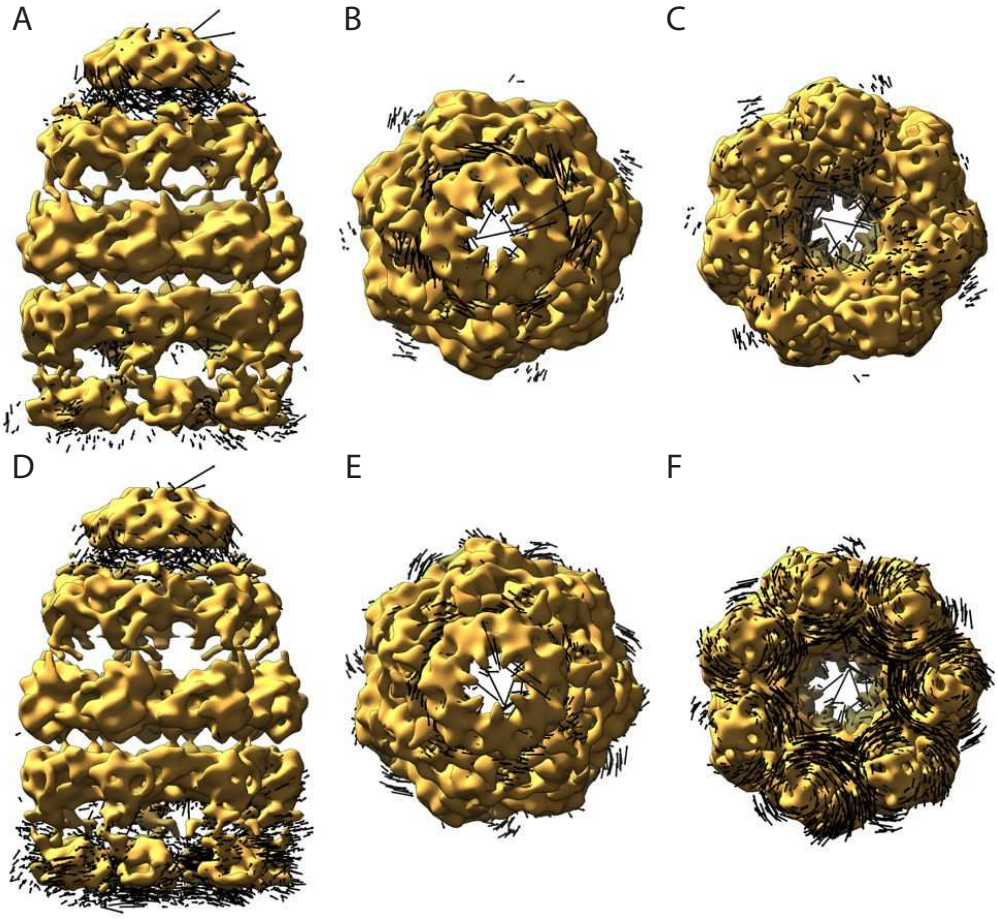

Figure 7.16: In the first row (A-C) the first eigenvector of a random GroEL/ES ensemble is superimposed on the average density map. (D-F) show the second eigenvector. The first eigenvector consists of an opening of the trans-ring A, which is not symmetric on all apical domains of the lower trans ring (C). A rotation of GroES with a downward shift is also included in this eigenvector. The second eigenvector encodes a rotation of the apical domains of the trans-ring (F) and a small rotation of GroES (D,E).

The second eigenvector (Fig. 7.17 C and D) is a rotation of only the apical domains around an axis that lies inside the apical domains. These rotational motions generate the large variance on the outside of the subunits.
To ensure the results of the refinement to be a valid and reliable set, we used the idea of the correlation matrix again. In Figure 7.18 the correlation matrix of the random ensembles is plotted as the projection on one of these ensembles. For GroEL/ES the matrix is not as ideal as in the refinement of the bootstrapped maps (cf. Fig. 7.4), in the projection of the model maps onto the target densities the refined is not always the best fitting density of the model maps, but at least among the best, which is acceptable due to the limitation of the random mixture of slightly different conformations.
Altogether this approach of a random ensemble of eigenvectors is consistent with the results of the PCA on the atomistic bootstrapped ensemble, but is not useful as a method to validate the results of a principal motion analysis.

