Image analysis methods development for in vitro and in situ cryo-electron tomography studies of conformational variability of biomolecular complexes : Case of nucleosome structural and dynamics studies

Mohamad Harastani

DOCTORAL THESIS OF SORBONNE UNIVERSITY

Doctoral school of Computer science, Communications, and Electronics (ED-130)

Doctoral Program Interfaces for the Living

Presented by

Mohamad HARASTANI

To obtain

Doctoral degree of Sorbonne University

In the field of Bioimage informatics

Thesis subject:

IMAGE ANALYSIS METHODS DEVELOPMENT FOR IN VITRO AND IN SITU CRYO-ELECTRON TOMOGRAPHY STUDIES OF CONFORMATIONAL VARIABILITY OF BIOMOLECULAR COMPLEXES: CASE OF NUCLEOSOME STRUCTURAL AND DYNAMICS STUDIES

Presented and defended publicly on October 4th, 2022

in front of a jury composed of:

Pr Alain TROUVÉ (ENS Paris-Saclay)	Reporter
Dr Patrick SCHULTZ (DR1 CNRS, IGBMC)	Reporter
Pr Friedrich FÖRSTER (Utrecht University)	Examiner
Dr Charles KERVRANN (DR1 Inria Rennes)	Examiner
Dr Niels VOLKMANN (DR Institut Pasteur)	Examiner
Dr Amélie LEFORESTIER (DR2 CNRS, LPS)	Thesis co-director
Dr Slavica JONIC (DR2 CNRS, IMPMC)	Thesis director

Thesis prepared at IMPMC-UMR 7590 (Sorbonne University) and LPS-UMR 8502 (Paris-Saclay University).

Dedication

To my family

Acknowledgment

I would like to thank all of:

My family, who has supported me financially and emotionally in the hard times. In particular, my mother, Amina Jarikji, and my father, Hassan Harastani. My sisters, Rima, Rania, and Ghada, for their support and motivation.

My Ph.D. supervisors and their teams, Slavica Jonic and Amélie Leforestier, Ilyes Hamitouche, Remi Vuillemot, Fatima Taiki, and Kahina Vertchik. Thanks to my collaborators at IGBMC, Mikhail Eltsov, and his student Fadwa Fatmaoui. And my collaborators at I2PC Madrid, the Scipion and Xmipp teams, Carlos Oscar Sorzano, Jose Maria Carazo, Pablo Conesa, David Herreros, David Strelak, and Jorge Jiménez, James Krieger and Jose Luis Vilas.

The IPV doctoral program of Sorbonne University for funding my thesis and the ANR project CRYOCHROM for funding parts of experimental works, open-access publications, and conference participation.

Those that helped me during my second master's degree, the people and associations that helped me pay my rent and get familiar with building a career in France, particularly Baydir Berrahal, Rima de Sahb, Kodiko, and EUF.

My sincere friends, I cannot list them all, but here are some that supported and motivated me through my Ph.D. journey. Ammar Mansour, Ramez Al-Hamwi, Yazan Al-Omari, Mahmoud Nazzal, Ramin & Nooshin Bakhshi, Liaisian Abdrakhmanova, Pouya Bolourchi, Faegheah Yaganli, and Tarek Abo Hassoun.

My previous supervisors, teachers, and mentors, particularly Suna Bolat, Mustafa & Resime Uyguroglu, Amine Nait-Ali, Hassan Demirel, Osman Kukrer and Runyi Yu.

This work would have never been done without the supporting people above and the many others I missed. I want to thank my mentors again, especially my Ph.D. supervisors Slavica Jonic and Amélie Leforestier, for their patience over the years and for helping me grow as a person and a scientist.

Sincerely, Mohamad

Summary

Preface

My thesis focuses on developing computational methods for biomolecular conformational variability analysis in cryo electron tomography (cryo-ET) and their use for the analysis of nucleosome conformational variability *in situ*. It is an interdisciplinary thesis between computer science and structural biology. I came to this field after studying electrical and electronics engineering and two master's programs in science and technology. During my second master's program, I completed a six-months internship at the IMPMC (Sorbonne University, Paris) in 2019. This internship was under the supervision of Dr. Slavica Jonic. My internship was to extend a method called HEMNMA to cryo-ET. HEMNMA was developed in the group of Dr. Jonic and is used for analyzing single-particle analysis (SPA) data in terms of biomolecular conformational variability. During my internship, we established a collaboration with Dr. Amélie Leforestier (LPS, University Paris-Saclay, Orsay) and Dr. Mikhail Eltsov (currently at IGBMC, Strasbourg). They were interested in our unique method for analyzing their tomographic data of nucleosomes *in situ*. My internship led to promising preliminary results. Dr. Jonic proposed a continuation of my internship for a Ph.D. under her supervision, co-supervised by Dr. Leforestier, and in collaboration with Dr. Eltsov in an interdisciplinary doctoral program at Sorbonne University called Interfaces for the Living (the French name is Interface pour le Vivant, abbreviated IPV). I applied for this project, obtained funding for a Ph.D. through a competition, and started my Ph.D. in October 2019.

During my Ph.D., my primary host lab was IMPMC, where I worked on methods development with Dr. Jonic's group. Also, I spent considerable time at LPS working with Dr. Leforestier's team, and I was involved in biological sample preparation (freezing, cryo-sectioning, observations) as well as data analysis and interpretation. My Ph.D. also took part in an interdisciplinary project called CRYOCHROM, funded by the French National Research Agency (abbreviated as ANR in French), directed by Dr. Leforestier, and in partnership with the group of Dr. Jonic, Dr. Eltsov, and Dr. Victor (specialist of chromatin modelization). CRYOCHROM investigates the conformations and distributions of nucleosomes in order to understand how chromatin functions. Although it is not the main focus of my Ph.D., I have gathered as much information as possible regarding the many stages of biological sample preparation and data acquisition and their associated characteristics and challenges. My participation in sample preparation and data collecting sessions is therefore documented in this manuscript.

Introduction

Biomolecules exhibit different forms of variability. Different copies of the same biomolecular complex can differ while interacting with their environment. A biomolecule may bind a substrate while another does not or binds another. They can be chemically modified when interacting with proteins. However, one notable source of variability is a gradual change of the conformation of biomolecules, commonly referred to as continuous conformational variability. My thesis was primarily concerned with developing image processing algorithms for assessing continuous conformational variability acquired in cryogenic Electron Tomography (cryo-ET) data.

Assume having some biomolecules trapped in a liquid medium or perhaps their mother cell. Assume that some of these biomolecules are copies of the same complex, and you are interested in visualizing them to understand their structure and dynamics. Cryogenic Electron Microscopy (cryo-EM) offers a way to observe these biomolecules, more precisely, after vitrifying (freezing) them. Imagine these biomolecules moving around, changing conformation elastically, and interacting with their surroundings before they freeze. Suddenly, freeze the whole. You will have a snapshot of these biomolecules at that pre-freezing moment. This snapshot is the kind of data you would expect from cryo-EM/ET, which can be in the form of an image called a micrograph in the case of cryo-EM 2D imaging or a volume called a tomogram in the case of cryo-ET. Cryo-ET is particularly useful for visualizing biomolecules *in situ*, thanks to its 3D data that helps disentangle the crowded cellular environment.

A micrograph or tomogram may contain many of these copies at different locations and orientations, besides their varying forms. These copies can be isolated into sub-images (called single particle images) or subvolumes (called subtomograms). For obtaining high-resolution 3D models for the biomolecule under study, single particle images or subtomograms should be sorted in such a way to get an average 3D structure. For a long time, specimen heterogeneity had a negative connotation and was only seen as a factor limiting the resolution of 3D reconstructions. However, research in the last decade has shown that identifying conformational transitions from heterogeneous samples can help study molecular mechanisms in action. One way to sort this biomolecular variability is discrete classification. However, classification is effective for discrete cases of variability (such as substrate binding) but not so much for continuous flexibility.

To understand this flexibility, we need methods to analyze a continuum of conformational states. Before this thesis, few methods considered continuous conformational changes explicitly for cryo-EM 2D data, but no method existed for cryo-ET 3D data. Subtomograms are noisy, low contrast, suffer from spacial anisotropies due to the image formation and the type of acquisition in the electron microscope, and thus are very difficult to analyze individually.

This thesis presents the first two methods that address the continuous conformational variability of biomolecules in cryo-ET data, HEMNMA-3D and TomoFlow. HEMNMA-3D analyzes experimental data with motion directions simulated by Normal Mode Analysis and allows the discovery of an extensive range of biomolecular motions hidden in the data. However, HEMNMA-3D depends on this prior (simulated motion directions), making it prone to misinterpretation and bias when misused. TomoFlow extracts movements from the data without prior information using a computer vision technique called the Optical Flow. Therefore, it is less prone to misuse. However, when it encounters large motion magnitudes, it results in a smooth and downscaled version of the actual biomolecular motion. Although the mathematical models used by HEMNMA-3D and TomoFlow differ, both can explore biomolecular conformational landscapes and are superior to classification into discrete classes. In this thesis, I systematically validate HEMNMA-3D and TomoFlow on synthetic datasets. Also, I demonstrate the utility of these two methods on experimental cryo-ET data of nucleosome conformational variability *in situ*, taking part in an ongoing study of nucleosome in cells. The two methods show coherent results, shedding insight into the conformational variability of nucleosomes, in line with previous visual and theoretical analyses of nucleosome conformations. I demonstrate that these methods produce valuable results with especially challenging *in situ* data, nucleosomes. They are thus also expected to be useful for conformational studies of other biomolecular complexes *in vitro* and *in situ*.

HEMNMA-3D and TomoFlow software are now publicly available as part of the cryo-ET data processing pipeline of the open-source software package ContinuousFlex, which is now a plugin of Scipion software and its backend software Xmipp that is widely used in the field. Throughout my Ph.D., I helped develop and maintain these three software packages, particularly ContinuousFlex.

This thesis manuscript is organized as follows:

Chapter 1 summarizes the essential instruments, technologies, and methods in electron microscopy used to collect data from biological samples, focusing on cryo-ET.

Chapter 2 covers biomolecular conformational variability captured by cryo-ET and reviews approaches other than those conducted in this thesis for dealing with it.

Chapter 3 introduces the nucleosome case study, summarizing previous findings and theoretical nucleosome conformation predictions. It explains how *in situ* nucleosome cryo-ET data are collected and what was uncovered from them prior to the start of this thesis.

Chapter 4 covers the mathematical background required for a thorough understanding of the methods developed in this thesis (i.e., HEMNMA-3D and TomoFlow), particularly the principles of Normal Mode Analysis and Optical Flow. Also, this chapter presents the cryo-ET subtomogram dataset of nucleosomes *in situ* that was analyzed in this thesis, highlighting practical considerations we followed for obtaining and pre-processing this dataset.

HEMNMA-3D and TomoFlow are explained in Chapters 5 and 6, respectively. These chapters describe their methodologies and validation using synthetic datasets and present the results produced by applying them to the *in situ* nucleosome dataset presented in Chapter 4. Chapter 5 also compares HEMNMA-3D to traditional data analysis methods for biomolecular structure and dynamics determination in cryo-ET.

Chapter 7 summarizes the software contribution of this thesis.

I end this manuscript with Chapter 8 on discussions, conclusions, and possible future works drawn from this thesis.

Chapter 1. Cryogenic electron microscopy and approaches for data collection of biological samples

A microscope is a tool that allows seeing details in objects that would otherwise be too small to see with the eyes. Microscopes started to be used as scientific tools centuries ago. One of the early promoters for microscopes to study biological material was Robert Hooke; in 1665, he published a book called Micrographia [1], where he reported several observations from daily life. Interestingly enough, he discovered that plants are made of small structures that he called cells. Several years later, a fabric merchant, Antonie van Leeuwenhoek designed a microscope to examine fabrics. His microscope was advanced enough, and his curiosity led him to discover microorganisms. He reported seeing tiny creatures that he called Animalcules that will later change our understanding of diseases [2]. Despite the advances in science and technology, these early microscopes are similar to light microscopes nowadays.

Light microscopes, shown in Figure 1a, operate on light, i.e., photons. A photon, which is a quantized fluctuation of the electromagnetic field has a wavelength that is larger than atoms, proteins, and most viruses.

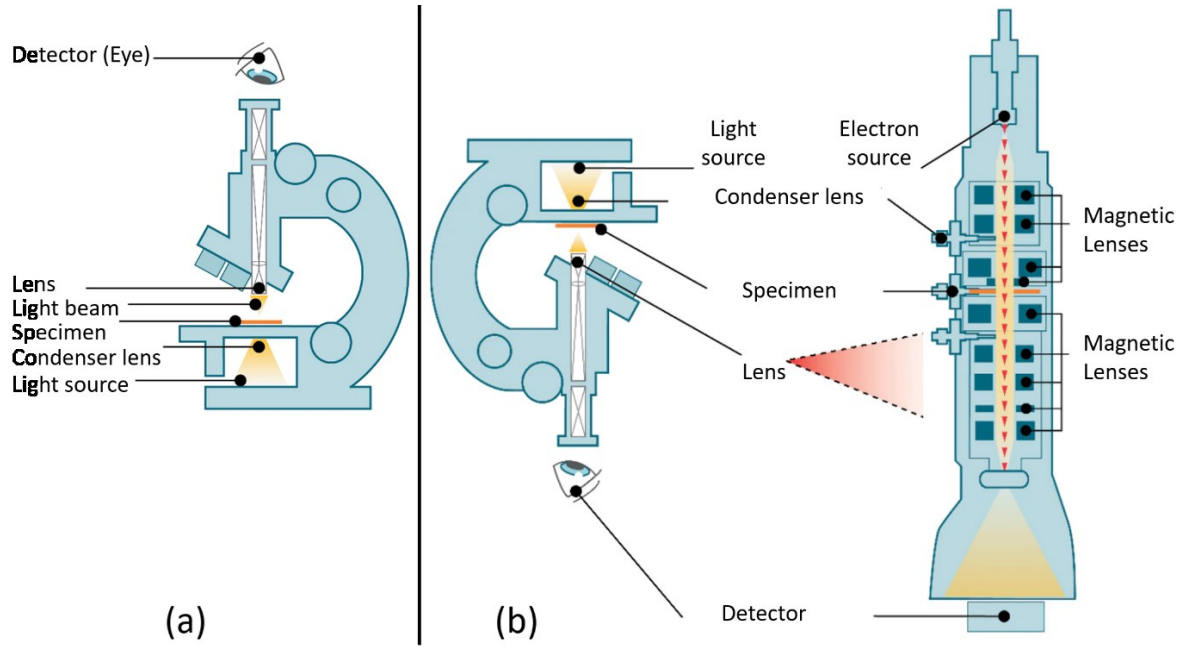


Figure 1 The general scheme of light and electron microscopes. The light microscope (a). Comparison between light and electron microscopes (b). Adapted from Thermofisher website (https://www.thermofisher.com).

Light microscopes can be used to see a range of microorganisms but are not powerful enough to see most viruses, molecules, or atoms (see Figure 2). Thus, to see smaller objects, the first electron microscope was invented by Max Knoll and Ernest Ruska in the 1930s [3]. A comparison between light and electron microscopes is shown in Figure 1b. These microscopes operate on electrons instead of photons. Electrons are up to 10^5 times smaller in wavelengths than visible light photons, giving them the ability to resolve individual atoms.

This chapter summarizes the physical and data processing principles for imaging biological samples using electron microscopes.

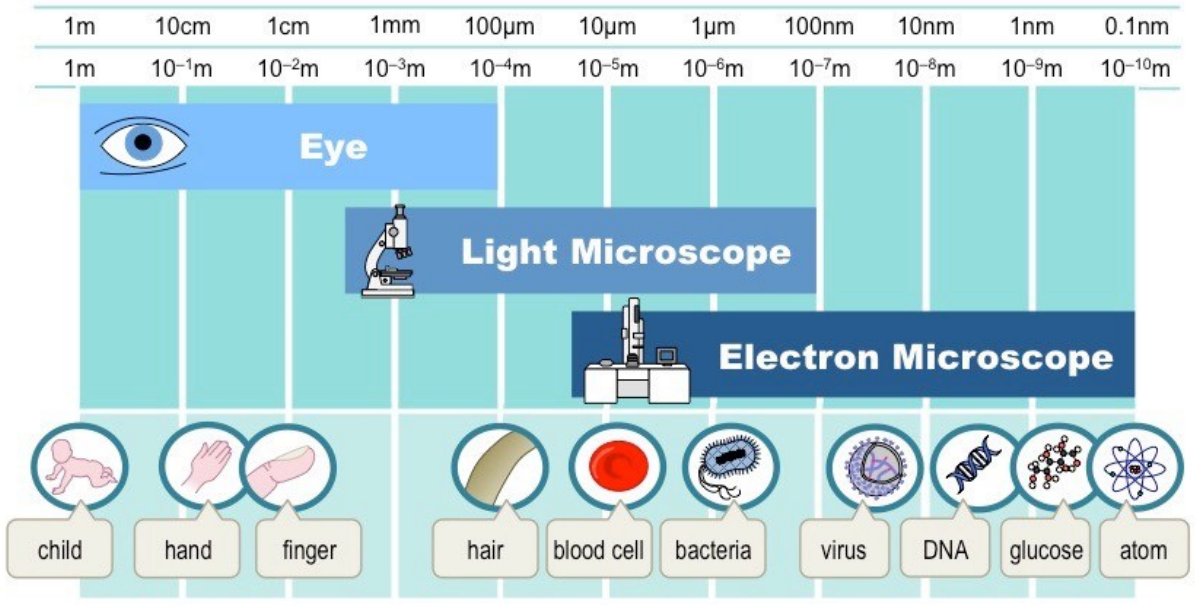


Figure 2 Range of sizes of different objects that can be resolved with eye, light, and electron microscopes. Adapted from BioNinja website (https://ib.bioninja.com.au/).

The molecules in the air scatter electrons; hence, the column of the electron microscope in which electrons travel should be vacuumed. The early usage of electrons for imaging macromolecular biological samples faced two limitations that led to the invention of cryo-EM. Firstly, the vacuum causes dehydration, thus the destruction of biological samples. Secondly, biological samples are very fragile to the high energy of the electron beam, which destroys the organic matter. An early attempt to cope with the first limitation was using dehydrated samples, which comes at the expense of damaging the native structures and leading to aggregation. An attempt to cope with the electron beam damage was the usage of heavy metals to increase the contrast and have a protective effect by coating the macromolecular complexes. The metal stains appeared darker in the electron microscopes and helped observe prints of the imaged

complexes. This technique was popular until the 1990s for observing proteins and viruses [4-6]. However, metal stains allow visualization of their interaction with macromolecules, and not the macromolecules themselves (besides other artifacts caused by the staining), thus limiting the resolution and the relevance of the observed structures [7].

In the 1980s, a technological breakthrough of imaging samples that are rapidly cooled to cryogenic temperatures to avoid dehydration and using phase contrast to avoid staining enabled resolving near-to-native high-resolution structural characteristics [8-10] in what is called cryogenic Electron Microscopy (cryo-EM).

The building blocks of cryo-electron microscopes

A schematic drawing on an electron microscope is shown in Figure 3. The column of the electron microscope contains:

1- An electron gun.
2- A condenser lens system; is used to condense the electron beam.
3- An objective lens system; is used to generate a magnified version of the object.
4- Several intermediate lens systems and a projector lens system; are used to magnify the object further.
5- An electron detector.

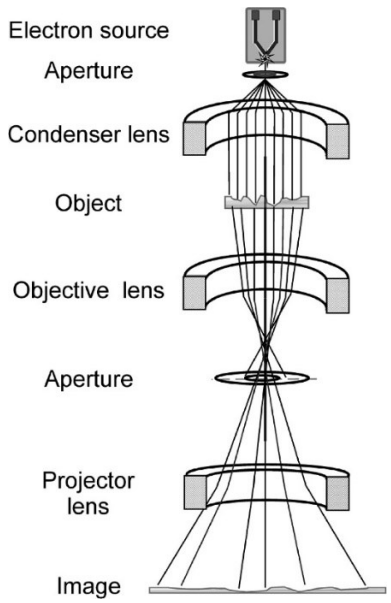


Figure 3 A schematic drawing of an electron microscope. Adapted from [11].

Image formation in the electron microscope

To form an image for a sample in the microscope, electrons interact with the sample. Amplitude contrast and phase contrast are the two types of contrast that contribute to image formation in electron microscopes [12]. To understand amplitude contrast, one must envision the incident electrons as particles scattered by objects in the sample. Some of these electrons will be absorbed by the sample, and others will be scattered at a high angle and removed by the objective lens aperture, hence, producing contrast (see Figure 4a). However, thin biological cryo-EM samples are composed of light elements (macromolecules: C, N, H, P, O, in a H, O, Ca, Mg, …environment) that scatter electrons at a low angle, leading to negligible amplitude contrast.