7.10 Conclusion

We showed that the analysis of principal motions from cryo-EM data based on bootstrapping is a powerful method to determine collective conformational fluctuations of large protein complexes and to investigate their conformational changes. The approach gives valid results which has been tested by the comparison of the variance maps and the correlation matrices. In the two examples we studied, GroEL/ES and Mm-CPN, we could not determine any strong coupling between components on principal motions. Further work might be necessary to find out whether a coupling of conformational motions can be found that was hidden behind statistical noise in our analysis.
The results for Mm-CPN do not seem to be as consistent as those for GroEL/ES. The problems with Mm-CPN are already seen in the correlation matrix where GroEL shows exactly the expected behavior. We therefore think the GroEL analysis is more reliable than the results for Mm-CPN. One of the main reasons is likely that the apical domains of Mm-CPN are very flexible and the density in those regions is a mixture of very dissimilar conformations, which cannot at all be described by just a single model. In the future it might be necessary to drastically increase the number of bootstrapped maps to obtain statistically more reliable results. In addition, it might be useful to to reduce the number of particles per bootstrapped reconstruction to increase the variance in the bootstrapped density ensemble, which could help to capture the variance by the model fitting with higher significance.
The calculation of the variance of the atomic position and from this the B-factors similar to the crystallographic data yields a measure for the uncertainty of atomic positions and is an important step to validate model accuracy in cryo-EM based flexible fitting. As in crystallography the atomic variance is not necessarily isotropic, so the isotropic B-factor yields only an approximation to the actual uncertainty. The big difference between the analysis of crystallographic and cryo-EM data is that for cryo-EM data we can directly access the correlations between atomic fluctuations, which is not possible in crystallography.

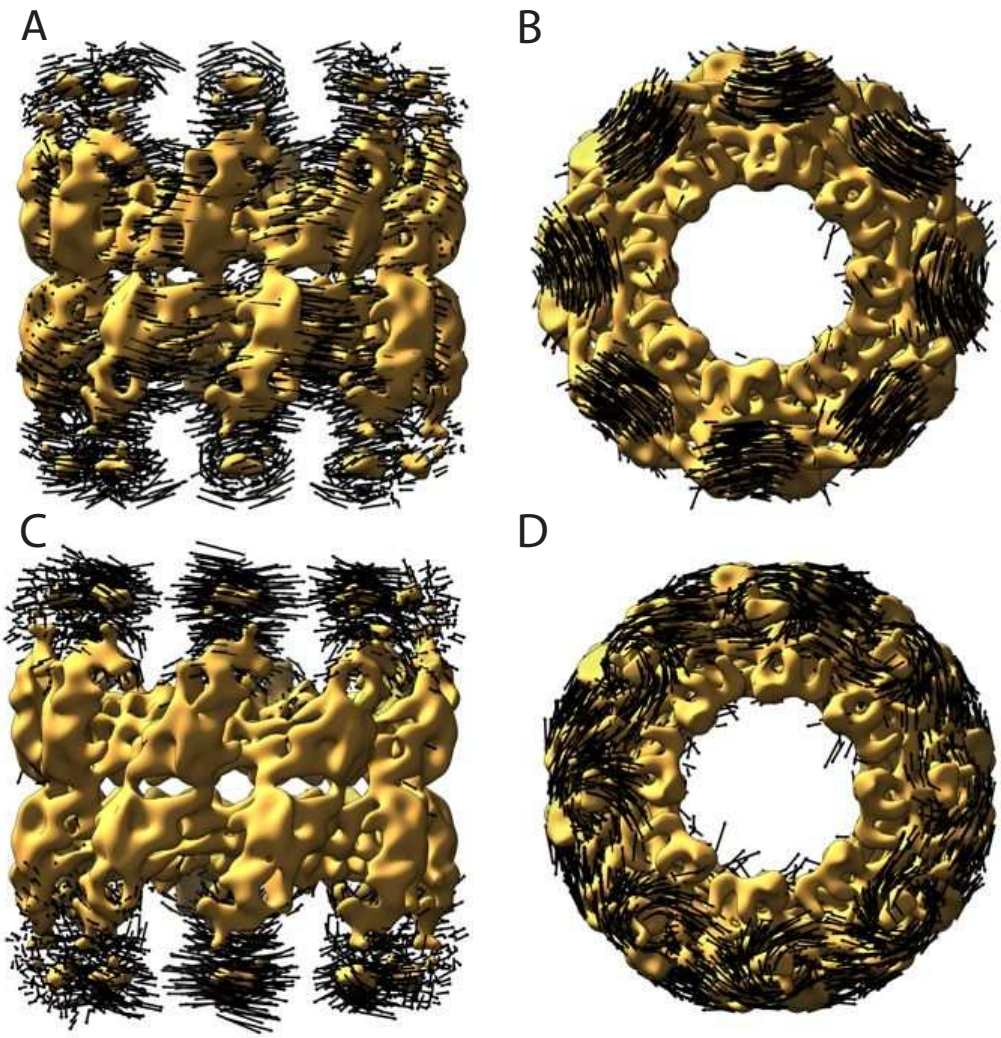

Figure 7.17: (A,B) show the first eigenvector which is a rotation of the entire subunits slightly dominated by a rotation in the apical domains. The amplitude of the rotation gets smaller from the apical through the intermediate to the equatorial domain and is only visible on the outside (A) close to the interface between the two rings. (C,D) shows the second eigenvector which is a rotation of only the apical domains.

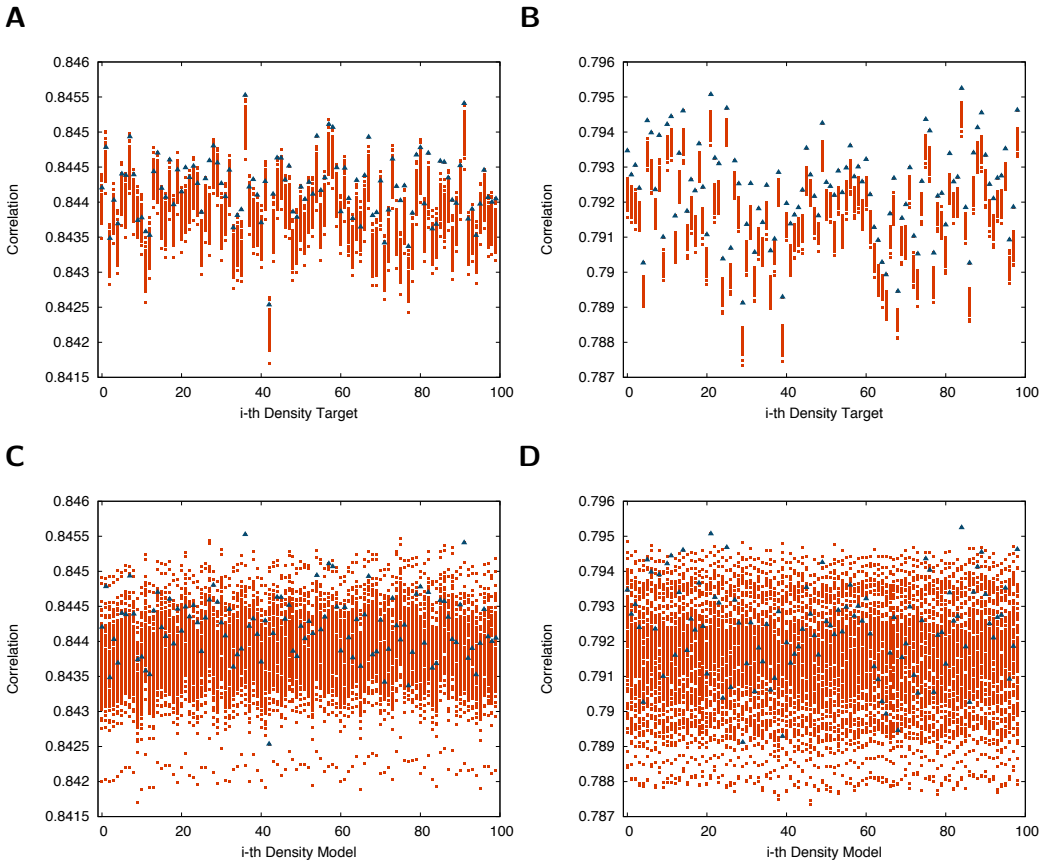

Figure 7.18: The correlation matrices of the refinement against the random maps are projected onto each ensemble. (A) and (C) shows the correlations for GroEL/ES and (B) and (D) for Mm-CPN. (A) and (B) show the correlations of each model map plotted versus each bootstrapped target. (C) and (D) show the correlations of the target density maps per model map. The identity correlations (correlation of the model map of the atomic structure refined against its target density) are marked in blue. Here the results are very clear for Mm-CPN, while the results for GroEL/ES are less perfect, but still good enough. For (C) and (D) the results are not as clear either, but this is expected for an isotropic Gaussian ensemble.