To understand phase contrast, one must envision the incident electrons as waves, scattered upon interaction with an object. Scattering results in phase shift. In the image plane, the scattered electron waves and the unscattered electron waves arrive at the detector recombined, and phase-contrast develops, resulting from their interference (see Figure 4b). However, scattered and unscattered electrons arrive at too small a phase shift when images are acquired precisely at focus. To increase the phase shift, lens aberrations and underfocus are used.

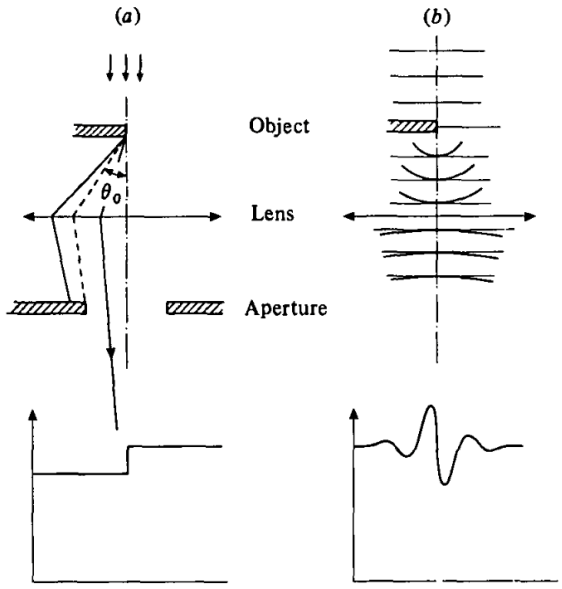


Figure 4 Schematic drawing of how the two types of contrast develop in the electron microscope (a) amplitude contrast which can be explained when the electron is envisioned as a particle, and (b) phase contrast which can be explained when the electron is envisioned as a wave. Adapted from [13].

Such an image is therefore not an exact projection of a biological sample. It is modified by the Contrast Transfer Function (CTF), summarized in the following section.

The Contrast Transfer Function (CTF) and the electron dose

CTF describes the modulation of the spatial frequency information of an image acquired by an electron microscope. CTF is related to the physical parameters of the microscope, including its accelerating voltage, lens aberrations, and, most importantly, the defocus. The CTF is a decaying two-dimensional sinusoid alternating about zero [14]. It is visualized as what is known as Thon rings in the power spectrum of the micrograph of an amorphous structure (as shown in Figure 5, bottom). The effects of the CTF include 1) loss of information at some spatial frequencies (the frequencies with zero amplitude shown by the black Thon rings in Figure 5), 2) dampening of the information of higher frequencies, and 3) phase inversion when CTF crosses zero.

The CTF effects can be partially corrected by flipping the inverted phases and boosting the higher spatial frequencies [15]. Still, CTF correction cannot retrieve frequencies where the CTF is zero.

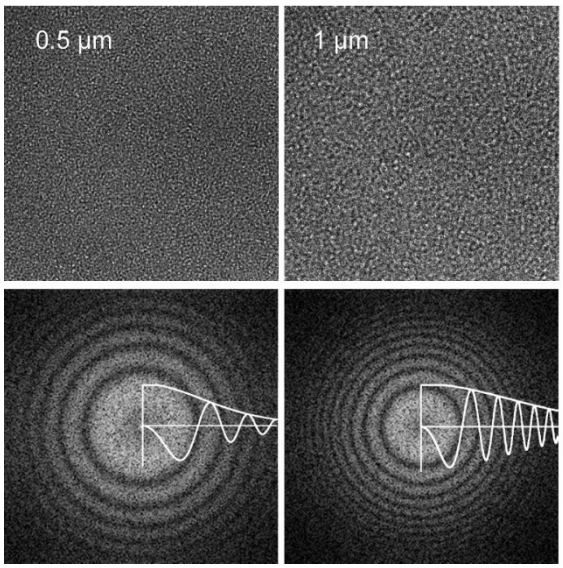


Figure 5 EM images of carbon film (on the top) and their corresponding Fourier transform (on the bottom). The Fourier transform shows the Thon rings and the corresponding CTF curves. The images were obtained with defocus values of 0.5 µm and 1 µm from left to right. Adapted from [11].

To cope with the information loss caused by the CTF, multiple image acquisitions of the sample from a fixed view can be collected at multiple defoci. However, biological samples

are fragile, and the electron dose used for imaging has to be limited to avoid significant radiation damage. A demonstration of the sample degradation with the accumulation of electron dose is shown in Figure 6. The high-frequency information, essential for observing the imaged samples at high resolution, is lost after the first few electrons per $Å^2$ ($e^-/ Å^2$) during image acquisition, and more evident damage is observed with higher electron dose accumulation [16].

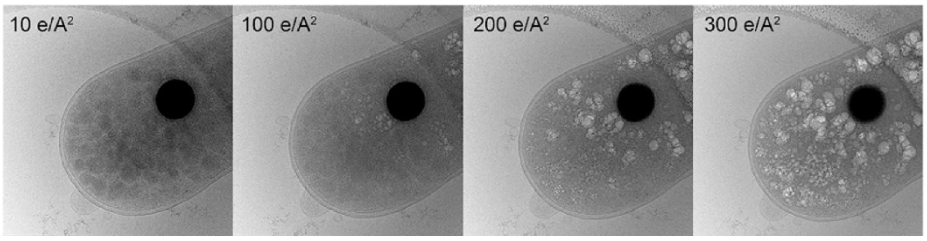


Figure 6: An example of radiation damage as a function of electron dose when imaging biological samples in cryo-EM. Adapted from [16].

Sample preparation and data collection approaches

Cryo-EM allows observing hydrated specimens in their native aqueous environment vitrified by rapid freezing. Most cryo-EM applications to date explore biomolecules' conformations *in vitro*, i.e., after purification. The scientific and technological advances in various fields allowed cryo-EM to become one of the pillars of structural biology with a method called Single Particle Analysis (SPA) [17], shown in Figure 7. SPA is based on imaging thousands to millions of copies of a biomolecule located in a vitrified thin film of the solution. Vitrification is based on rapidly freezing a thin film (typically 50-100 nm) suspended on a holey carbon film covering an EM grid (or EM grid in short) obtained from a solution containing the purified copies of the biomolecule. This rapid freezing is usually done by plunging the sample into a cryogenic fluid (e.g., liquid ethane). This method, known as plunge-freezing [13, 18], allows forming vitreous ice, which preserves the native structures and has desired properties for cryo-EM imaging. Cryo-EM images of the vitrified sample, corresponding to electron-beam projections of different holes of the EM grid for electrons that travel through the sample, is called a micrograph. A single micrograph is a two-dimensional (2D) image, possibly containing copies of the biomolecule, at random locations and orientations. For SPA, the locations of these copies are then "picked" (isolated in smaller-sub images), manually or via computer algorithms, into images containing single biomolecules, which are called single-particle images, or in short single particles [19]. Despite the challenge, algorithms for sorting the orientations of single

particles are widespread and allow high-resolution 3D reconstructions for various biomolecules [20].

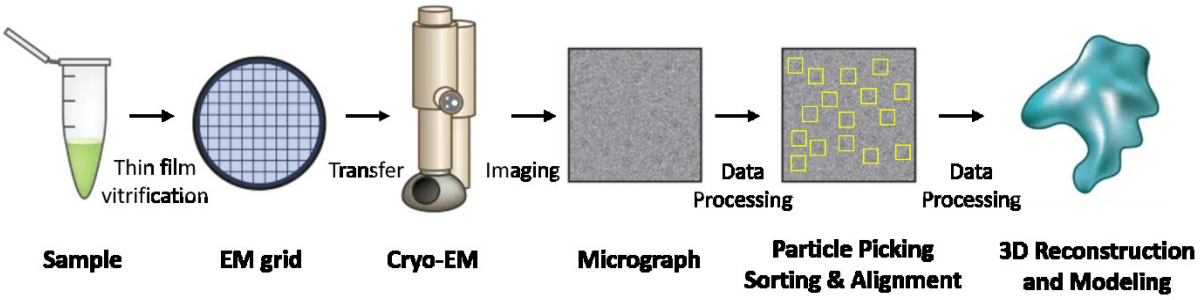


Figure 7 Cryo-EM and SPA pipeline. Adapted from [17].

Thin-film cryo-EM followed by SPA is a technique for biomolecules *in vitro*. Hence, it is not designed to image biomolecules in their cellular environment (*in situ*). On the one hand, some biomolecules cannot be purified or thus lose some of their properties. On the other hand, purifying biomolecules does not shed enough light on their functions and interactions with other biomolecules inside the cells.

In order to visualize biomolecules *in situ*, one needs to vitrify their containing cells or tissues and obtain thin enough samples for them to be imaged by cryo-EM. Small cells, like bacteria or picoeukaryotes, or thin cellular extensions can be vitrified by plunge-freezing [21], whereas thicker samples (particularly large cells or tissues) cannot. A solution to vitrify thick samples (up to 200 µm) is high-pressure freezing (HPF), where the temperature is dropped, and the pressure is increased simultaneously for a few milliseconds [22]. Vitrified cells must then be thinned using two techniques, namely, cryogenic Focused Ion-Beam milling (cryo-FIB milling) [23] or ultrathin cryo-sectioning [24].

Cryo-FIB milling is usually used for obtaining sections from cells that can be grown and frozen on EM grids. It employs a beam of gallium ions to obtain 100-250 nm thin sections called lamellae of the cellular material. Cryo-sectioning allows obtaining sections from any type of sample, including larger ones frozen by HPF, then transferred onto an EM grid. The sectioning is done using a diamond knife that mechanically cuts serial 30-100 nm thin slices of the sample. Although this technique can provide thinner sections and larger surface areas than cryo-FIB milling, it is coupled with compression artifacts caused by the mechanical cutting process, absent in cryo-FIB milled samples. Compression is a major artifact that deforms the

sample in the cutting direction and depends on the nature and density of the material and on the object scale. Deformations can be severe on the level of the long-range structures, such as entire cells or organelles [25, 26]. However, on the molecular level, deformations were reported to be negligible or inexistent [27, 28].

The crowded cell environment, with its overlapping molecules, remains an obstacle when analyzing cryo-EM micrographs corresponding to vitreous cell sections compared to thin-film samples of purified biomolecules in dilute solution. Hence, 2D SPA is usually not the go-to method in this case, but cryogenic Electron Tomography (cryo-ET) thanks to the 3D information that allows disentangling the crowded environment [29].

Cryo-ET is a 3D imaging technique based on obtaining cryo-EM micrographs of a sample at multiple tilting angles, usually from -60 to 60 degrees with 1 to 4 degrees angular step [30]. The set of micrographs of different tilted views is called a "tilt-series" and is then reconstructed into a 3D volumetric image called a tomogram. This principle of cryo-ET is illustrated in Figure 8.

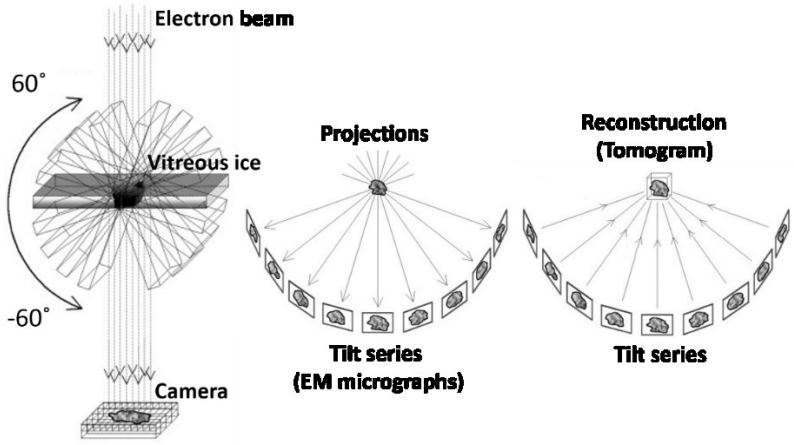


Figure 8 Principle of cryo-ET image acquisition scheme. The left shows the tilt series acquisition of a 3D object. The middle shows the tilt series and what they represent relative to the 3D object. The right shows how the tilt series can be reconstructed into a 3D volume called a tomogram representing the 3D object. Adapted from [29].

A tomogram, possibly corresponding to a vitrified cell section or lamella can contain copies of biomolecules at locations and orientations that can shed light on their biological function and mutual interactions. The locations of these copies can be picked, manually or via computer algorithms, into sub-volumes containing single copies of the biomolecules, which are called subtomograms. Algorithms (explained in the next chapter) for aligning the

subtomograms and obtaining an average of them are called Subtomogram Averaging (STA or StA) [31]. STA allows high-resolution 3D reconstructions of biomolecular structures [32].

Tomographic reconstruction and missing wedge artifacts

The fundamentals of obtaining 3D reconstructions of biomolecular complexes from cryo-EM single-particle images and obtaining tomographic reconstructions from cryo-ET tilt series images are best explained by the Fourier central slice theorem [33, 34]. According to this theorem, a central slice through the Fourier transform of a 3D object corresponds to the Fourier transform of the object's 2D projection at the same orientation as the orientation of the central slice (see Figure 9). In cryo-ET tomographic reconstruction, the 2D Fourier transform of an aligned tilt-series [35] is positioned to create a 3D Fourier transform of the reconstructed tomogram. A tomogram is obtained by the 3D inverse Fourier transform.

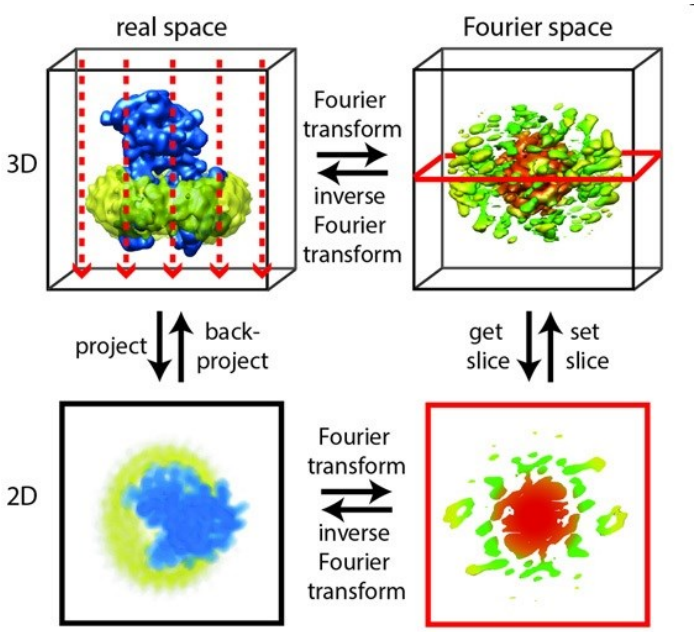


Figure 9 Fourier central slice theorem. The 2D Fourier Transform of a 2D projection image resulting from projecting a 3D object corresponds to a central slice across the original 3D object's Fourier transform with the same projection angle. Adapted from [36].

Several tomographic reconstruction algorithms are useful for different applications. The most common approach is Weighted Back-Projection (WBP) [37], which calculates for each image in the tilt series a backprojection body, and the tomographic reconstruction becomes the sum of all the backprojection bodies. WBP uses a weighting filter that reduces the contribution of low spatial frequencies. Several other reconstruction algorithms exist and are usually

algebraic-based; two of the most common ones are Simultaneous Iterative Reconstruction Technique (SIRT) [38] and Simultaneous Algebraic Reconstruction Technique (SART) [39]. SIRT and SART offer a good contrast which can be useful for particle picking, whereas WBP provides a higher resolution for subtomogram averaging.

Due to the limitation of the tilting angle used in acquiring the tile series, which is usually limited in the range $\pm 60°$, a tomogram in Fourier space has a missing wedge region (shown in Figure 10 on the top). The missing views in the tilt series used to reconstruct tomograms, i.e., the missing wedge in the Fourier space of a tomogram, cause data anisotropies known as missing wedge artifacts [40, 41].

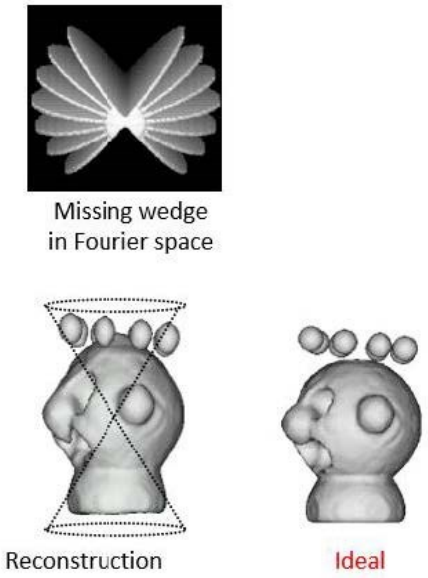


Figure 10 Missing wedge artifacts in cryo-ET tomographic reconstructions. From [42].

Chapter 2. State of the art: cryo-ET methods for biomolecular conformational variability analysis

Conformational variability of biomolecular complexes

The structures (their shapes) and chemistry of biomolecular complexes determine their functions to a large extent. Hence, understanding their structures is crucial to understanding their working mechanisms. However, biomolecules are flexible, and they exhibit gradual conformational transitions, referred to as continuous conformational variability, as shown in Figure 11A. As will be seen in the next chapter, this variability can be intrinsic or induced by other factors exerting forces on them. Besides continuous variability, biomolecules can bind or unbind substrates, giving rise to a form of discrete variability, shown in Figure 11B. Continuous and discrete variabilities may happen together, such as binding and unbinding different substrates while continuously changing conformations, as shown in Figure 11C.

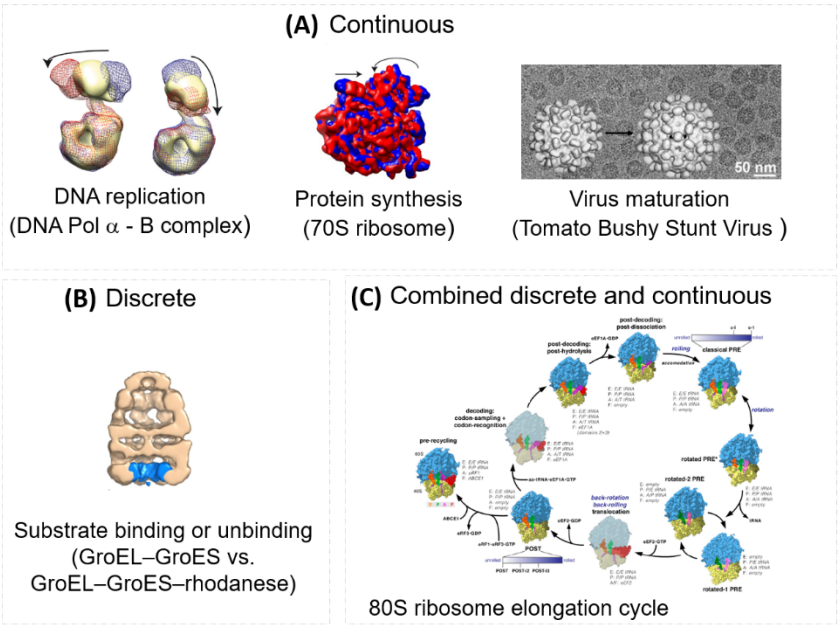


Figure 11 Types of conformational and compositional variabilities. (A) Examples of continuous conformational variabilities for biomolecules: DNA Pol α-B complex continuous conformational movements, 70S ribosome continuous ratchet-like movement, and tomato bushy stunt virus swelling-like movement. (B) An example of discrete compositional variability of substrate binding GroEL-GroES vs. GroEL-GroES-rhodanese. (C) combined discrete and continuous variabilities of 80S ribosome elongation cycle, binding and unbinding different ligands while continuously changing conformations. Adapted from [43-45].

Data processing methods for determining biomolecular structure and dynamics from cryo-ET specimens

Subtomograms contain copies of biomolecular structures at different locations, orientations, and conformations. Subtomograms suffer from a low signal-to-noise ratio (SNR) due to the small electron dose used to obtain the tilt series to avoid radiation damage to the fragile biological sample. The MW artifacts are often observed as elongation along the beam axis, blurring, and distracting caustics in the subtomograms. Due to the low SNR and the MW, cryo-ET data processing is mainly based on rigid-body aligning and averaging many subtomograms to enhance the data quality and reveal the targeted biomolecular structure.

STA is the process that allows obtaining high-resolution models of biomolecular structures from cryo-ET subtomograms after aligning them to a reference orientation and averaging them, as shown in Figure 12.