Altogether the presented technique can be an easy way to analyze conformational variances in cryo-EM data and can be easily implemented into the workflow as reconstructing the bootstrapping is relatively fast compared to the refinement of the density and the optimization of the particle orientation in the initial density reconstruction step. The fast PCA algorithm we developed allows to analyze very large data sets like ensembles of density maps or ensembles large macromolecular models.

8 Chapter

Conclusion

Single-particle is a cryo-EM is powerful method to study the structure of large biomolecules. In contrast to NMR or X-ray crystallography where ensemble averages are observed, cryo-EM collects information on single particles and at least in principle provides access to the full distribution of conformational states. This work investigated how the variance in the data set of single particle images can be interpreted in terms of macromolecular dynamics. This knowledge is important for understanding functionally relevant protein motions and for revealing the molecular mechanisms.

We investigated two protein complexes, the chaperonins GroEL/ES and Mm-CPN in collaboration with the lab of Wah Chiu (Baylor College of Medicine, TX, USA). The data were recorded by Junjie Zhang and Donghua Chen.

By using a bootstrapping approach an ensemble of density maps was generated from which the variance in the density was studied. A variance map was computed, which has been described in the literature before, and which simply visualizes the regions in the protein that are most flexible. Here we were interested in learning about correlated fluctuations. For this purpose a PCA was performed on the ensemble of volumes yielding eigenvolumes which describe the principal components of the density fluctuations. However, we faced a big challenge as standard implementations to solve the eigenvalue problem in the PCA failed here simply because of the enormous size of the data set. For the density maps studied in this work the dimensionality (the number of grid points) is in the order of 10^7, which would mean that the eigenvalue of a $10^7 \times 10^7$ needs to be computed; this is due to both CPU and memory requirements not tractable. Fortunately, the number of samples (bootstrapped density maps) is not large (order of 10^2). To solve this problem a fast sparse approach has been developed whose speed and memory requirements depend mostly on the number of samples and not on the dimensionality. This approach can be expected to be useful in a large variety of applications.

The ultimate goal of this work was to determine the dynamics of the protein machinery As was discussed, the density fluctuations do not directly translate into atomic fluctuations since amount of density changes heavily depends on the type of motion and the shape of the molecule; small motion could lead to large changes

in density, whereas large motions could lead to only small density changes. It was therefore necessary to build an ensemble of atomic models that fully captures the variance of the density ensemble. For this purpose atomic models were fitted against the bootstrapped density maps, yielding an ensemble of atomic structures. Since the chaperonin structures we studied are large ($\sim$ 60.000 atoms) and the resolution of the density maps were low ($\sim$ 10 Å) overfitting is a big problem at such low resolution. While for X-ray crystallographic refinement a cross-validation approach has been introduced more than 20 years ago, no such approach has so far been described to cross-validate refinement against cryo-EM density maps. In this work a cross-validation approach has been developed and thoroughly tested, which defines a Fourier shell as a free data set that is not used for the actual refinement but only for validating the fitted model. Future research will focus on improving the selection of this Fourier shell to optimally trade off between using as much information as possible for the refinement while still ensuring robustness of the validation measure.

For the first time large scale conformational motions of protein complexes could be determined from cryo-EM data. This opens completely new possibilities to study the conformational dynamics of very large macromolecular complexes. While we presented a variety of approaches to validate the observed principal motions, there are still open questions about the validation. A more rigorous approach to decide whether the identified principal motions are indeed significant would be desirable. A possible approach would be to assess the principal motions by comparing their projections directly to the single particle images.

Sample heterogeneity remains one of the main challenges in the analysis of cryo-EM data. Solving this problem is mostly a computational challenge and has the potential to not only improve the resolution but at the same time to also yield a picture of the conformational dynamics. While the bootstrapping is a straightforward and elegant approach its limitation is that each bootstrapped density map is still an average over a large number of particles. Future work will focus on integrating the bootstrapping with the reconstruction process and on analyzing the variance directly in terms of the single-particle images.

The Cryo-EM technique has seen tremendous improvements in resolution in the past years. Together with the analysis of dynamics this technique contributes fundamentally to understanding the mechanisms of complex macromolecular machines and will even more do so in the future.

List of Figures

List of Tables

Bibliography

[1] Collaborative Computational Project Number 4. The ccp4 suite: programs for protein crystallography. *Acta Cryst D*, 50:760–763, 1994.

[2] M. L. Baker, J. Zhang, S. J. Ludtke, and W. Chiu. Cryo-em of macromolecular assemblies at near-atomic resolution. *Nat. Protoc.*, 5:1697–1708, 2010.

[3] P. R. Baldwin and P. A. Penczek. Estimating alignment errors in sets of 2-d images. *J Struct Biol*, 150(2):211–225, May 2005.

[4] P. R. Baldwin and P. A. Penczek. The transform class in sparx and eman2. *J Struct Biol*, 157(1):250–261, Jan 2007.

[5] H. H. Barrett. Objective assessment of image quality: effects of quantum noise and object variability. *J Opt Soc Am A*, 7(7):1266–1278, Jul 1990.

[6] W. T. Baxter, A. Leith, and J. Frank. Spire: the spider reconstruction engine. *J Struct Biol*, 157(1):56–63, Jan 2007.

[7] E. Behrmann, G. Tao, D. L. Stokes, E. H. Egelman, S. Raunser, and P. A. Penczek. Real-space processing of helical filaments in sparx. *J Struct Biol*, 177(2):302–313, Feb 2012.

[8] K. Braig, P. D. Adams, and A. T. Brunger. Conformational variability in the refined structure of the chaperonin groel at 2.8 a resolution. *Nat Struct Mol Biol*, 2:1083–1094, 1995.

[9] K. Braig, Z. Otwinowski, R. Hegde, D. C. Boisvert, A. Joachimiak, A. L. Horwich, and P. B. Sigler. The crystal structure of the bacterial chaperonin groel at 2.8 a. *Nature*, 371(6498):578–586, Oct 1994.

[10] A. T. Brünger. Free r value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355:472–475, 1992.

[11] P. Chacón and W. Wriggers. Multi-resolution contour-based fitting of macromolecular structures. *J Mol Biol*, 317(3):375–384, 2002.

[12] D. H. Chen, K. Luke, J. Zhang, W. Chiu, and P. Wittung-Stafshede. 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. *J Mol Biol*, 381(3):707–717, Sep 2008.

[13] J. Z. Chen, J. Fürst, M. S. Chapman, and N. Grigorieff. Low-resolution structure refinement in electron microscopy. *J Struct Biol*, 144(1-2):144–151, Oct-Nov 2003.

[14] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. s Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[15] C. R. Crawford. Ct filtration aliasing artifacts. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 10(1), 1991.

[16] B. L. de Groot, D. M. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29(2):240–251, 1997.

[17] B. L. de Groot, D. M. F. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. C. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29:240–251, 1997.

[18] M. Delarue and P. Dumas. On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc Natl Acad Sci*, 101:6957–62, 2004.

[19] F. DiMaio, M. D. Tyka, M. L. Baker, W. Chiu, and D. Baker. Refinement of protein structures into low-resolution density maps using rosetta. *Journal of Molecular Biology*, 392:181–90, 2009.

[20] L. Ditzel, J. Löwe, D. Stock, K. O. Stetter, H. Huber, R. Huber, and S. Steinbacher. Crystal structure of the thermosome, the archaeal chaperonin and homolog of cct. *Cell*, 93(1):125–138, Apr 1998.

[21] O. Dror, K. Lasker, R. Nussinov, and H. Wolfson. Ematch: an efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-em maps of large macromolecular assemblies. *Acta Crystallographica D D*, 63:42–49, 2007.

[22] B. Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68, 1981.

[23] F. Fabiola and M. S. Chapman. Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure*, 13:389–400, 2005.

[24] F. Fabiola, A. Korostelev, and M. S. Chapman. Bias in cross-validated free r factors: mitigation of the effects of non-crystallographic symmetry. *Acta Cryst D D*, 62:227–238, 2006.

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

[26] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, 1st edition, 1996.

[27] J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith. Spider and web: processing and visualization of images in 3d electron microscopy and related fields. *J Struct Biol*, 116(1):190–199, Jan-Feb 1996.