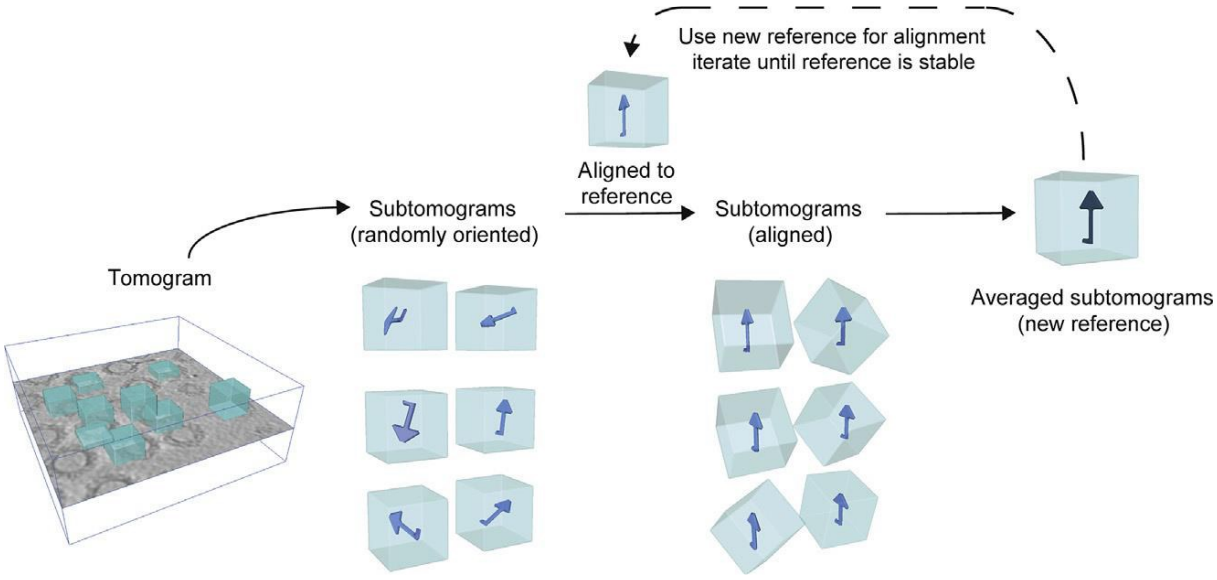


Figure 12 The process of subtomogram averaging. Adapted from [31].

However, as biomolecules are flexible, the biomolecular copies in subtomograms vary due to this flexibility, giving rise to two STA problems. On one side, this variability limits the resolution of 3D reconstructions since subtomograms do not contain identical copies. On the other side, STA hides the information relevant to conformational transitions that help study biomolecular mechanisms captured in cryo-ET. In order to account for these two problems, previous computational methods simplify the problem by discretizing the gradual transitions of biomolecular structures and addressing the problem via classification. Classification can be

done during STA or posterior to it. Hereafter, I review STA and classification techniques used to determine structure and dynamics in typical cryo-ET data processing pipelines other than the methods developed in this thesis.

Subtomogram Averaging (STA)

STA workflow (shown in Figure 12) is an iterative alignment and averaging procedure: subtomograms are aligned against a reference to maximize a scoring function; the aligned subtomograms are averaged to produce a volume that becomes the reference for the next iteration of alignment. The starting reference is preferably chosen as the average of all or a subset of the subtomograms that are randomly oriented; however, in some cases, another starting reference can be justified, e.g., a very low-resolution version of the anticipated average shape, which gets refined iteratively during STA. The scoring function used in STA is usually the Constrained Correlation Coefficient (CCC) [46], which restricts the Correlation Coefficient (CC) calculation to the Fourier space region that excludes the MW.

The search for the angular and shift alignments of subtomograms can be done in Real or Fourier spaces [47, 48], allowing the determination of the six-dimensional rigid-body parameters, i.e., three rotational angles and three shifts for each of the subtomograms, applied to align them to the global subtomogram average.

STA is a computationally demanding process, especially for large datasets; however, nowadays, datasets of subtomograms are small compared to SPA in terms of the number of instances (subtomograms), especially *in situ*.

The symmetry of an analyzed complex can be beneficial for structural determination in STA since each subtomogram can be assigned to multiple views of the global average, depending on the axes of symmetry.

Before diving into classification practice, it is worth mentioning that the MW artifacts must be carefully considered when producing subtomogram classes. Otherwise, one might fall into the trap of producing erroneous results by misinterpreting these MW artifacts as different conformational classes. Moreover, when a subtomogram is aligned against a global reference (the subtomogram average), the shifting parameters resulting from the alignment can be applied to the subtomogram. In practice, after STA, the shifting information can be used to extract centered subtomograms that will require only angular information (three angles, commonly

addressed as Euler angles) to be applied to each subtomogram to align it to the global average. In some classification techniques, in particular post-alignment classifications (explained next), centered subtomograms are used.

Subtomogram classification posterior to STA

The simplest form of classification is to set a threshold value for the CC between the aligned subtomogram and the reference; all subtomograms below the threshold are removed from the dataset (not taken into the average), which can be helpful when most subtomograms contain the properly aligned structure of interest; in such cases, CC thresholding can be used to remove misaligned subtomograms or those containing other structures.

Nevertheless, when there is a high degree of structural heterogeneity or misalignment, such approaches may fail because the CC is calculated using a reference averaged from highly heterogeneous particles. A more sophisticated approach to post-alignment classification is to compare the subtomograms in the dataset and sort out the dataset into several different classes. This clustering is usually performed on a correlation matrix using Principal Component Analysis (PCA) [49] and k-means clustering or hierarchical clustering [50, 51].

In the following, a brief background [51] on post-alignment classification is summarized.

Let S_i and S_j be two subtomograms associated with their corresponding sets of Euler angles $(\alpha_i, \beta_i, \gamma_i)$ and $(\alpha_j, \beta_j, \gamma_j)$ obtained via STA to align the subtomograms to a common reference system (that of the global subtomogram average) and let W be a binary missing wedge window.

Then $\grave{S}_i$ and $\grave{S}_j$ can be defined as the aligned version of S_i and S_j respectively, i.e. $\grave{S}_i$ and $\grave{S}_j$ are found by applying the corresponding set of Euler angles to S_i and S_j. Similarly, let w_i and w_j be the missing-wedge windows that correspond to $\grave{S}_i$ and $\grave{S}_j$, i.e. w_i and w_j are found by applying the corresponding set of Euler angles to W. Let $w_{ij} = w_i * w_j$ be the intersection of the two w_i and w_j, i.e. $w_{ij} = 1$ only for the region where the Fourier space of both $\grave{S}_i$ and $\grave{S}_j$ is not missing. Let M be a mask for the region of interest in the aligned subtomograms, where M can be binary or a more sophisticated mask constructed based on morphological operations on the global subtomogram average.

We can now obtain a normalized and common missing wedge constrained version of the two subtomograms. The normalization can be done by subtracting the mean μ and dividing by the standard deviation; however, the mean and standard deviation should be constrained both to w_{ij} in Fourier space (obtained via Fourier Transform FT) and to M in the real space:

$$\widetilde{S_\iota} = \frac{M * FT^{-1}\left(FT\left(\grave{S}_\iota\right) * w_{ij}\right) - \mu_i}{\sum_{x,y,z} M \sqrt{\sum_{x,y,z}\left(M * FT^{-1}\left(FT\left(\grave{S}_\iota\right) * w_{ij}\right) - \mu_i\right)^2}}$$
$$\widetilde{S_\jmath} = \frac{M * FT^{-1}\left(FT\left(\grave{S}_\jmath\right) * w_{ij}\right) - \mu_j}{\sum_{x,y,z} M \sqrt{\sum_{x,y,z}\left(M * FT^{-1}\left(FT\left(\grave{S}_\jmath\right) * w_{ij}\right) - \mu_j\right)^2}} \quad (2.1)$$

The covariance matrix CM can then be written as the matrix for which its elements are:

$$\mathrm{CM_{ij}} = \sum_{xyz} \widetilde{S_i}\,(x,y,z)\,\widetilde{S_J}(x,y,z) \quad (2.2)$$

Where i and j are for all the subtomograms in the dataset.

This covariance matrix in eq (2.2) is then fed to a classification algorithm (e.g., Hierarchical clustering) or is fed first to a dimensionality reduction technique (e.g., PCA) followed by a clustering algorithm (e.g., k-means).

Simultaneous subtomogram classification and alignment

Classification of subtomograms during alignment is usually performed through multireference alignment or maximum likelihood estimation.

In simple terms, multireference alignment starts from known references of low resolution that are used to separate a heterogeneous dataset. Multireference-based approaches require a number of references, which determines the number of classes [50, 52]. There are two strategies to perform a multireference alignment. In both strategies, each subtomogram is aligned against all the given references. In the first strategy, each subtomogram is assigned to a class to which it is most similar, and at the end of each iteration, the members of each class are averaged to generate a new set of references. The second strategy uses a scoring function to determine the contribution weight (between 0 and 1) for each subtomogram to each class, i.e., each subtomogram contributes to the average of all the classes with different weights. In both

strategies, the differences between the classes are amplified through iterative alignment and averaging, and the starting references are refined.

Subtomogram classification based on maximum likelihood estimation [53-55] is rather more sophisticated than multireference alignment. It is based on a different data model that does not require starting references but only choosing a number of desired classes. The data model used for maximum-likelihood estimation [53] is given by:

$$X_i^0 + X_i^m = R_{\Phi_i} A_{k_i} + G_i \tag{2.3}$$

Where: X_i^0 is Fourier transform of an input subtomogram i. X_i^m is the missing Fourier components from the X_i^0 due to the MW. k is the number of classes chosen by the user, and k_i is the class that the subtomogram i will be assigned to it. A_{k_i} is one of k unknown 3D structures in Fourier space. $R_{\Phi_i}$is the transformation matrix that represents the six-dimensional rigid-body alignment parameters Φ_i (three angles and three shifts) that aligns A_{k_i} to the subtomogram (and its inverse can align the subtomogram to A_{k_i}). G_i is independent white Gaussian noise with mean zero and unknown standard deviation σ. An alteration of this data model can be done in a few ways, such as the absence of the term X_i^m which leads to a different optimization problem [56] or the change of the noise model from white to colored [57].

The task is the determination for each subtomogram i the class assignment k_i and the six-dimensional rigid-body alignment parameters Φ_i. An optimization algorithm was originally implemented in [53] for the task above, and advanced versions were recently implemented in the software package Relion [58].

Methods summary for analyzing conformational variabilities in subtomograms

Cellular cryogenic electron tomography (cryo-ET) is currently undergoing its "resolution revolution" reaching a near-atomic resolution *in situ* [59] and allowing studying macromolecules in their physiological environment that affects their conformational landscape [28, 60].

This section summarizes and discusses practical considerations, advantages, and disadvantages of families of STA and classification methods, summarized in Figure 13.

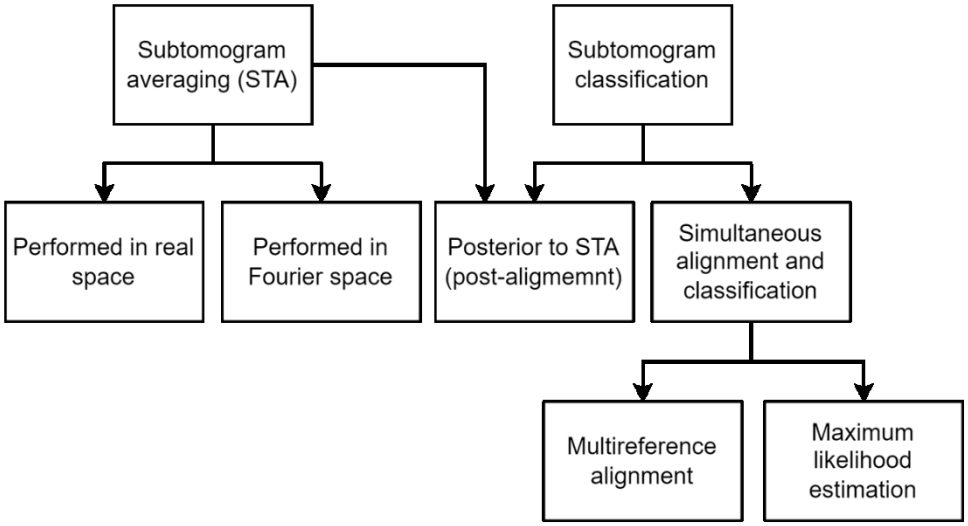


Figure 13 Families of subtomogram averaging and classification.

Mainstream STA techniques nowadays perform what we can call an exhaustive search for the six-dimensional parameters for every subtomogram in a dataset to align all subtomograms to a common reference system (the one of the subtomogram average). This six-dimensional search problem is computationally expensive. Thus, the search for these parameters is usually divided into consecutive optimizations by coupling two angles and searching for the shifts separately. However, STA remains computationally expensive and usually requires a few days to align small-size subtomogram datasets (ranging between hundreds to tens of thousands), even on powerful computers using algorithms that use new technologies (multiple CPU and GPU computing). Faster alignment of subtomograms can be performed using Fast Rotational Matching (FRM) [48] by aligning the subtomograms by taking advantage of Spherical Harmonics (Fourier space of spherical coordinates), but it has not yet gained popularity in readily available cryo-ET packages despite its high potential in reducing the computational time. Other techniques implemented this search based on machine learning [61], but it is still an open field and has not yet produced significant results up to my knowledge.

Subtomogram averaging followed by post-alignment classification, and multireference alignment, are widespread techniques for subtomogram alignment and classification and are offered by the most readily available packages for cryo-ET data analysis, e.g., Dynamo [50]. Post-alignment classification is simple to use and, at least for advanced users, does not necessitate defining the number of classes in advance. The resultant dendrogram (tree) can be viewed before deciding the number of classes when utilizing post-alignment classification via Hierarchical clustering, for example. Furthermore, it is not computationally expensive. Because

it is based on reconstructing a covariance matrix (correlation matrix), the matrix can be reconstructed once, and a varied number of classes can be tested. However, post-alignment classification has a disadvantage: it is strongly reliant on STA alignment quality, which degrades with widely varied specimens.

Multireference alignment methods require prior knowledge of the specimen's anticipated conformations by providing a set of starting references. Although these references are refined during the alignment, these methods are still prone to bias by the references' choice, resulting in overfitting and misinterpretations. Another problem in multireference alignment is called the “attractor problem”, where the class that contains the larger number of subtomograms results in the highest similarity score with the subtomograms in the dataset since its average has the highest SNR. The attractor problem often results in empty classes, i.e., some starting references die during iterative alignment due to the attractor problem.

Maximum likelihood estimation methods for subtomograms alignment and classification started to gain popularity in the past few years since they were offered by the Relion package [58] (commonly used for SPA and recently extended to cryo-ET). Nevertheless, maximum likelihood estimation methods suffer from two major drawbacks. First, the number of classes must be set (decided) in advance. A new choice of the number of classes requires repeating the entire maximum likelihood-based alignment. Second, they suffer from the attractor problem, as in multi-reference alignment techniques explained above.

Generally speaking, classification helps remove outliers and separate discrete types of variabilities. However, the drawbacks discussed above for subtomograms classification techniques limit their capacity to provide classes showing transitions of biomolecular conformations. The limited number of classes averages out rare conformations necessary to understanding the biomolecular landscape. Also, in general, biomolecules are flexible, with continuous conformational transitions; hence, particles assigned to the same class will rarely, if ever, have perfectly identical conformations, resulting in lower resolution averages.

With recent instrumentation and software development, more research moves toward studying single-particle subtomograms individually (with no or a minimum of averaging) by developing new methods for denoising, missing wedge correction, and 3D reconstruction [40, 41, 62].

Chapter 3. State of the art: Chromatin and nucleosome studies

This thesis analyzes cryo-ET data obtained from Drosophila embryos frozen by HPF and cryosectioned to explore the conformational variability of nucleosomes *in situ*.

The following chapter provides a brief literature review on nucleosome structure and dynamics, introduces the biological model used in the study, and explains how HPF and cryo-ET of vitreous sections (CETOVIS) are employed for data acquisition.

Chromatin and the Nucleosome Core Particle (NCP)

Eukaryotic cells organize their genome, reaching several billions of Deoxyribonucleic Acid (DNA) base pairs (bp), by packing it into a nucleus [63]. The genome is not randomly packed but is organized into chromatin. Chromatin is formed by a series of monomers called nucleosomes, forming the so-called “beads-on-a-string” filament [64], as shown in Figure 14.

Chromatin repeats nucleosomes every 160 to 240 bp of DNA, which varies between species, cells, and even within cells. The Nucleosome Core Particle (NCP) [65] shown in Figure 14 comprises ~146 bp of DNA wrapped ~1.65 times, forming a left-handed superhelix around a histone octamer. The histone octamer consists of two copies of each of the histone proteins H2A, H2B, H3, and H4. Histones contain flexible N-terminal regions called histone tails, which extend from the core and may carry post translational modifications (PTMs), and have critical functions in regulating chromatin [66].

Chromatin and its nucleosome building blocks regulate the major genome processes of transcription, replication, and repair [67, 68]. Different chromatin regions can be functionally defined (Figure 15): gene-rich and transcriptionally active euchromatin (EuC), and more compact and littler transcribed heterochromatin (HC) domains, including constitutive heterochromatin (cHC), gene-poor and enriched in repetitive sequences, and facultative heterochromatin (fCH) which can switch between active and repressed states. They have different chemical PTMs on the histone tails (acetylation, methylation, and phosphorylation). PTMs and other regulatory factors remodel nucleosome distribution in space and time and can

incorporate histone variants. For example, PTMs allow chromatin to be more or less condensed [69] or bind specific regulatory factors.

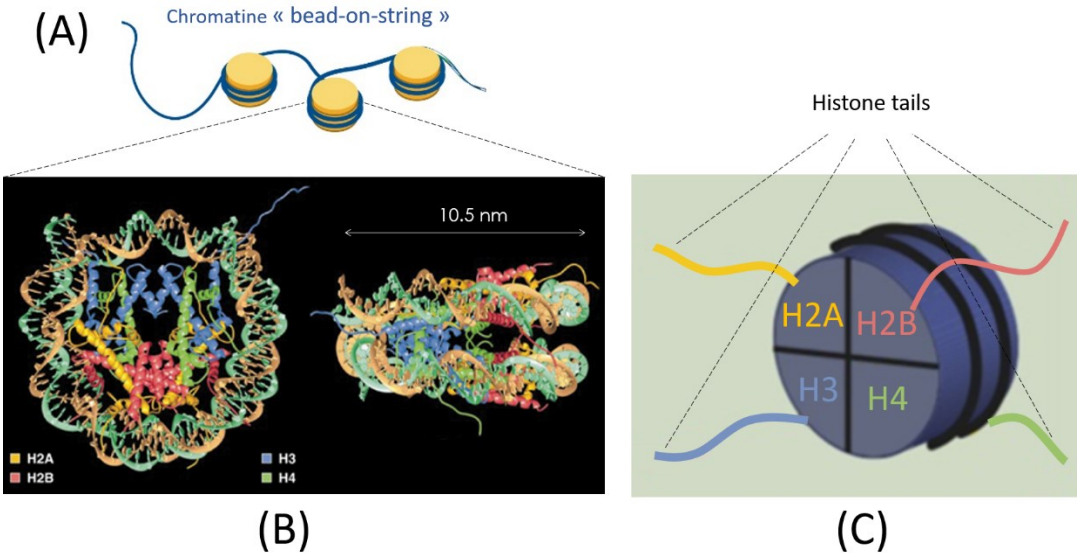


Figure 14 (A) "Bead-on-string" structure of chromatin, with nucleosomes linked by a DNA segment. (B) The nucleosome core particle consisting of ~146 bp of DNA wrapped ~1.65 turns around a histone octamer as a building block of chromatin in bead-on-string form. (C) Sketch of the histone tails. Adapted from [65, 66].

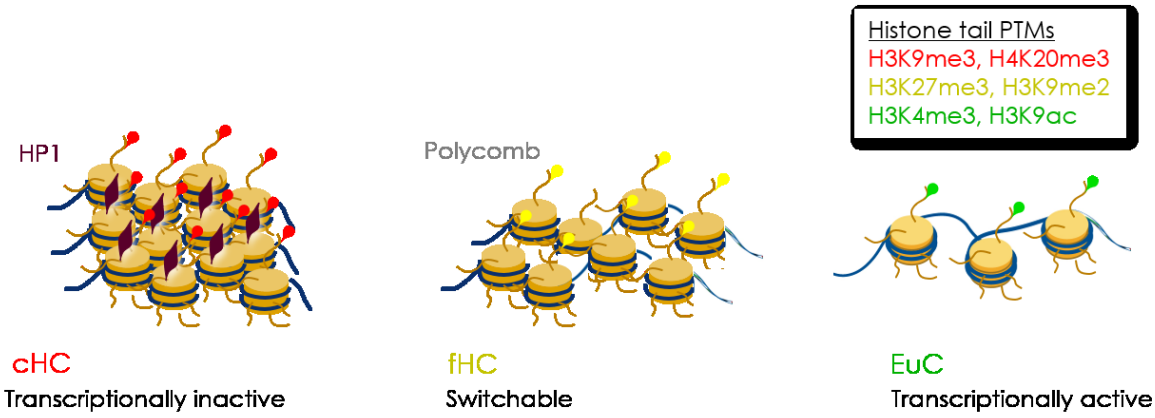


Figure 15 Major chromatin regions: cHC, fHC, and EuC differ by histone tail PTMs. Courtesy of Amélie Leforestier.