[28] J. Fu, H. Gao, and J. Frank. Unsupervised classification of single particles by cluster tracking in multi-dimensional space. *J Struct Biol*, 157(1):226–239, Jan 2007.

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

[30] N. Grigorieff and S. C. Harrison. Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy. *Curr Opin Struct Biol*, 21(2):265–273, Apr 2011.

[31] D. A. Haley, J. Horwitz, and P. L. Stewart. The small heat-shock protein, alphab-crystallin, has a variable quaternary structure. *J Mol Biol*, 277(1):27–35, Mar 1998.

[32] K. J. Hansen. *Advances in Optical and Electron Microscopy*, volume 4, chapter The optical transfer theory of the electron microscope: fundamentalprinciples and applications, pages 1 – 84. Academic Press, New York, 1971.

[33] M. H. Hansen, W. N. Hurwitz, and W. G. Madow. *Sample survey methods and theory*. Wikey, New York, 1953.

[34] G. Harauz and M. Heel van. Exact filters for general geometry three dimensional reconstruction. *Optik*, 73:146–156, 1986.

[35] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[36] P. W. Hawkes and E. Kasper. *Principles of Electron Optics: Wave Optics*. Academic Press, London, 1994.

[37] S. Helgason. *The Radon Transform*. Birkhäuser, 2nd edition, 1980.

[38] G. Herman and A. Lent. Iterative reconstruction algorithms. *Comput. Biol. Med.*, 6:273–294, 1976.

[39] K. Hinsen, N. Reuter, J. Navaza, D. L. Stokes, and Lacapére J-J. Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum ca-atpase. *Biophysical journal*, 88:818–27, 2005.

[40] M. Hohn, G. Tang, G. Goodyear, P. R. Baldwin, Z. Huang, P. A. Penczek, C. Yang, R. M. Glaeser, P. D. Adams, and S. J. Ludtke. Sparx, a new environment for cryo-em image processing. *J Struct Biol*, 157(1):47–55, Jan 2007.

[41] A. Jack and M. Levitt. Refinement of large structures by simul- taneous minimization of energy and r factor. *Acta Crystallogr.*, A34(931–935), 1978.

[42] A. C. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging*. IEEE Press, New York, 1988.

[43] K. Karhunen. Über lineare methoden in der wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys*, 37:1–79, 1947.

[44] E. J. Kirkland. *Advanced Computing in Electron Microscopy*. Springer, 2nd edition, 2010.

[45] G. J. Kleywegt and T. Jones a. xdlmapman and xdldataman - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Cryst D*, 52:826–828, 1996.

[46] M. Knoll and E. Ruska. Das elektronenmikroskop. *Z. für Physik*, 78:318–339, 1932.

[47] A. R. Kusmierczyk and J. Martin. Nucleotide-dependent protein folding in the type ii chaperonin from the mesophilic archaeon methanococcus maripaludis. *Biochem J*, 371(Pt 3):669–673, May 2003.

[48] K. Lasker, M. Topf, A. Sali, and H. J. Wolfson. Inferential optimization for simultaneous fitting of multiple components into a cryoem map of their assembly. *J Mol Biol*, 388:180–194, 2009.

[49] K. H. Lee, H. S. Kim, H. S. Jeong, and Y. S. Lee. Chaperonin groesl mediates the protein folding of human liver mitochondrial aldehyde dehydrogenase in escherichia coli. *Biochem Biophys Res Commun*, 298(2):216–224, Oct 2002.

[50] W. Liu, N. Boisset, and J. Frank. Estimation of variance distribution in three-dimensional reconstruction. ii. applications. *J Opt Soc Am A Opt Image Sci Vis*, 12(12):2628–2635, Dec 1995.

[51] W. Liu and J. Frank. Estimation of variance distribution in three-dimensional reconstruction. i. theory. *J Opt Soc Am A Opt Image Sci Vis*, 12(12):2615–2627, Dec 1995.

[52] M. Loeve. *Probability theory*, volume 2 of *Graduate Texts in Mathematics*. Springer-Verlag, 4 edition, 1978.

[53] S. J. Ludtke, P. R. Baldwin, and W. Chiu. Eman: semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol*, 128:82–97, 1999.

[54] S. J. Ludtke, P. R. Baldwin, and W. Chiu. Eman: semiautomated software for high-resolution single-particle reconstructions. *Journal of Structural Biology*, 128:82–97, 1999.

[55] S. J. Ludtke, D. H. Chen, J. L. Song, D. T. Chuang, and W. Chiu. *Structure*, 12(7):1129 – 1136, 2004.

[56] S. J. Ludtke, J. Jakana, J. L. Song, D. T. Chuang, and W. Chiu. A 11.5 a single particle reconstruction of groel using eman. *J Mol Biol*, 314:253–262, 2001.

[57] J. A. Mindell and N. Grigorieff. Accurate determination of local defocus and specimen tilt in electron microscopy. *J Struct Biol*, 142(3):334–47, Jun 2003.

[58] G. Moliere. Theorie der streuung schneller gelandener teilchen i. einzelstreuung am abgeschirmten coulomb-field. *Z fur Naturforscher*, 2:133 – 145, 1947.

[59] M. Orzechowski and F. Tama. Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys J*, 95:5692–5705, 2008.

[60] P. A. Penczek. Variance in three-dimensional reconstuctions from projections. *Proceedings of the IEEE*, 2002.

[61] P. A. Penczek. Fundamentals of three-dimensional reconstruction from projections. *Methods Enzymol*, 482:1–33, 2010.

[62] P. A. Penczek. Image restoration in cryo-electron microscopy. *Methods Enzymol*, 482:35–72, 2010.

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

[64] P. A. Penczek, C. Yang, J. Frank, and C. M. Spahn. Estimation of variance in single-particle reconstruction using the bootstrap technique. *J Struct Biol*, 154(2):168–183, May 2006.

[65] P. A. Penczek, J. Zhu, and J. Frank. A common-lines based method for determining orientations for n $>$ 3 particle projections simultaniously. *Ultramicroscopy*, 63:205–218, 1996.

[66] P. A. Penczek, J. Zhu, and J. Frank. Three dimanesinal reconstructionwith contrast transfer function from defocus series. *Scanning Microscopy*, 11:147–154, 1997.

[67] P. A. Penczek, J. Zhu, R. Schröder, and J. Frank. Three dimensional reconstruction wit contrast transfer compensation from defocus series. *Scanning Microscopy*, 11:147–154, 1997.

[68] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. Ucsf chimera–a visualization system for exploratory research and analysis. *J Comp Chem*, 25:1605–12, 2004.

[69] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.

[70] J. Radon. Über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Berichte über die Verhandlungen der Sächsische Akademie der Wissenschaften*, 69:262 – 277, 1917.

[71] A. M. Roseman. Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Cryst D*, 56:1332–1340, 2000.

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

[73] M. G. Rossmann. Fitting atomic models into electron-microscopy maps. *Acta Cryst D D*, 65:1341–1349, 2000.

[74] M. G. Rossmann, M. C. Morais, P. G. Leiman, and W. Zhang. Combining x-ray crystallography and electron microscopy. *Structure*, 13:355–62, 2005.

[75] C. Sachse, J. Z. Chen, P. D. Coureux, M. E. Stroupe, M. Fändrich, and N. Grigorieff. High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *J Mol Biol*, 371(3):812–835, Aug 2007.

[76] H. Sasaki, M. van Heel, E. Zeitler, and T. Suzuki. Fine structure of mitochondrial helical filaments revealed by computer image analyses. *J Electron Microsc (Tokyo)*, 39(5):388–395, 1990.

[77] O. Scherzer. The theoretical resolution limit of the electron microscope. *Journal of Applied Physics*, 20:20 – 29, 1949.

[78] G. F. Schröder, A. T. Brunger, and M. Levitt. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure*, 15:1630–41, 2007.

[79] G. F. Schröder, A. T. Brunger, and M. Levitt. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure*, 15:1630–1641, 2007.

[80] G. F. Schröder, M. Levitt, and A. T. Brunger. Super-resolution biomolecular crystallography with low-resolution data. *Nature*, 464:1218–1222, 2010.

[81] M.F. Smith and J.P. Langmore. Quantitaion of molecular densities by cryoelectron microscopy-determination of the radial density distribution of tobacco mosaic-virus. *J. Mol. Bio.*, 226:763–774, 1992.

[82] C. M. Spahn and P. A. Penczek. Exploring conformational modes of macromolecular assemblies by multiparticle cryo-em. *Curr Opin Struct Biol*, 19(5):623–631, Oct 2009.