The nucleosome family of conformations and variants

The canonical nucleosome structure shown in Figure 14 was obtained by X-ray crystallography of engineered nucleosome particles that are identical, symmetric, highly stable, and assembled from recombinant histones and optimal DNA sequences [65, 70-73]. However,

there is increasing evidence that nucleosomes are a family conformations [28, 67, 74-83]. However, most studies were performed *in vitro* or *in silico*.

In vitro, nucleosomes exhibit several types of variations categorized into salt-induced transitions, intrinsic dynamics, chemical variations, and conformational changes upon interaction with other proteins. An interested reader is referred to recent reviews in [84-86]. Examples are illustrated in Figure 16.

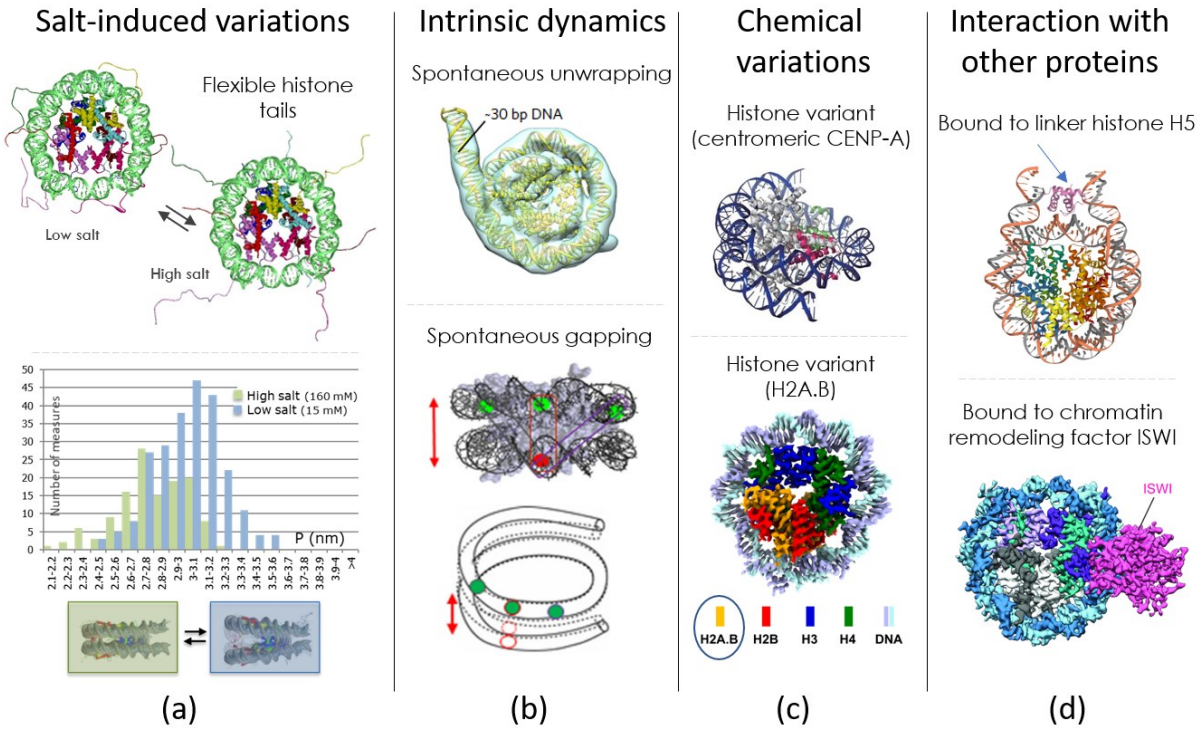


Figure 16 Examples of four categories of nucleosome variations obtained from *in vitro* studies. (A) At different salt concentrations, (on top) nucleosomes extend and retract histone tails [82], and (on bottom) change the distance between the DNA gyres (P) [28]. (B) Nucleosomes can exhibit spontaneous dynamics, such as unwrapping a segment of their DNA (or breathing) and gaping, or edge opening [75, 81]. (C) Two examples of histone variants; as centromeric CENP-A and H2A.B [77, 87]. (D) Nucleosomes analyzed interacting with linker histone H5 and the chromatin remodeling factor ISWI [88, 89]. Adapted from [28, 75, 77, 81, 82, 87-89].

Theoretical models and simulations also predict nucleosome conformational variations related to their dynamics and/or and histone composition, as shown in Figure 17 [90], with the gaping, opening, and breathing motions, subnucleosomal particles such as “hemisomes” (half nucleosomes) [91] or “tetrasomes”, and “reversomes”, nucleosomes with right-handed DNA wrapping [92], to mention a few.

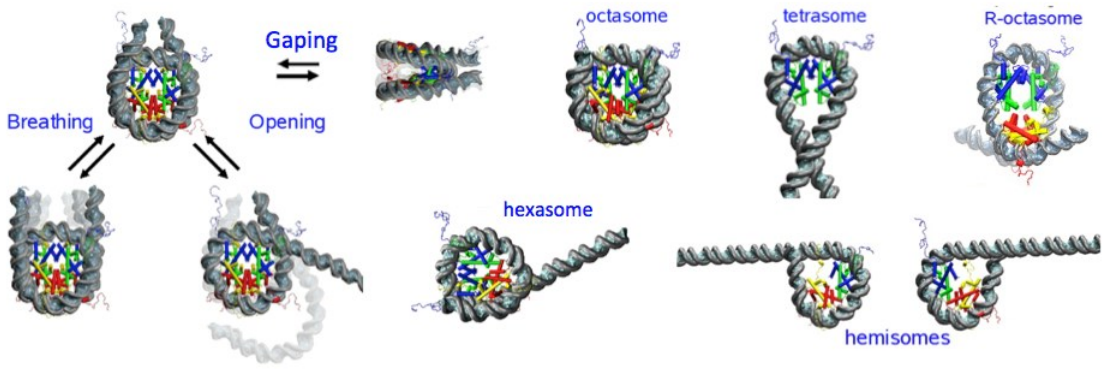


Figure 17 Examples of hypothetical *in silico* nucleosome conformational changes showing nucleosome gaping, opening (also referred to as DNA unwrapping in other literature), and breathing. Non octameric nucleosomes and right-handed particles are also predicted. Adapted from [90].

In situ, most studies are indirect [93, 94]. Recent cryo-ET approaches provided the first direct insights into nucleosome structure and variability. At the start of this thesis, two research studies documented nucleosome variability *in situ* [28, 95], one of which was performed in the research group of the co-supervisor of this thesis, Dr. Amélie Leforestier [28].

In [95], nucleosomes from a HeLa cell thinned by cryo-FIB milling were observed using cryo-ET. The resultant tomogram, shown in Figure 18A, shows the nuclear envelope and the nucleoplasm with nucleosomes in densely and loosely populated regions associated with HC (Figure 18C) and EuC (Figure 18D). Two averages were also obtained via subtomogram averaging and classification shown in Figure 18B, one of which resembles a canonical nucleosome and the other associated with longer DNA, as in a chromatosome [96].

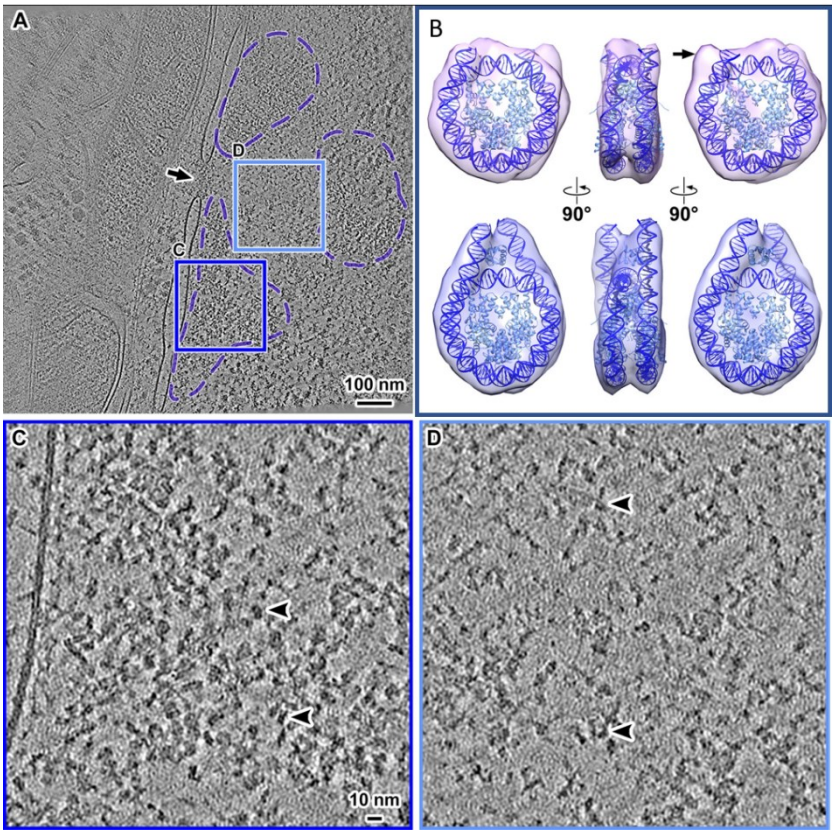


Figure 18 (A) Cryotomogram of a cryo-FIB-thinned HeLa cell. The nuclear envelope with a nuclear pore complex (arrow) is recognized. Chromatin regions with different nucleosome concentrations are observed, corresponding to HC (enclosed by purple dotted lines) and EuC. (B) Two subtomogram averages are shown at 50% transparency with the edited crystallographic structures docked PDB 1AOI on top and PDB 5NLO (chromatosome with linker histone) at the bottom. (C, D) Tomographic slices (10 nm) of the (C, HC) and (D, EuC) positions boxed in A, enlarged (zoomed in) by a factor of 4.5. Adapted from [95].

Drosophila embryonic brain: a biological model to study nucleosome *in situ*

The primary model used in [28] is Drosophila embryonic brain at late developmental stages (12-15) shown in Figure 19. Embryos, protected by their vitelline membrane, can be vitrified by HPF without damaging risks of osmotic stress induced by the surrounding cryoprotective solution. Cell nuclei occupy a relatively large volume fraction in the Drosophila embryonic brain cells, thus providing a higher chance of finding chromatin regions in cryo-sections. In addition, cHC forms a unique compact domain that can be distinguished from EuC/fHC upon purely morphological criteria, thus facilitating the identification of chromatin compartments, opening perspectives for the retrieval of functional information. These advantages render the Drosophila embryos a well-characterized experimental system for studying nucleosomes *in situ*.

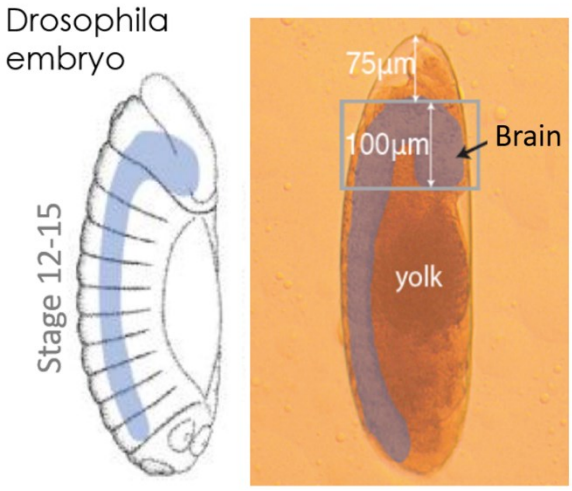


Figure 19 Drosophila embryo at stages 12-15 used as a biological model for studying nucleosomes *in situ*. Courtesy of Mikhail Eltsov.

The tools and methods, from Drosophila flies to nucleosome-containing tomograms

This section summarizes the workflow used to obtain nucleosome-containing cryo-tomograms of Drosophila embryonic brain cells, starting from the culture of the Drosophila flies.

During my Ph.D., I participated in experimental sessions that aim to record new cryo-tomogram for further analyses. I describe here the procedures used during these experimental sessions. Similar procedures were followed by M. Eltsov and A. Leforestier before the beginning of my thesis to obtain the tomographic data used in this thesis (see Chapter 4). The results available before I started are summarized at the end of this section.

Drosophila melanogaster flies (Bloomington Stock number 30564) are maintained in a standard Bloomington medium, and Embryos are collected on grape juice agar plates as shown in Figure 20a. The embryos are floated in water and then plunged in 50% v:v bleach for a few seconds in order to dechorionate them, i.e., for dissolving their chorion, to allow their vitrification as shown in Figure 20b. The dechorionated embryos are still viable and develop normally. They are inspected under a stereomicroscope, and late development stages (14-15) are identified [97] and transferred into copper carriers filled with dextran 25% in phosphate-buffered saline, as embedding medium. The carrier is mounted on the HPF holder (Figure 20c) and is transferred to the HPF machine (Wohlwend Compact 03), which applies high pressure

(2045 bars) and low temperature (liquid nitrogen at -196°C) simultaneously, as shown in Figure 20d.

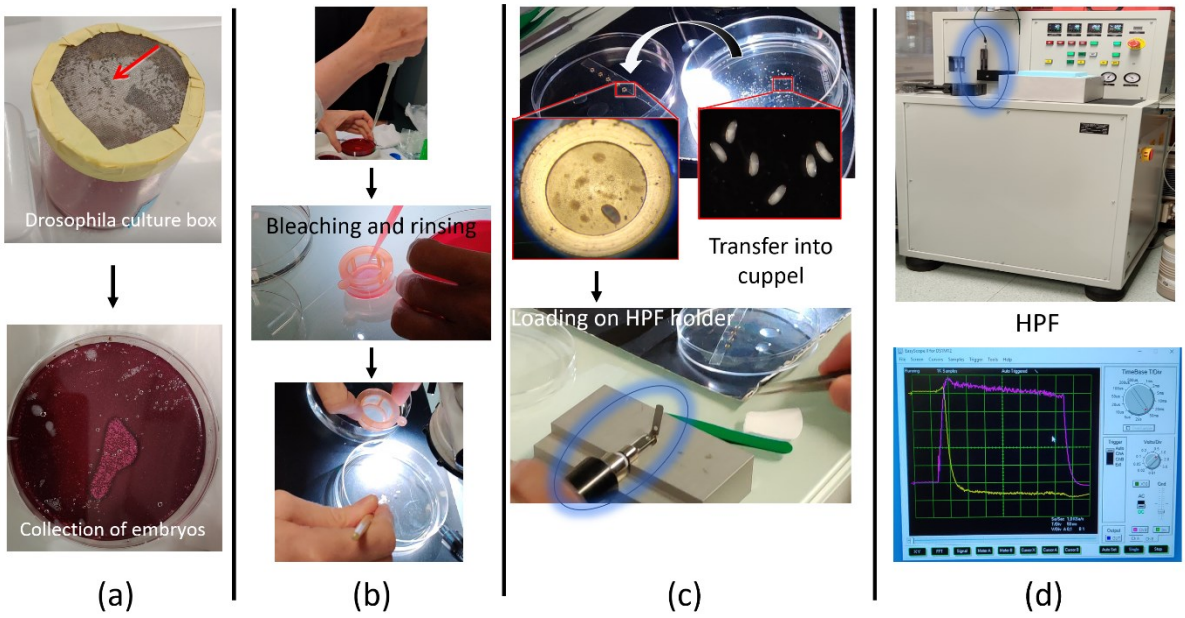


Figure 20 Experimental procedures followed to culture Drosophila embryos (a) rinsing and dechorionated them (b) identifying late development stages and transferring them on carriers then HPF rod (c) vitrifying the embryos using HPF (d). See text for details.

Frozen embryos are sectioned at -145°C using an ultramicrotome (Leica Ultracut FC6/UC6) installed in a controlled environment (humidity level is maintained below 20%) using diamond knives (25°, Diatome) to obtain sections of 50 or 75 nm thickness, that are collected on EM grids (e.g., Quantifoil R2/2), as shown in Figure 21a. The sections are checked by 2D imaging using an available EM (JEOL 2010F at LPS) until nuclei-rich cell regions are reached (embryonic brain cells); serial sections of these regions are then prepared and stored for cryo-ET (tilt series acquisition) using a high-end EM (Titan Kryos 300kV), as shown in Figure 21b. The raw data (tilt series movies) are processed and reconstructed into a tomogram (see Chapter 4 for more details).

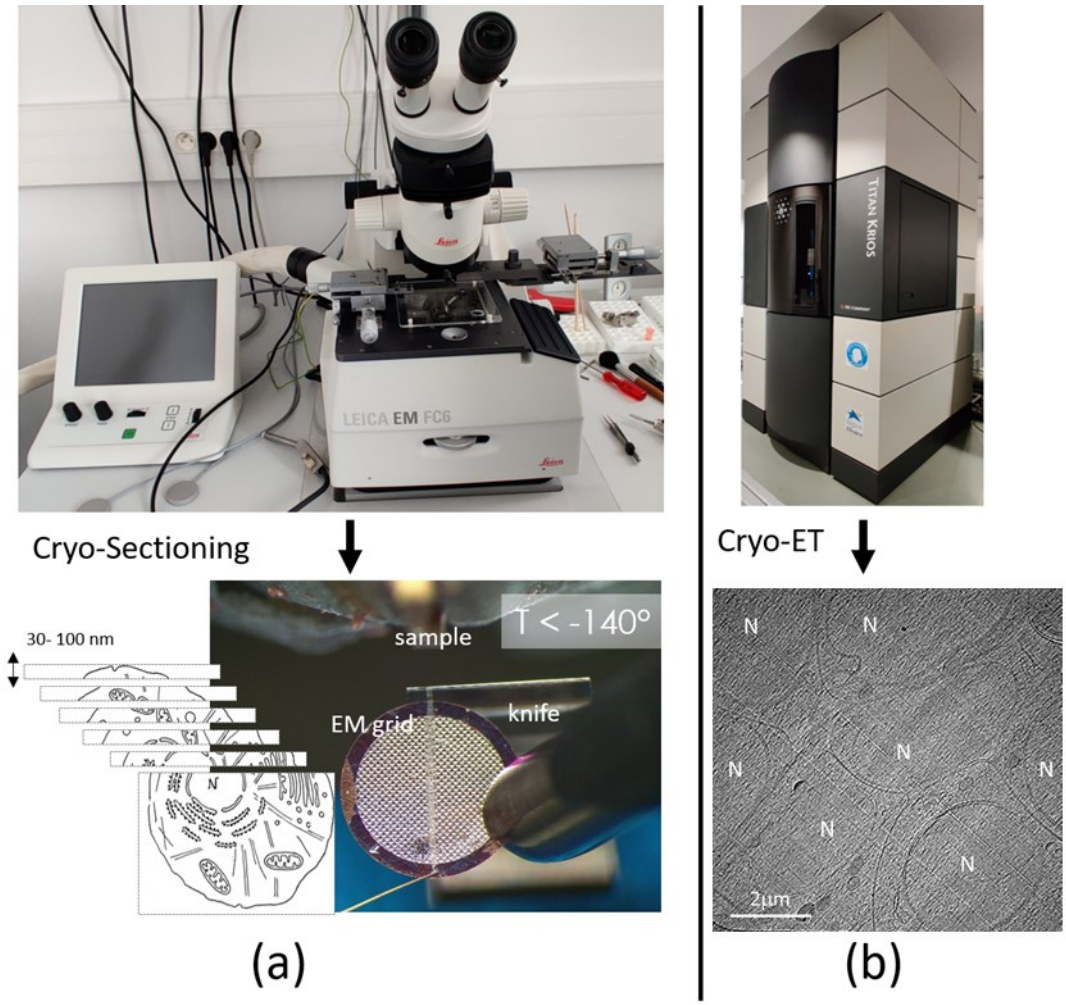


Figure 21 Cryo-Sectioning of Drosophila embryos (a) and cryo-ET tilt-series acquisition on high-end cryo-EM (b). See text for details. Source: ((a), bottom) Courtesy of M.Eltsov & A.Leforestier. ((b), bottom) Courtesy of Fatima Taiki.