[83] S. M. Stagg, G. C. Lander, J. Quispe, N. R. Voss, A. Cheng, H. Bradlow, S. Bradlow, B. Carragher, and C. S. Potter. A test-bed for optimizing high-resolution single particle reconstructions. *J Struct Biol*, 163:29–39, 2008.

[84] A. Stewart and N. Grigorieff. Noise bias in the refinement of structures derived from single particles. *Ultramicroscopy*, 102:67–84, 2004.

[85] K. Suhre, J. Navaza, Sanejou, and Y-henri. Norma: a tool for flexible fitting of high-resolution protein structures into. *Acta Cryst D D*, 62:1098–1100, 2006.

[86] F. Tama, O. Miyashita, and C. L. Brooks. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J Mol Biol*, 337:985–99, 2004.

[87] G. Tang, L.i Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke. Eman2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38 – 46, 2007.

[88] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali. Protein structure fitting and refinement guided by cryo-em density. *Structure*, 16:295–307, 2008.

[89] C. Toyoshima and N. Unwin. Contrast transfer for frozen-hydrated specimens. *Ultramicroscopy*, 25:279–291, 1988.

[90] L. G. Trabuco, E. Villa, K. Mitra, J. Frank, and K. Schulten. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, 16:673–683, 2008.

[91] L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, and K. Schulten. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and x-ray crystallography. *Methods*, 49(2):174–180, Oct 2009.

[92] M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt, and M. Schatz. A new generation of the imagic image processing system. *J Struct Biol*, 116(1):17–24, Jan-Feb 1996.

[93] A. J. C. Wilson. Largest likely values for the reliability index. *Acta Crystallographica*, 3:397–398, 1950.

[94] M. Wolf, D. J. DeRosier, and N. Grigorieff. Ewald sphere correction for single-particle electron microscopy. *Ultramicroscopy*, 106(4-5):376–382, Mar 2006.

[95] W.Owen and Saxton. Semper: Distortion compensation, selective averaging, 3-d reconstruction, and transfer function correction in a highly programmable system. *Journal of Structural Biology*, 116(1):230 – 236, 1996.

[96] W. Wriggers and S. Birmanns. Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. *J Struct Biol*, 133(2-3):193–202, 2001.

[97] W. Wriggers, R. A. Milligan, and J. A. Mccammon. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J Struct Biol*, 125:185–195, 1999.

[98] Z. Xu, A. L. Horwich, and P. B. Sigler. The crystal structure of the asymmetric groel-groes-(adp)7 chaperonin complex. *Nature*, 388(6644):741–750, Aug 1997.

[99] C. Yang, E. G. Ng, and P. A. Penczek. Unified 3-d structure and projection orientation refinement using quasi-newton algorithm. *J Struct Biol*, 149(1):53–64, Jan 2005.

[100] Z. Yang, J. Fang, J. Chittuluru, F. J. Asturias, and P. A. Penczek. Iterative stable alignment and clustering of 2d transmission electron microscope images. *Structure*, 20(2):237–247, Feb 2012.

[101] Z. Yang and P. A. Penczek. Cryo-em image alignment based on nonuniform fast fourier transform. *Ultramicroscopy*, 108(9):959–969, Aug 2008.

[102] J. Zhang, M. L. Baker, G. F. Schröder, N. R. Douglas, S. Reissmann, J. Jakana, M. Dougherty, C. J. Fu, M. Levitt, S. J. Ludtke, J. Frydman, and W. Chiu. Mechanism of folding chamber closure in a group ii chaperonin. *Nature*, 463(7279):379–383, Jan 2010.

[103] J. Zhang, B. Ma, F. DiMaio, N. R. Douglas, L. A. Joachimiak, D. Baker, J. Frydman, M. Levitt, and W. Chiu. Cryo-em structure of a group ii chaperonin in the prehydrolysis atp-bound state leading to lid closure. *Structure*, 19(5):633–639, May 2011.

[104] W. Zhang, M. Kimmel, C. M. Spahn, and P. A. Penczek. Heterogeneity of large macromolecular complexes revealed by 3d cryo-em variance analysis. *Structure*, 16(12):1770–1776, Dec 2008.

[105] X. Zhang, E. Settembre, C. Xu, P. R. Dormitzer, R. Bellamy, S. C. Harrison, and N. Grigorieff. Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc Natl Acad Sci U S A*, 105(6):1867–1872, Feb 2008.

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