In my thesis, I analyzed a dataset obtained by M. Eltsov with a Titan Krios (FEI, Thermofischer, Eindhoven, The Netherlands) operated at 300 kV equipped with a GATAN GIF Quantum SE post-column energy filter and K2 Summit direct electron detector (Gatan, Pleasanton, USA). Tilt series were recorded using Serial EM software (https://bio3d.colorado.edu/SerialEM) [98] at a nominal magnification of 64000 × (2.2 Å/pixel), and a target defocus of -3.5 µm. The dose-symmetric recording scheme [99] was applied within an angular range from 60˚ to +60˚, with a starting angle 0˚ and an angular increment of 2˚. The electron dose was set to 1.5 e^{-}/ Å^2 for individual tilt images, corresponding to the total dose of 91.5 e^{-}/ Å^2 for the complete tilt series. A marker-less tilt series alignment was done in IMOD [100], three dimensional CTF correction and weighted backprojection with the voxel size of 4.4 Å were performed using EmSART (https://github.com/uermel/Artiatomi)

[101] provided by Achilleas Frangakis [99, 100].–The reconstructed volumes were denoised using 3D non-linear anisotropic diffusion filter of Etomo of IMOD (k = 1, 15 iterations).

These data allowed visualization of individual nucleosomes *in situ* at a level of detail sufficient to follow the DNA wrapped around and measure the distance between the DNA gyres (see Figure 22). This has in particular revealed a variation of this distance, indicative of a gaping-like conformational variability *in situ*.

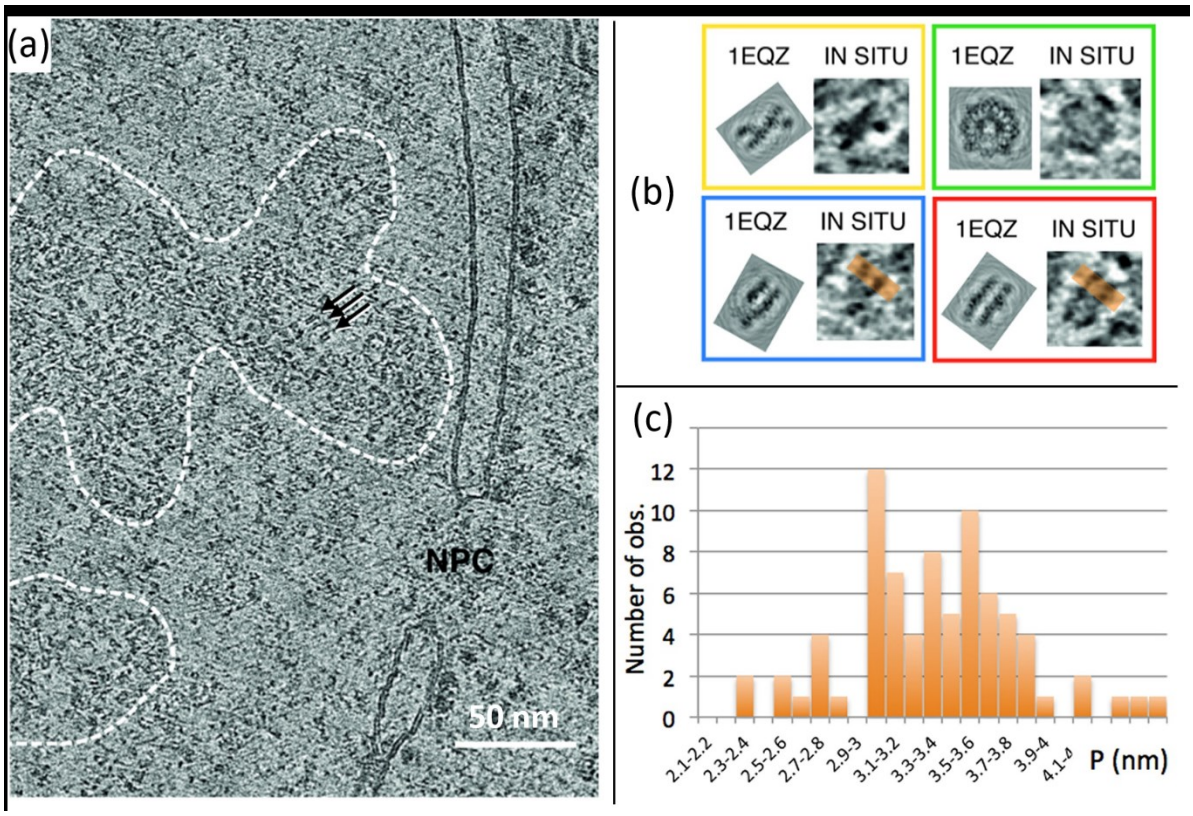


Figure 22 CETOVIS of Drosophila embryonic brain interphase nuclei. (a) A virtual slice of a tomogram (5 nm) shows the nuclear envelope and a nuclear pore complex (NPC); regions enclosed in dashed lines are highly populated in nucleosomes (arrows). (b) Four nucleosome views (3 side views and the top view) identified in the tomogram are compared to cryo-ET images simulated from the crystallographic structure PDB 1EQZ. The distance P between the DNA gyres around the nucleosome can be measured in side views (orange line profiles). (c) Histogram of the distance between the DNA gyres. Adapted from [28].

Chapter 4. Background for the methods developed in this thesis and an experimental dataset of nucleosomes *in situ* analyzed with these methods

In this thesis, two cryo-ET data processing methods for analyzing continuous conformational variability of biomolecular complexes were developed, namely, HEMNMA-3D and TomoFlow, and they were used to analyze a dataset of nucleosomes *in situ*.

HEMNMA-3D is based on matching simulated movements with subtomograms based on Normal Mode Analysis (NMA), whereas TomoFlow extracts such movements from the data based on 3D dense optical flow.

This chapter reviews NMA and optical flow and explains how the dataset for nucleosomes *in situ* was obtained and pre-processed before it was analyzed in terms of conformational variability using HEMNMA-3D and TomoFlow.

Normal Mode Analysis (NMA)

When an atomic structure with N atoms is represented in the Cartesian coordinate system, three coordinates represent every atom in space. The Cartesian representation gives each atom three degrees of freedom and the overall structure 3N degrees of freedom. These degrees of freedom explain how this structure might change the conformation when interacting with its environment. However, the atoms forming a macromolecular structure are connected via chemical bonds; hence, not all these degrees of freedom represent physically feasible motions. In other words, the structure is associated with its energy.

Two methods are commonly used to simulate molecular mechanics, namely, Molecular Dynamics (MD) [102] and Normal Mode Analysis (NMA) [103].

MD simulations are based on exploring different conformations using the Cartesian coordinates of atoms while considering the energy cost caused by the modification of bond strengths, angles, and electrostatic interactions of the atomic structure with its simulated environment. MD simulations can be accurate, but they are computationally demanding, and the simulation setting is a tedious task.

NMA simplifies macromolecular degrees of freedom that describe the motions of the structure by changing the coordinate system to a set of elastic motion degrees called normal modes. This section summarizes the computation of normal modes based on the elastic network model [104, 105]. This model represents the interaction between the atoms as locally connected by elastic springs within a cutoff distance, giving rise to the following energy function.

$$E_p = \sum_{r_{i,j}^0 < R} E_p(r_i, r_j) \tag{4.1}$$

Where R denotes the radius of interaction between the atoms, $r_{i,j} = |r_i - r_j|$ denotes the distance between the atoms *i* and *j*, the zero superscripts indicate the given initial configuration, and $E(r_i, r_j)$ denotes the Hookean pairwise potential between atoms *i* and *j*, and is given by:

$$E_p(r_i, r_j) = \frac{C}{2}\left(r_{i,j} - r_{i,j}^0\right)^2 \tag{4.2}$$

Where C is the bond strength between the connected atoms (spring stiffness constant).

It is evident in eq (4.2) that this model assumes that the potential energy is equal to zero at the initial configuration (substituting $r_{i,j}$ by $r_{i,j}^0$). In other words, this model assumes that the initial structure is given at minimum energy conformation before calculating NMA.

For a conformation $\boldsymbol{r}$, this potential energy can be expanded around its initial conformation $\boldsymbol{r^0}$ as the following Taylor series:

$$\begin{aligned} E_p(\boldsymbol{r}) = E_p(\boldsymbol{r^0}) + \sum_i \left(\frac{\partial E_p}{\partial r_i}\right)^0 (r_i - r_i^0) \\ + \frac{1}{2}\sum_{i,j} \left(\frac{\partial^2 E_p}{\partial r_i \partial r_j}\right)^0 (r_i - r_i^0)(r_j - r_j^0) + \cdots \end{aligned} \tag{4.3}$$

Where superscripts of zero indicate the initial conformation. The first term is zero because it is the potential energy of the initial structure, which is assumed to be zero. The second term is zero because the first derivative of the potential energy is evaluated at the initial conformation, which is assumed to be the minimum of the potential energy.

Hence, an estimate of the potential energy to second-order is the sum of pairwise potentials:

$$E_p(\boldsymbol{r}) \approx \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 E_p}{\partial r_i \partial r_j}\right)^0 (r_i - r_i^0)(r_j - r_j^0) \\ = \frac{1}{2}\sum_{i,j}(r_i - r_i^0)\, H_{i,j}\, (r_j - r_j^0) = \frac{1}{2}\,\Delta\boldsymbol{r}^{\boldsymbol{T}}\boldsymbol{H}\,\Delta\boldsymbol{r} \tag{4.4}$$

Where $\boldsymbol{H}$ is the Hessian matrix obtained from the second derivatives of the potential with respect to the displacements $\boldsymbol{r}$, and $\Delta\boldsymbol{r}$ is the vector of the atomic displacements of the structure relative to the initial conformation.

Now, by substituting eq (4.1) in the second derivative of the potential energy $\left(\frac{\partial^2 E_p}{\partial q_i \partial q_j}\right)$ we get the following derivatives [106]:

$$\frac{\partial^2 E_p}{\partial x_i \partial y_j} = -\frac{C(x_i - x_j)(y_i - y_j)}{{r_{i,j}}^2}, \qquad \frac{\partial^2 E_p}{\partial x_i \partial z_j} = -\frac{C(x_i - x_j)(z_i - z_j)}{{r_{i,j}}^2}, \\ \frac{\partial^2 E_p}{\partial y_i \partial z_j} = -\frac{C(y_i - y_j)(z_i - z_j)}{{r_{i,j}}^2} \tag{4.5}$$

Where x, y and z represent the Cartesian components of the atoms.

Hence, $H_{i,j}$ which is equal to $\left(\frac{\partial^2 E_p}{\partial q_i \partial q_j}\right)^0$ for i not equal to j, and $r_{i,j}^0 < R$:

$$H_{i,j} = -\frac{C}{{r_{i,j}^0}^2}\begin{bmatrix} {x_{i,j}^0}^2 & x_{i,j}^0 y_{i,j}^0 & x_{i,j}^0 z_{i,j}^0 \\ x_{i,j}^0 y_{i,j}^0 & {y_{i,j}^0}^2 & y_{i,j}^0 z_{i,j}^0 \\ x_{i,j}^0 z_{i,j}^0 & y_{i,j}^0 z_{i,j}^0 & {z_{i,j}^0}^2 \end{bmatrix} \tag{4.6}$$

Where $H_{i,j} = 0$ for $r_{i,j}^0 > R$ and the diagonal submatrices $H_{i,i}$ are given by:

$$H_{i,i} = -\sum_{j;\, j \neq i} H_{i,j} \tag{4.7}$$

It is obvious that is of size 3N x 3N, where N is the number of atoms in the structure.

The eigenvectors of the Hessian matrix $\boldsymbol{H}$ are normal mode vectors, and its eigenvalues are squares of normal mode frequencies:

$$\boldsymbol{H}\,\boldsymbol{a_j} = \omega_j{}^2\boldsymbol{a_j} \tag{4.8}$$

Where $\boldsymbol{a_j}$ and $\omega_j{}^2$ are the eigenvectors and eigenvalues of $\boldsymbol{H}$, respectively. Recall that $\boldsymbol{H}$ is symmetrical, which renders its eigenvalues real and positive, and its eigenvectors real and orthogonal [107].

Solving the problem in eq (4.8) is equivalent to the diagonalization of $\boldsymbol{H}$, which is equivalent to setting the cross-products i, j of the second-order term of the potential energy functions to zero and keeping only quadratic terms i, i, i.e., a harmonic approximation of the potential energy function:

$$\boldsymbol{A^T}\,\boldsymbol{H}\,\boldsymbol{A} = \boldsymbol{L} \tag{4.9}$$

Where $\boldsymbol{A} = [\boldsymbol{a_1} \quad \cdots \quad \boldsymbol{a_{3N}}]$ and $\boldsymbol{L} = \begin{pmatrix} \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_{3\mathrm{N}} \end{pmatrix}$

$\boldsymbol{H}$ has six zero eigenvalues in general, i.e., zero frequency modes, corresponding to rigid-body shifts and rotations. Therefore, this reduces the number of normal modes representing non-rigid-body motions to 3N-6.

NMA simulates molecular mechanics for atomic structures. However, it is also possible to perform NMA on EM maps after converting them into a collection of Gaussian functions (pseudoatoms) that describe well the shape of the molecule [108, 109].

The diagonalization of the Hessian matrix is the most computationally demanding part of NMA. In the case of atomic structures, one way to reduce the size of the Hessian is to use the rotation-translation block (RTB) method, which divides the structure into blocks (one or a few consecutive residues per block) whose rotations and translations are considered rather than all degrees of freedom for all atoms [110, 111]. Since the RTB method reduces the basis for Hessian diagonalization, it allows fast computing of normal modes.

Normal modes represent a basis for molecular elastic deformation, i.e., a macromolecular structure at conformation A displaced using a linear combination of given

amplitudes along its M normal modes will reach a conformation B. Hence, one of the common applications of NMA is its usage for the elastic deformation of an existing atomic (or pseudoatomic) structure of one conformation to fit an EM map of a different conformation of the same macromolecule, which is usually known as normal mode flexible fitting (shown in Figure 23) and allows obtaining atomic (or pseudoatomic) resolution models for the EM map [105, 112-114].

We will see an extension of normal mode elastic fitting to subtomographic data in Chapter 5. For more details on NMA and other methods to calculate normal modes, the reader is referred to [106].

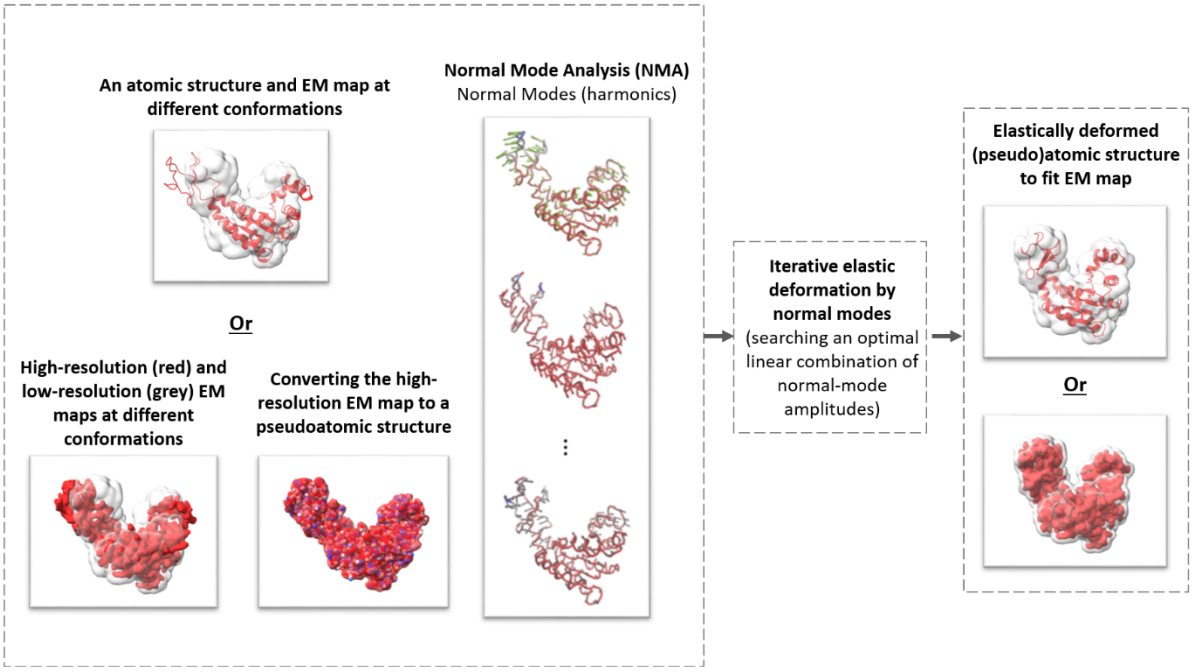


Figure 23 The general scheme of elastic deforming of a reference structure (atomic or pseudoatomic) using normal modes to fit a density map (e.g., an EM map or a subtomogram average).

Optical flow

Optical flows are a family of computer vision algorithms representing movements in image sequences [115]. A typical optical flow algorithm takes as input two video frames and finds the corresponding pixel-to-pixel displacements between the two images. Some optical flow algorithms track the displacements of features that are extracted from the images (features can be edges, corners, etc.) and are called sparse optical flows. Whereas dense optical flow

algorithms, which will be discussed in this chapter and serve as background for one of the methods developed in this thesis (TomoFlow in Chapter 6), find the displacement corresponding to all the pixels between the two images. The term optical flow will refer to dense optical flow hereafter.

Given two images, I_1 and I_2, of moving 3D objects. A typical example for I_1 and I_2 is when they represent two consecutive video frames of a video I, i.e., $I_1 = I(t)$ and $I_2 = I(t + 1)$; however, they can also be two similar enough images where the notion of time is not essential, such as in the problem of deformable image registration [116] (see Figure 24).

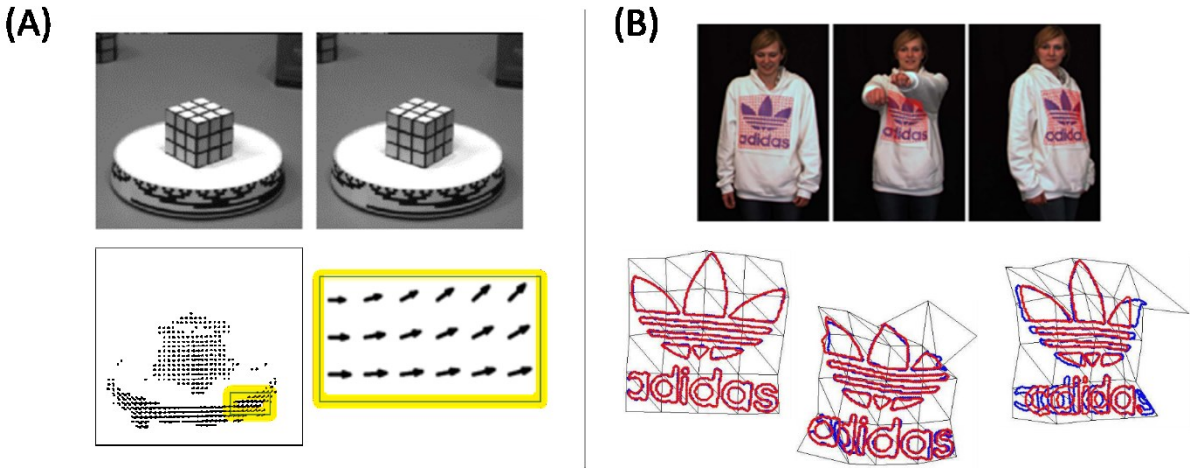


Figure 24 Optical flow between two video frames shows each pixel's displacement (A). An optical flow application on finding an object's deformation in multiple views (B). Adapted from [117, 118].

To find the relationship between the pixels of two images I_1 and I_2, one can start by assuming that the brightness of pixels does not change during the motion between the pixels of the two images, i.e., for a pixel (x, y) in I_1, the same pixel brightness should be found at some distance (u, v) from that pixel in I_2, as follows:

$$I_1(x, y) = I_2(x + u, y + v) \tag{4.10}$$

A second assumption is that pixel displacement is small and can be approximated with the first term of Tailor expansion (one-pixel range), as follows:

$$I(x + u, y + v) \approx I(x, y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v \tag{4.11}$$

Dense optical flow can be defined as the set of (u, v) for each pixel (x, y) between I_1 and I_2 to satisfy (4.10 – 4.11).

Those two assumptions were reasonable to establish a practical computational base for the optical flow. Still, they are limiting factors because brightness values can change during the motion, and the motion can be beyond the one-pixel range. The work of Gunner Farneback in [119] provides solutions for those two problems based on quadratic expansions by i) assuming that the flow field is smooth locally (close pixels move in the same direction) and estimating a displacement vector for each pixel while taking into account its neighborhood helps in minimizing brightness errors and sensitivity to noise; and ii) calculating optical flow iteratively, i.e., an optical flow found at iteration *n* becomes a prior used to estimate optical flow at iteration *n+1*, and on multiple scales (pyramids) of the input images help estimate motion fields that account for larger pixel displacement.

The next section summarizes the work of optical flow calculation proposed by Gunner Farneback in [119].

Estimating optical flow based on quadratic expansions (Farneback optical flow)

Farneback optical flow [119] represents a pixel neighborhood in an image by a quadratic function. A quadratic function is given by:

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\boldsymbol{x} + c \tag{4.12}$$

Where $\boldsymbol{A}$ is a symmetrical matrix, $\boldsymbol{b}$ is a vector, and c is a scalar, and these parameters ($\boldsymbol{A}$, $\boldsymbol{b}$, and c) can be found based on the weighted least square fit of the signal (pixel neighborhood).

Now, given a quadratic expansion f_1 if translated by a displacement d, and the result is f_2, then:

$$\begin{aligned} f_2(\boldsymbol{x}) = f_1(\boldsymbol{x} - \boldsymbol{d}) &= (\boldsymbol{x} - \boldsymbol{d})^T \boldsymbol{A_1}(\boldsymbol{x} - \boldsymbol{d}) + \boldsymbol{b_1}(\boldsymbol{x} - \boldsymbol{d}) + c_1 \\ &= \boldsymbol{x}^T \boldsymbol{A_2}\boldsymbol{x} + \boldsymbol{b_2}\boldsymbol{x} + c_2 \end{aligned} \tag{4.13}$$

Where:

$$\boldsymbol{A_2 = A_1} \tag{4.14}$$

$$\boldsymbol{b_2 = b_1 - 2A_1 d} \tag{4.15}$$

$$c_2 = \boldsymbol{d^T A_1 d - b_1^T d} + c_1 \tag{4.16}$$

Then, the displacement d can be found as:

$$\mathbf{d} = -\frac{1}{2}\boldsymbol{A_1}^{-1}(\boldsymbol{b_2 - b_1}), \qquad \boldsymbol{A_2 = A_1} \tag{4.17}$$

Now in the cases where we have two image neighborhoods, one would first find the quadratic approximations of both images in some neighborhoods, i.e., $\boldsymbol{A_1(x)}$, $\boldsymbol{b_1(x)}$, $c_1(\boldsymbol{x})$, $\boldsymbol{A_2(x)}$, $\boldsymbol{b_2(x)}$, and $c_2(\boldsymbol{x})$. Then, the task is to model the translations $\boldsymbol{d(x)}$ for every pixel.

We can set $\boldsymbol{A(x)}$ as:

$$\boldsymbol{A(x)} = \frac{1}{2}(\boldsymbol{A_1(x) + A_2(x)}) \tag{4.18}$$

Then $\boldsymbol{d(x)}$ can be found by substituting $\boldsymbol{A(x)}$ in eq (4.15) by:

$$\boldsymbol{A(x)\, d(x)} = -\frac{1}{2}(\boldsymbol{b_2(x) - b_1(x)}) = \boldsymbol{\Delta b(x)} \tag{4.19}$$

Where $\boldsymbol{\Delta b(x)} = -\frac{1}{2}\big(\boldsymbol{b_2(x) - b_1(x)}\big)$.

Note that $\boldsymbol{d(x)}$ is a spatially varying displacement field that replaces the global displacement $\boldsymbol{d}$ used in eq (4.13).

In principle, eq (4.19) can be solved pointwise, but further improvement in the accuracy of the resultant displacement field can be achieved when integrating the information over the neighborhood by satisfying eq (4.19) with a solution that minimizes:

$$\sum_{i \in N} w_i \, \|\boldsymbol{A_i \, d(x) - \Delta b_i}\|^2 \tag{4.20}$$

Where i represents the indices of pixels in a neighborhood N, and $\boldsymbol{w}$ represents the weights given for the neighborhood when estimating the displacement for a specific pixel. The

central pixel should have the highest weight, and other pixels are weighted according to their distances.

The solution $\boldsymbol{d}(\boldsymbol{x})$ that minimizes eq (4.20) is given by:

$$\boldsymbol{d}(\boldsymbol{x}) \;=\; (\sum_{i \in N} w_i \, \boldsymbol{A_i}^{\boldsymbol{T}} \boldsymbol{A_i})^{-1} \sum_{i \in N} w_i \, \boldsymbol{A_i}^{\boldsymbol{T}} \Delta \boldsymbol{b_i} \tag{4.21}$$

Now assume that an a priori displacement vector $\widetilde{\boldsymbol{d}}(\boldsymbol{x})$ is known for the displacement between f_1 and f_2. Then eq (4.18 – 4.19) can be updated by:

$$\mathbf{A}(\mathbf{x}) \;=\; \frac{1}{2}(\boldsymbol{A_1}(\boldsymbol{x}) \;+ \boldsymbol{A_2}(\widetilde{\boldsymbol{x}})) \tag{4.22}$$

$$\Delta \boldsymbol{b}(\boldsymbol{x}) \;=\; -\frac{1}{2}(\boldsymbol{b_2}(\widetilde{\boldsymbol{x}}) \;-\; \boldsymbol{b_1}(\boldsymbol{x})) + \boldsymbol{A}(\boldsymbol{x})\widetilde{\boldsymbol{d}}(\boldsymbol{x}) \tag{4.23}$$

Where $\widetilde{\boldsymbol{x}} \;=\; \boldsymbol{x} + \widetilde{\boldsymbol{d}}(\boldsymbol{x})$. Obviously, substituting $\widetilde{\boldsymbol{d}}(\boldsymbol{x})$ with zero in eq (4.22 – 4.23) gives back eq (4.18 – 4.19). This prior displacement field can be used in two ways; the first is to find the displacement field iteratively, starting with $\widetilde{\boldsymbol{d}}(\boldsymbol{x})$ equals zero at the first iteration. And the second is to find the optical flow at multiple scales, with multiple iterations at every scale, starting from the coarsest scale and moving up, by using the optical flow from one scale as a prior at the following scale.

In this thesis, I have used a 3D version of Farneback optical flow, which directly extends the original implementation of 2D Farneback optical flow. The following sections explain the multiscale iterative approach used to find Farneback-3D optical flow and examine its robustness to noise when finding optical flow between noisy EM maps.

Multiresolution pyramidal approach for 3D optical flow calculation

This section describes the multiresolution pyramidal approach for 3D optical flow (OF) calculation with Farneback-3D toolbox (https://pypi.org/project/farneback3d), which is used in TomoFlow (in Chapter 6).

The 3D OF pyramidal approach involves (i) creating a multiresolution volume pyramid by downsampling the volume at each pyramid level (see Figure 25), (ii) calculating OF iteratively at each pyramid level, and (iii) propagating the OF calculated at a coarser level to

the next finer level in order to refine it, until the finest (original volume) level is reached. Between the pyramid scales, the OF propagation is done by upsampling the OF found on a coarser level to the next finer level and applying this upsampled OF onto the reference volume at the finer level to create a warped reference that is then used to find the OF at that finer level.

In Farneback-3D, a Gaussian anti-aliasing filtering is applied to the volume at each pyramid level before the volume is downsampled (the Gaussian standard deviation is adjusted to the scaling factor selected for downsampling). The scaling factor of 0.5 was used in the experiments in this section and in TomoFlow, meaning that each volume dimension was reduced by 2 at each pyramid level. The coarsest volume pyramid level is 32x32x32 voxels, which is the coarsest level allowed by Farneback-3D; also, we used 2-level pyramids for volumes of size 64^3 voxels and 3-level pyramids for volumes of 128^3 voxels. We used a window size of 10 × 10 × 10 voxels for integrating the displacement field over a neighborhood of each voxel and 10 iterations of the algorithm at each pyramid level. All other parameters of Farneback-3D were used with their default values. TomoFlow graphical interface, integrated in ContinuousFlex (in Chapter 7) allows modifying these values.

The OF is first calculated on the coarsest pyramid level (lowest scale) and, then, it is refined on the first finer pyramid level (larger scale), followed by the refinement on the next one etc., until the refinement on the finest pyramid level (original scale, i.e., the input volumes). For each pyramid level, the OF is calculated iteratively. In each iteration, the calculated OF is applied to the reference volume to warp it; this warped reference is then used to find the OF in the next iteration and produce the reference for the following iteration, etc., until the convergence is achieved (the OF between two successive iterations does not change significantly).

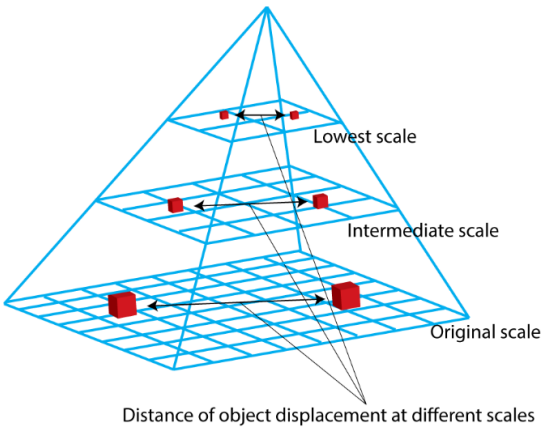


Figure 25 Multiresolution data pyramid scheme.

The robustness of Farneback-3D to noise

In this section, we show the performance of the Farneback 3D optical flow (OF) method in matching different conformational variability magnitudes, challenged by noise only. Also, we provide a quantitative assessment of the algorithm for mapping conformations while disentangling it from the subtomographic-approach limitations such as missing wedge and rigid-body (angular and shift) variability. The PDB:4AKE chain A structure (obtained at 2.2 Å resolution by X-ray crystallography and referred here to as AK) was used to synthesize three conformations by elastic deforming AK using its normal mode 7 and gradually increasing the amplitude of the mode. The three synthesized conformations were then converted into volumes (volume size: 128^3 voxels; voxel size: 1 $Å^3$) and noise was applied directly onto these volumes (without low-pass filtering of the volumes or synthesizing tilt series and calculating 3D reconstructions).

The conformational distance of each of the three synthetic conformations is reflected by the selected amplitude of normal mode 7. The following three values of the amplitude were used: 1) -75 (the structure referred to as AK_75), 2) -125 (the structure referred to as AK_125), and 3) -200 (the structure referred to as AK_200). The four atomic structures (AK, AK_75, AK_125 and AK_200) are shown in Figure 26. These structures converted into volumes are shown in Figure 27.

The root mean square deviations (RMSDs) of the AK_75, AK_125 and AK_200 structures with respect to the AK structure are shown in Table 1, along with the cross-correlations (CC) between the AK volume and each of the AK_75, AK_125 and AK_200 volumes.

Random Gaussian noise was added to each of the AK_75, AK_125 and AK_200 volumes in such a way to obtain the following 6 values of the signal-to-noise ratio (SNR): 1) 0.5, 2) 0.1, 3) 0.05, 4) 0.01, 5) 0.005, and 6) 0.001. In Figure 28, we show the different SNR values of the volumes using central slices of the noisy AK_125 volumes as an example.

The OF was calculated using Farneback-3D with a 3-level volume pyramid of a scaling factor of 0.5 (meaning a pyramid with the levels of 128^3, 64^3 and 32^3 voxels for the test datasets analyzed in this section, where 32^3 voxels is the coarsest pyramid level allowed by Farneback-3D).

We calculated the OFs between the non-noisy AK volume and the noisy AK_75, AK_125 and AK_200 volumes (6 SNR values for each of AK_75, AK_125 and AK_200). Each OF was used to warp the AK volume. The obtained warped AK volumes are the non-noisy estimates of the noisy AK_75, AK_125 and AK_200 volumes and are called “matched” volumes (the term introduced in the TomoFlow, see Chapter 6).

The “matched” volumes for AK_75, AK_125 and AK_200 are shown Figure 29, Figure 30, and Figure 31, respectively. The CCs between the non-noisy versions of each of the AK_75, AK_125 and AK_200 volumes and the corresponding “matched” volumes are presented in Table 2.

The visual comparison in Figure 29, Figure 30, and Figure 31, and the corresponding results in Table 2 indicate that the matched volume approached the conformation in all different noisy volumes (the CC between the matched volume and each of the non-noisy versions of the AK_75, AK_125 and AK_200 volumes is always higher than the original cross correlation before the matching). Better results were obtained for smaller magnitudes of the conformational change and lower noise levels.

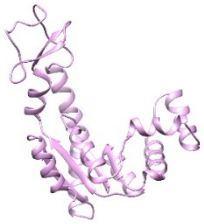
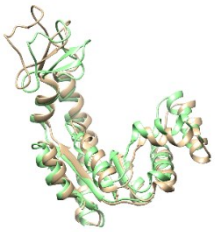
AK	AK_75	AK_125	AK_200
			
AK + AK_75 + AK_125 + AK_200	AK + AK_75	AK+ AK_125	AK + AK_200
			

Figure 26 Atomic structures used in the experiment. See the text in this section for details on how they were obtained.

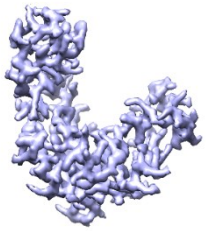
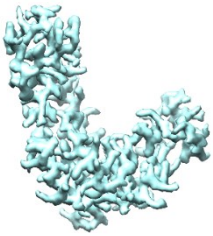
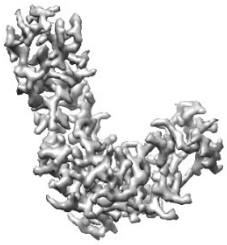
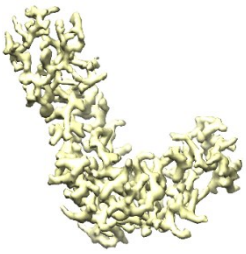
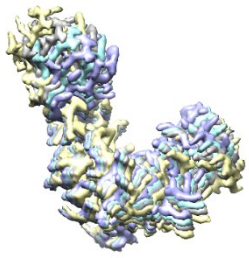
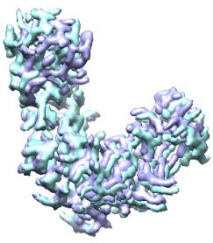
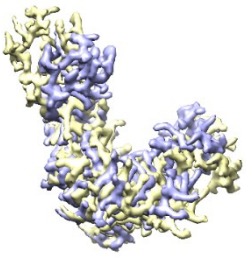
AK	AK_75	AK_125	AK_200
			
AK + AK_75 + AK_125 + AK_200	AK + AK_75	AK+ AK_125	AK + AK_200
			

Figure 27: Volumes used in the experiment. See the text in this section for details on how they were obtained.

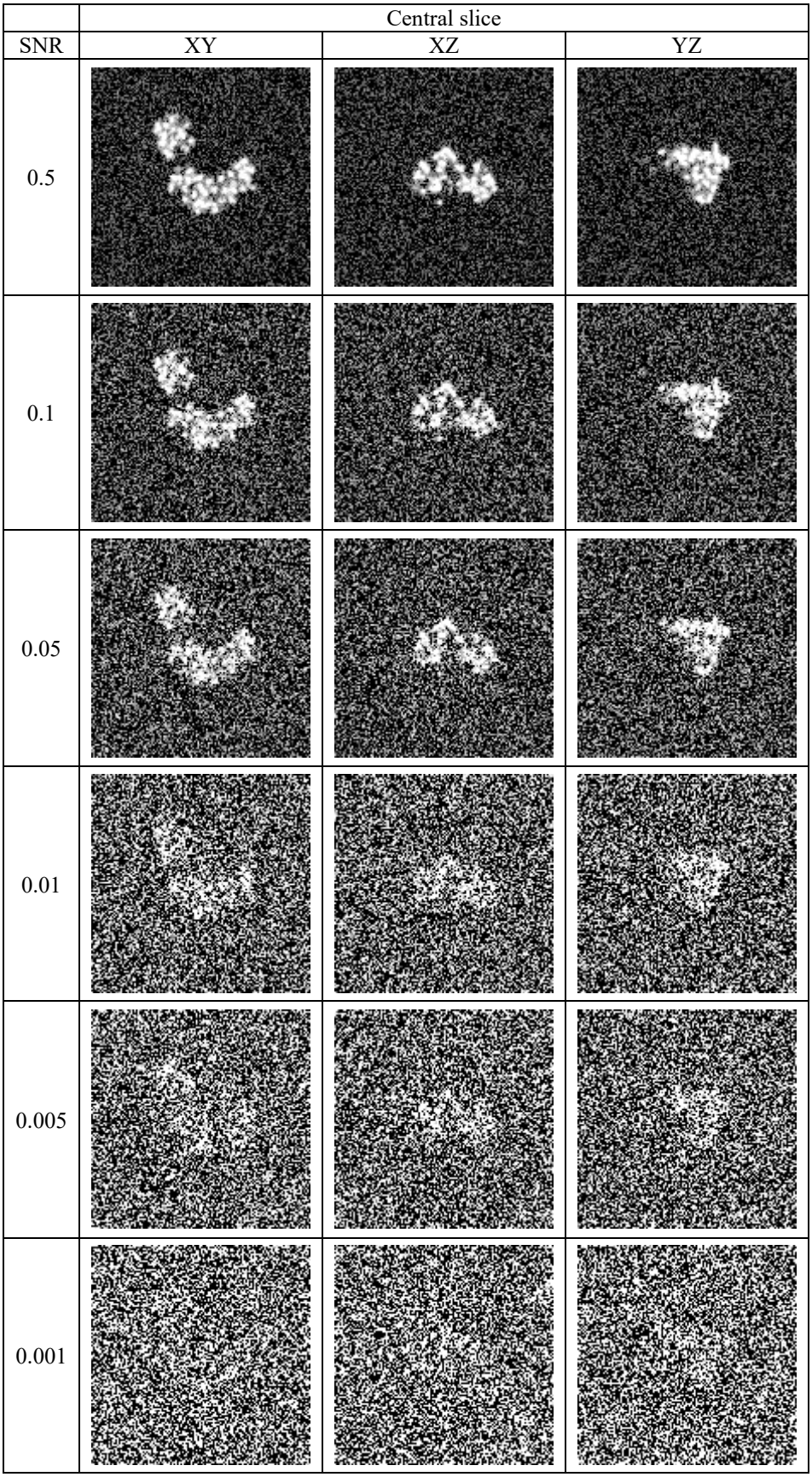
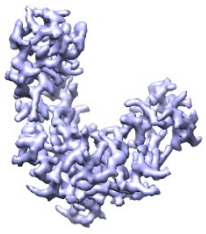
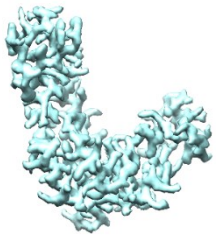
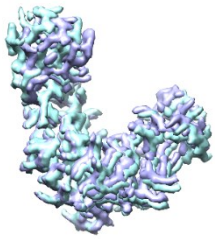


Figure 28 Central slices of the volume AK125 at different values of the signal-to-noise ratio (SNR).

	AK	AK_75	AK + AK_75
Before OF matching			

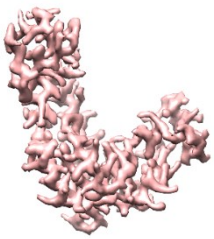
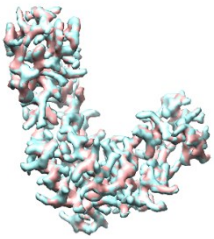
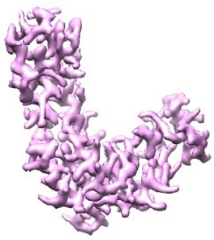
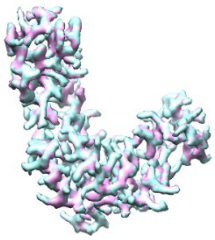
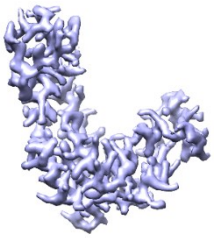
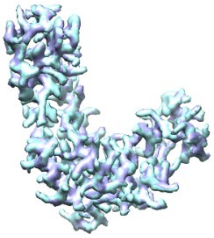
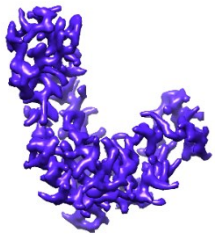
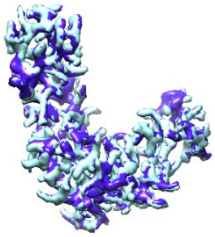
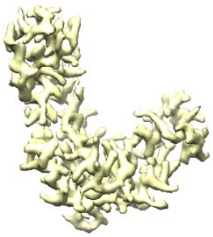
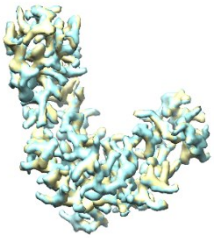
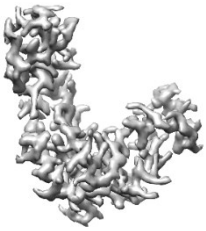
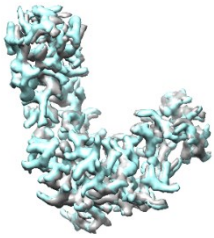
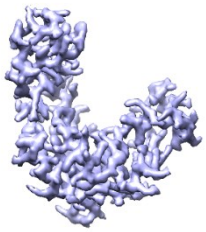
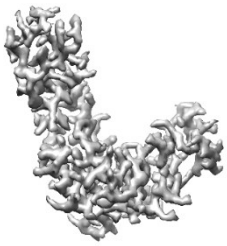
After OF matching			
Matched SNR 0.5	Matched SNR 0.5 + AK_75	Matched SNR 0.1	Matched SNR 0.1 + AK_75
			
Matched SNR 0.05	Matched SNR 0.05 + AK_75	Matched SNR 0.01	Matched SNR 0.01 + AK_75
			
Matched SNR 0.005	Matched SNR 0.005 + AK_75	Matched SNR 0.001	Matched SNR 0.001 + AK_75
			

Figure 29 OF-based matching of the non-noisy AK volume to different noisy versions of the AK_75 volume.

	AK	AK_125	AK + AK_125
Before OF matching			

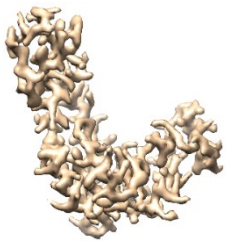
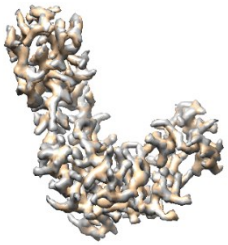
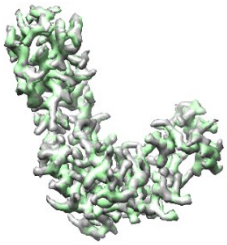
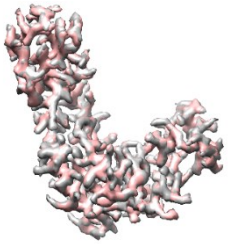
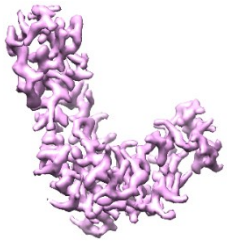
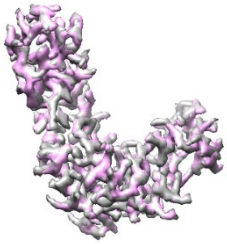
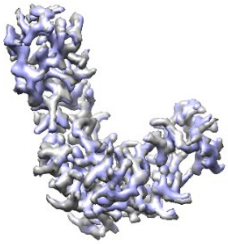
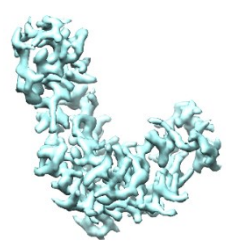
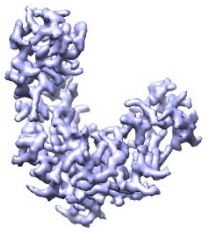
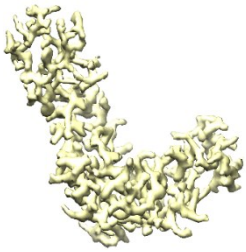
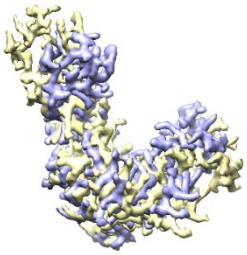
After OF matching			
Matched SNR 0.5	Matched SNR 0.5 + AK_125	Matched SNR 0.1	Matched SNR 0.1 + AK_125
			
Matched SNR 0.05	Matched SNR 0.05 + AK_125	Matched SNR 0.01	Matched SNR 0.01 + AK_125
			
Matched SNR 0.005	Matched SNR 0.005 + AK_125	Matched SNR 0.001	Matched SNR 0.001 + AK_125
			

Figure 30 OF-based matching of the non-noisy AK volume to different noisy versions of the AK_125 volume.

	AK	AK_200	AK + AK_200
Before OF matching			

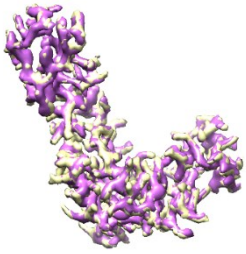
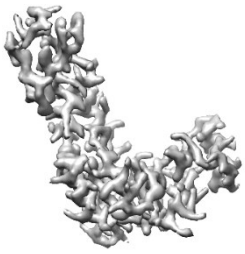
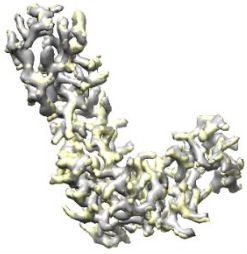
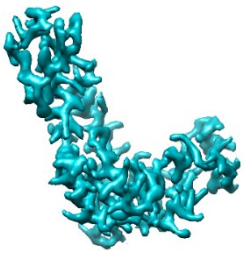
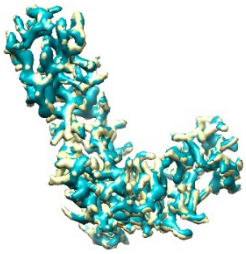
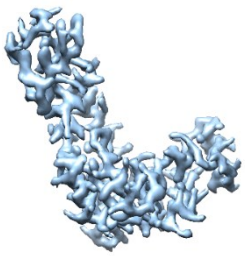
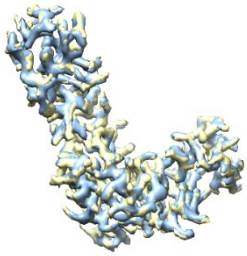
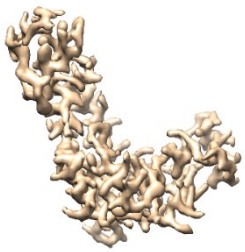
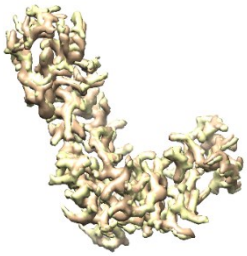
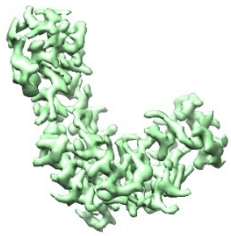
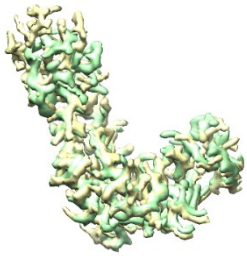
After OF matching			
Matched SNR 0.5	Matched SNR 0.5 + AK_200	Matched SNR 0.1	Matched SNR 0.1 + AK_200
			
Matched SNR 0.05	Matched SNR 0.05 + AK_200	Matched SNR 0.01	Matched SNR 0.01 + AK_200
			
Matched SNR 0.005	Matched SNR 0.005 + AK_200	Matched SNR 0.001	Matched SNR 0.001 + AK_200
			

Figure 31 OF-based matching of the non-noisy AK volume to different noisy versions of the AK_200 volume.

Table 1 Quantitively measure of the difference between the AK conformation and each of the three synthetic conformations used here (AK_75, AK_125 and AK_200), expressed in terms of the root mean square deviation (RMSD) between the atomic structures and in terms of the cross-correlation (CC) between the volumes from these atomic structures. Note here that the CC in this table is calculated for non-noisy volumes.

	AK_75	AK_125	AK_200
RMSD compared to AK [Å]	1.8431	3.0719	4.915
CC with AK	83.78%	73.27%	61.59%

Table 2 Cross-correlation between the "matched" AK volume and the non-noisy AK_75, AK_125 and AK_200 volumes. The AK volume "matching" was done with respect to different noisy versions of the AK_75, AK_125 and AK_200 volumes.

	SNR 0.001	SNR 0.005	SNR 0.01	SNR 0.05	SNR 0.1	SNR 0.5
AK_75	94.75%	98.18%	98.50%	98.96%	99.01%	99.06%
AK_125	92.71%	97.23%	97.90%	98.52%	98.65%	98.65%
AK_200	86.23%	95.27%	96.56%	97.46%	97.53%	97.63%

An experimental dataset used to analyze nucleosomes *in situ*

The Drosophila embryo cryo-sample preparation, vitreous sectioning, tilt series acquisition, and tomogram reconstruction were performed as described in Chapter 3 and in [28].

The dataset preprocessing described in this section was provided by Dr. Mikhail Eltsov (IGBMC, Strasbourg University, Strasbourg).

A slice of the experimental nucleosome tomographic data is shown in Figure 32.

Nucleosomes were manually picked in IMOD. Then, 64^3 voxel subtomograms (voxel size of 4.4 Å) were extracted from the original non-denoised volumes. To refine manually-picked nucleosome coordinates, subtomogram alignment and averaging were performed with SubTomogramAveraging (https://github.com/uermel-/Artiatomi) script using a sum of the randomly rotated subtomograms as an initial reference.

Alignment of subtomograms was performed in two steps. Initially, a bandpass filter was applied with a low cutoff frequency of 3 reciprocal-space pixels, a high cutoff frequency of 8 reciprocal-space pixels, and a Gaussian edge smoothing with a standard deviation of 3 reciprocal-space pixels. Ten iterations of an unconstrained rotational search (three rotational degrees of freedom) were performed with an angular sampling step of 10, and a translational search (three translational degrees of freedom) was performed within a radius of 5 real-space pixels. In the second step, a bandpass filter was applied with low and high cutoff frequencies

of 15 reciprocal-space pixels and 3 reciprocal-space pixels, respectively, and a Gaussian edge smoothing with a standard deviation of 3 reciprocal-space pixels. At that step, 20 iterations of the rotational search were performed with an angular sampling step of 2, constrained to 20° around the orientation found in the previous step, and the translational search radius was reduced to 3 real-space pixels. The cross-correlation between the last several iteration averages (0.994) indicated the stabilization of the subtomogram alignment. A new set of subtomograms of the same dimensions and voxel size was extracted at the refined nucleosome positions were exported to be analyzed by HEMNMA-3D and TomoFlow (in Chapters 5 and 6). The nucleosome data used in this thesis have been deposited in EMPIAR and EMDB databases under the accession codes EMPIAR-10679 and EMD-12699, respectively.

The nucleosome conformational variability detected in this dataset in previous works [28, 120] was mainly described as gapping and breathing motions of the nucleosome [90]. However, these conformational variabilities were previously identified only via manual measurements [28].

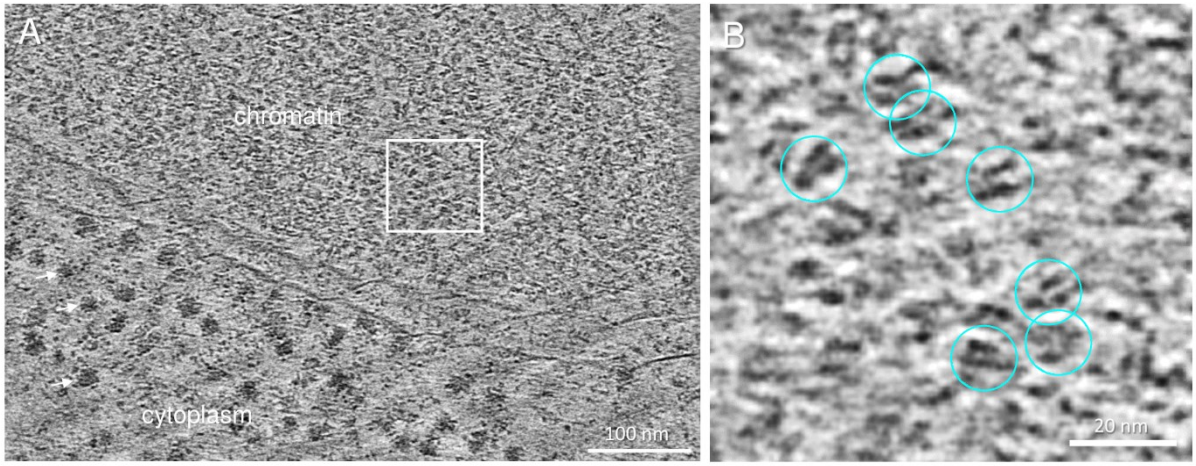


Figure 32 A slice of the experimental nucleosome tomographic data: (A) A 5-nm thick slice through a tomographic reconstruction showing an area of compact chromatin at a nuclear periphery (chromatin) that is easily distinguished from cytoplasm filled with ribosomes (arrows). (B) An enlargement of the chromatin area is outlined with a white square in (A). Circles indicate positions of nucleosomes selected for subtomogram extraction.

Chapter 5. HEMNMA-3D: Cryo-ET data processing method based on NMA to analyze continuous conformational variability of biomolecular complexes

This chapter presents HEMNMA-3D, the first method for analyzing cryo-electron subtomograms in terms of continuous conformational changes of complexes.

HEMNMA-3D combines elastic and rigid-body 3D-to-3D iterative alignments of a flexible 3D reference (atomic structure or electron microscopy density map) to match the conformation, orientation, and position of the complex in each subtomogram.

The elastic matching combines molecular mechanics simulation (Normal Mode Analysis of the 3D reference) and experimental, subtomogram data analysis. The rigid-body alignment includes compensation for the missing wedge due to the limited tilt angle of cryo-ET. The conformational parameters (amplitudes of normal modes) of the complexes in subtomograms obtained through the alignment are processed to visualize the distribution of conformations in a space of lower dimension (typically, 2D or 3D), referred to as the space of conformations. This allows a visually interpretable insight into the dynamics of the complexes, by calculating 3D averages of subtomograms with similar conformations from selected (densest) regions and by recording movies of the 3D reference's displacement along selected trajectories through the densest regions.

HEMNMA-3D was published in 2021 in two manuscripts. The first manuscript [120] describes HEMNMA-3D and shows its validation using synthetic datasets and its application to the experimental dataset describing *in situ* nucleosome conformational variability (presented in Chapter 4). The second manuscript [121] (a more detailed version available on bioRxiv [122]) compares HEMNMA-3D to conventional methods of subtomogram averaging and classification on a synthetic dataset of nucleosome conformational variability.

This chapter presents the method and results originally published in the mentioned articles.

HEMNMA-3D method

A graphical summary of HEMNMA-3D is presented in Figure 33. The flowchart in Figure 34 describes the workflow of the proposed method, which was inspired by the workflow of HEMNMA [43, 123].

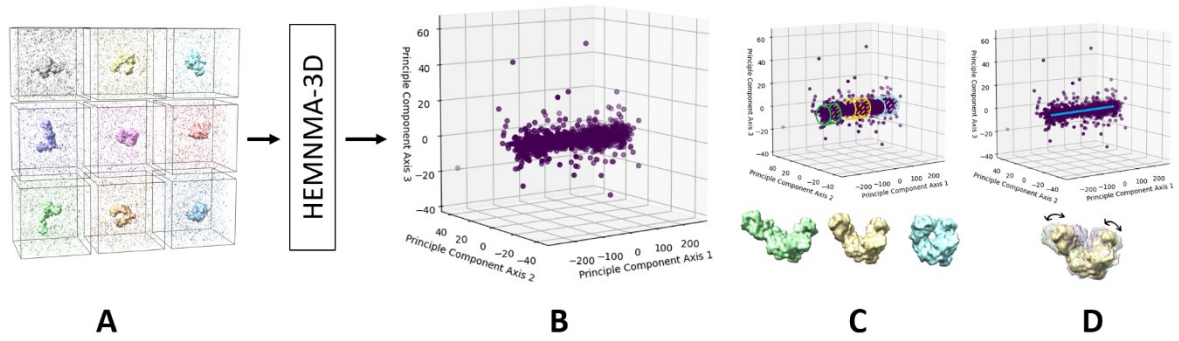


Figure 33 A graphical summary of the data flow of HEMNMA-3D. (A) Input subtomograms containing the same biomolecule but at different orientations, positions, and conformations (here represented with a low noise level for illustration). (B) Input subtomograms projected onto a low-dimensional "space of conformations," describing and visualizing the biomolecular conformational variability contained in the subtomograms. (C) Grouping of close points (subtomograms with similar biomolecular conformations) and averaging of subtomograms in these groups. (D) Animating biomolecular motion along trajectories identified in the densest regions.

The workflow of HEMNNA-3D comprises the following steps:

(1) Input: the input to the method are a reference structure and a set of subtomograms. In the case where the reference structure is a density map (a 3D volume such as an EM map or a subtomogram average), a conversion to 3D Gaussian functions (pseudoatoms) takes place.

(2) Normal mode analysis of the reference atomic structure or the reference pseudoatomic structure (obtained by converting the reference density map into 3D Gaussian functions in the previous step).

(3) Combined iterative elastic and rigid-body 3D-to-3D alignment of the reference structure with the input subtomograms, with missing wedge compensation.

(4) Visualization of computed conformations.

In the remaining part of this section, we describe these steps in more detail. Please note that the first two steps of the workflow are exactly as those of HEMNMA and were thoroughly presented, tested and discussed in previously published works on HEMNMA, its tools and applications [43, 108, 109, 123, 124], one of which I contributed [123]. Here, we recall their basic principles.

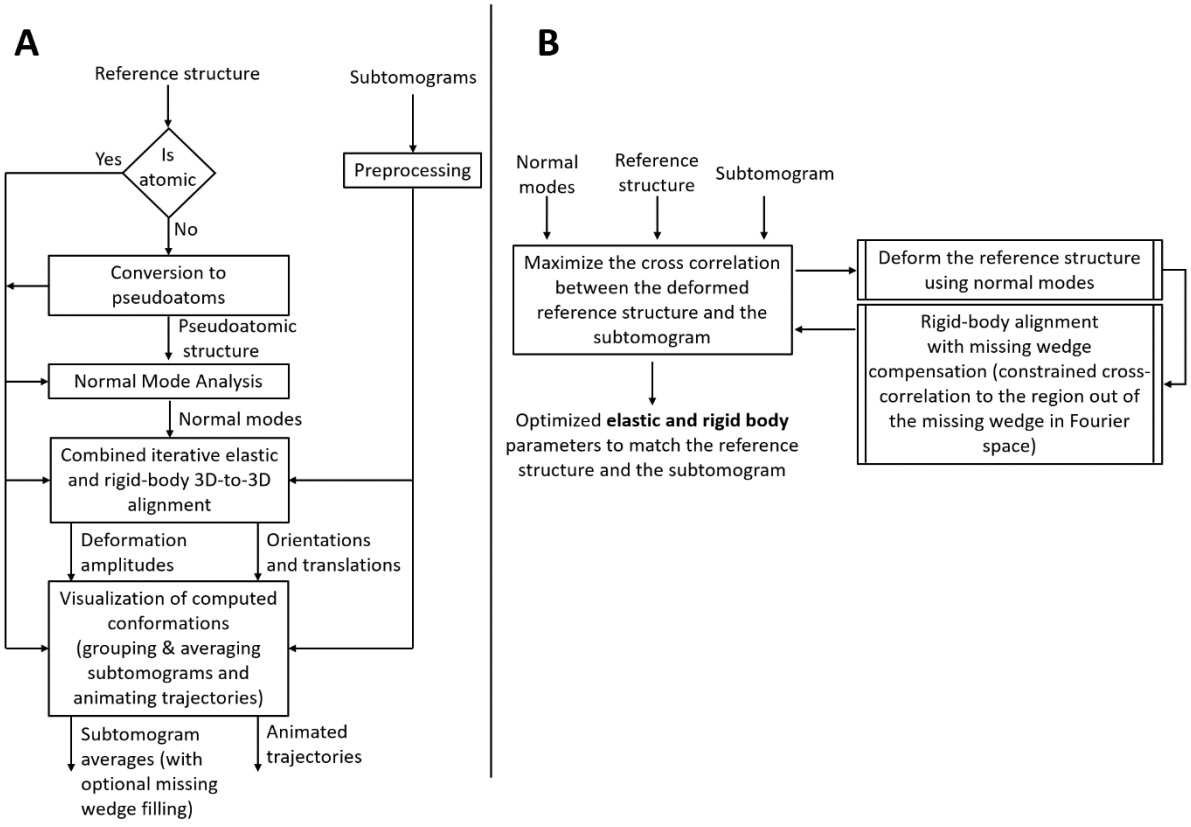


Figure 34 Flowchart of HEMNMA-3D. (A) Workflow. (B) Combined iterative elastic and rigid-body 3D-to-3D alignment step (the core module of HEMNMA-3D).

Input reference and conversion of reference density maps into pseudoatoms

A reference structure of the molecule targeted in the subtomograms can be used in the form of an atomic model (PDB formatted files) or a density map, such as an EM map (SPA reconstruction) or a subtomogram average (obtained using classical StA without considering conformational heterogeneity). Although our method can be used with both atomic and density-map reference structures, one should prefer using a reference density map from the data at hand, if it can be obtained.

If a reference density map is used, it must be converted into a collection of Gaussian functions (pseudoatoms) with a carefully selected standard deviation (pseudoatom size, whose

default value is 1 voxel [108, 109]). The pseudoatom size should lead to a structure (called pseudoatomic structure) that, converted back to a density map, approximates the input density map with a small error (given a target approximation error, whose default value is 5%).

Optionally, a mask on the density map can be used prior to the conversion into pseudoatoms (e.g., a spherical binary mask of a given radius) to reduce background noise. Such masks may also be useful if applied on input cryo-ET subtomograms to maximize the chance of having a single molecular complex in each subtomogram (Preprocessing block in the workflow in Figure 34A.

Normal mode analysis

This step involves computing normal modes of a reference atomic or pseudoatomic structure for the 3D-to-3D elastic alignment in the next step. The computation of normal modes is based on the elastic network model [104, 105] by representing the interaction between the (pseudo-)atoms as if they are locally connected by elastic springs (within a cutoff distance). Normal Mode Analysis requires the diagonalization of a 3N * 3N matrix of second derivatives of the potential energy (Hessian matrix), where N is the number of nodes in the elastic network model determined by the total number of atoms (or pseudoatoms) in the input reference. In the case of atomic structures, we use the rotation-translation block (RTB) method, which divides the structure into blocks (one or a few consecutive residues per block) whose rotations and translations are considered rather than all degrees of freedom for all atoms [110, 111]. Since the RTB method reduces the basis for Hessian diagonalization, it allows fast computing of normal modes. Since pseudoatomic structures usually contain fewer nodes (pseudoatoms) than atomic structures, normal modes can be obtained by a direct diagonalization of the 3N * 3N Hessian, which is referred to as the Cartesian method. Larger values of the interaction cutoff distance (the distance below which atoms or pseudoatoms do not interact) lead to more rigid motions. The atomic interaction cutoff distance may be set manually (by default 8 A) and the pseudoatomic cutoff distance is recommended to be computed automatically based on the distribution of the pseudoatomic pairwise distances (e.g., as the value below which is a given percentage of all distances as in [43, 108, 109, 123]. The modes are computed along with their respective collectivity degrees, which count the number of atoms or pseudoatoms affected by the mode as in [125].

To allow faster data analysis and avoid noise overfitting in the 3D-to-3D elastic alignment in the next step, we select a subset of normal modes (usually, less than 10) with the lowest frequencies and highest collectivities, as previously described [43, 123, 124].

Low-frequency high-collectivity normal modes have been shown to be relevant to functional conformational changes [126-130]. The first six (lowest-frequency) normal modes are related to rigid-body transformations and are thus not used for the 3D-to-3D elastic alignment in the next step. The rigid-body 3D-to-3D alignment is done without using these rigid-body normal modes, as explained in the next paragraph.

Combined iterative elastic and rigid-body 3D-to-3D alignment

This step, represented in Figure 34B, is the backbone of the proposed method. It has been inspired by the combined iterative elastic and rigid-body 3D-to-3D alignment step of StructMap method [131], which was proposed for pairwise similarity analysis of SPA high-resolution EM maps (no missing wedge). In HEMNMA-3D proposed here, this step comprises simultaneous NMA-based elastic alignment (search for amplitudes of a linear combination of normal modes) and rigid-body alignment (search for orientation and position, meaning 3 Euler angles and x, y, and z shifts) of the reference structure with each given subtomogram. It refines the amplitudes of displacement along each used normal mode (elastic parameters) as well as the angles and shifts (rigid-body parameters) of the reference structure until the best match is obtained between this reference structure and the given subtomogram. The latter is achieved by maximizing the similarity between the subtomogram and the density volume from the elastically deformed, oriented and shifted reference, and includes missing wedge compensation. The missing wedge compensation is done by calculating the cross-correlation between the reference and subtomogram density maps only in the region of the Fourier space where the data can be trusted, i.e., by constraining the cross-correlation evaluation to the Fourier space region that excludes the missing wedge region (the region outside of the one specified by the tilt angle range, e.g., -60˚ to +60˚. To maximize this constrained cross-correlation (CCC), we use a variant of Powell's UOBYQA method, which subjects the objective function to a trust-region radius [132]. For each subtomogram, the normal mode amplitudes are initiated with zeros, meaning that the non-deformed reference is used in the first iteration. As the iterations evolve, the reference model is displaced with the new guesses of the normal mode displacement amplitudes, converted into a volume and rigid-body aligned with the subtomogram using the method of fast rotational matching [48]. At the end of each iteration, the CCC is found and fed

to the numerical optimizer [132]. The iterations repeat until the final value of the trust-region radius or the maximum number of iterations is reached.

Visualizing and utilizing the space of conformations

The number of elastic alignment parameters (normal mode amplitudes) is determined by the number of selected normal modes for the 3D-to-3D elastic alignment. The ensemble of normal mode amplitudes (for all subtomograms) can be projected onto a lower-dimensional space, so-called conformational space, using a dimensionality reduction technique. Here, we use linear Principal Component Analysis (PCA) as it is the most widely known and intuitively clear dimensionality reduction method, but other dimension reduction methods could also be used (linear or nonlinear). The dimensionality reduction is usually performed to two or three dimensions, which allows a global data display and easier modeling of conformational changes. Each point in the conformational space represents a subtomogram and close points correspond to similar conformations in the subtomograms. The points that differ significantly from the remaining observations (too isolated, outlier points) may be excluded from further analysis by excluding the points below a certain p-value based on the Mahalanobis distance (the distance between each point and the whole distribution) [133]. The space of conformations can then be analyzed to reveal molecular dynamics. This can be done by averaging subtomograms of similar conformations in the densest regions of the conformational space or by exploring the densest regions by fitting curves (approximation by line segments) through the data and displacing the reference structure along these curves (referred to as trajectories) to animate the motion along them.

Averaging subtomograms of similar conformations

Close points in the conformational space can be grouped, which results in grouping subtomograms of similar conformations and averaging them. Before computing group averages, the rigid-body alignment parameters found along with the 3D-to-3D elastic alignment are applied to the subtomograms. Optionally, before computing group averages, the missing-wedge Fourier space region of individual subtomograms may be filled in with the corresponding region of the global average computed from all subtomograms. A similar procedure of missing wedge filling of individual subtomograms is used in EMAN2 software package [134]. The subtomogram averages obtained from the selected groups of subtomograms

can be overlapped and compared to understand the conformational changes of the complex in the given set of subtomograms.

Animating motions (trajectories)

Distinct trajectories can be determined through the data in the conformational space, and animated to see the motion of the biomolecule while it is displaced along the trajectory. To animate a trajectory, several points (e.g., 10) along the trajectory should be mapped back to the original displacement space (e.g., using inverse PCA), resulting in elastic alignment parameters that can be used to deform the reference atomic or pseudoatomic structure. Concatenating and displaying the resulting structures can show a movie-like animation of the reference biomolecule traveling across the specified trajectory.

Results and discussion

In this section, we present and discuss the results of HEMNMA-3D with synthetic and experimental subtomograms.

Synthesizing datasets for testing the method performance

For testing HEMNMA-3D in general, and the combined elastic and rigid-body 3D-to-3D alignment module in particular (which is the core module of the proposed method), we synthesized two datasets of conformationally heterogeneous subtomograms that mimic discrete and continuous conformational variability, called "Discrete" and "Continuous" datasets respectively. The flowchart for the data generation procedure is shown in Figure 35 and is detailed in the following.

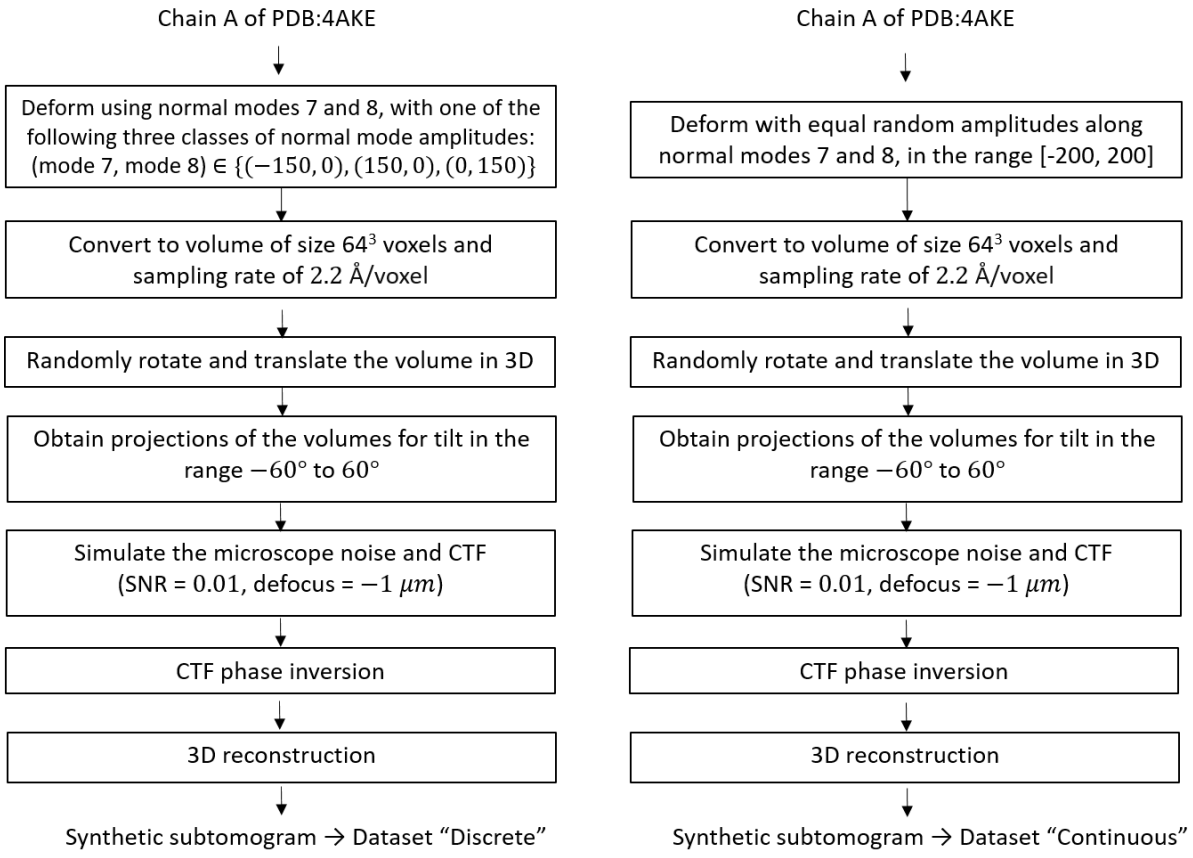


Figure 35 Flowcharts of synthesis of the datasets used for testing and validating HEMNMA-3D, namely "Discrete" dataset (left) and "Continuous" dataset (right).

The "Discrete" dataset comprises 900 synthetic subtomograms representing three different (synthetic) conformations of the atomic PDB:4AKE structure [135] of adenylate kinase chain A (1656 atoms), i.e., 300 subtomograms per conformation. We generated this dataset using the atomic PDB:4AKE structure and its first two non-rigid-body normal modes, i.e., modes 7 and 8 (note here that the mode number corresponds to the frequency of the mode and that higher numbers correspond to higher frequencies). Precisely, the three conformations are represented by the following amplitudes of modes 7 and 8: (mode 7, mode 8) ∈ {(-150, 0), (+150, 0), (0, +150)}.

The "Continuous" dataset comprises 1000 synthetic subtomograms representing a continuum of conformations of the same PDB:4AKE structure. We generated this dataset using this atomic structure and its modes 7 and 8 using a linear relationship between the amplitudes of the two modes. More precisely, the synthesized amplitudes of modes 7 and 8 were identical and randomly distributed in the range [-200, +200] (uniform distribution).

To generate a subtomogram, first, we deform the atomic structure using appropriate amplitudes for the selected normal modes depending on the dataset in hand, i.e., we use (mode 7, mode 8) = (+150, 0) or (-150, 0) or (0, +150) to create a subtomogram in the "Discrete" dataset, while we assign a random value in the range [-200, 200] for both mode 7 and mode 8 to generate a subtomogram in the "Continuous" dataset. Then, we convert the deformed structure to a volume of size 64^3 voxels and a voxel size of 2.2 Å [136]. Afterwards, we rotate and shift this volume in 3D space using random Euler angles and random x, y, z shifts, and we project the rotated and shifted volume using tilt values -60˚ to +60˚ to obtain a tilt series. We simulate microscope conditions by adding noise, modulating the images with the contrast transfer function (CTF) of the microscope (using the defocus of -1 μm), then adding noise again in such a way that a part of the noise will be modulated by the CTF, and the other part will not with a total SNR = 0.01.

Finally, we reconstruct a volume (our synthetic subtomogram) from the tilt series using a Fourier reconstruction method [20]. A few examples of the synthesized subtomograms (SNR = 0.01) and their less noisy version (SNR = 0.5, for illustration) is presented in Figure 36.

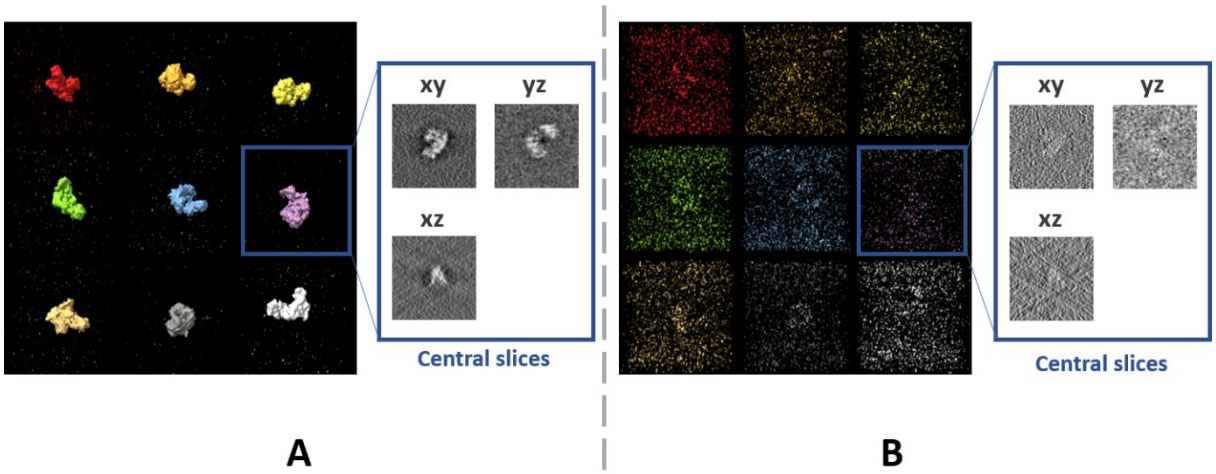


Figure 36 Examples of synthetic subtomograms containing the same molecule but at different orientations, positions and conformations for two different noise levels. (A) Low level of noise (SNR = 0.5). (B) High level of noise (SNR = 0.01).

Synthetic discrete-type conformational variability

In this experiment, our goal is to retrieve the ground-truth amplitudes of normal modes 7 and 8 by the combined elastic and rigid-body alignment (the core module of HEMNMA-3D) of a reference model with the subtomograms in the "Discrete" dataset. In other words, the goal

is to find a solution for the challenging inverse problem of finding the conformation of the structure in each subtomogram. Since the proposed method can use two choices for the reference model, namely, an atomic structure and a density map (e.g., an EM map or a subtomogram average), we performed two types of tests. In the first test type, the atomic structure used to generate the synthetic subtomograms (chain A of the PDB:4AKE) was used as a reference for retrieving normal mode amplitudes of the synthetic subtomograms. In the second test type, we converted the atomic structure into a density map (volume) of size 128^3 voxels and voxel size of 1 $Å^3$ and we used this density map as a reference for retrieving normal mode amplitudes of the synthetic subtomograms. In the case of the reference density map, normal modes were computed from the corresponding structure obtained by converting the density map into pseudoatoms (1675 pseudoatoms for the given pseudoatom radius of 1.25 voxels and the target approximation error of 5%). In both cases (reference atomic structure and reference pseudoatomic structure, with their corresponding normal modes), we used three modes (modes 7, 8 and 9) instead of only two modes (modes 7 and 8 that were used to generate synthetic subtomograms), to make the 3D-to-3D elastic and rigid-body alignment task even more challenging.

Figure 37 presents the estimated amplitudes of normal modes 7 and 8 (the estimated amplitude of normal mode 9 is close to 0 and is therefore not shown graphically). Table 3 presents the mean absolute error and the standard deviation between the estimated and ground-truth normal-mode amplitudes. In both test cases, the three distinct synthetic groups of subtomograms are correctly separated, considering the extreme noise level. The results show a less accurate alignment in the second case, which is expected since, in that case, the atomic structure was used to generate the dataset and the pseudoatomic structure was used as the reference model for the method to estimate the normal-mode amplitudes from this generated dataset. This is in contrast to the first test case where the same atomic structure was used to create the dataset and as the reference for the method to estimate the normal-mode amplitudes from this dataset. Figure 38 shows grouping and averaging the subtomograms in the first test type (atomic reference). We compared the obtained subtomogram averages with the corresponding ground-truth subtomograms. The visual comparison shows no significant difference between them, and the cross-correlation values vary between 97-98%.

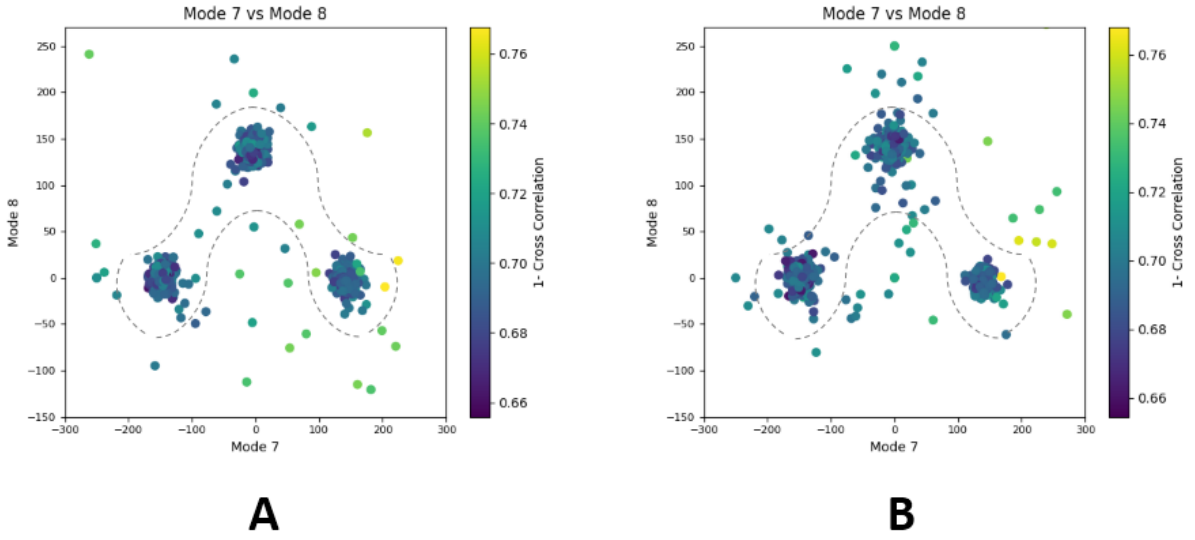


Figure 37 Plots showing the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with "Discrete" dataset (synthetic subtomograms are simulating discrete conformational heterogeneity). (A) Use of the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. (B) Use of a pseudoatomic structure (from a simulated density map) and its normal modes to estimate the conformational and rigid-body parameters of the molecules in the input synthetic subtomograms. The goal was the retrieval of the ground-truth relationship between the amplitudes along normal modes 7 and 8; ideally, all data should lay in one of the following three clusters of normal-mode amplitudes: (mode 7, mode 8) ∈ {(-150, 0), (150, 0), (0, 150)}; each point in the plot represents a subtomogram and close points represent similar conformations. Note that the dashed curves enclose the data points where p-value > 0.01 in Table 3. See the text for more details on this experiment.

Table 3 Mean absolute error and standard deviation between the estimated and ground-truth normal-mode amplitudes along with the angular and shift distances obtained with HEMNMA-3D and "Discrete" synthetic dataset, using an atomic structure (Atomic) and simulated EM map (Volume) as input references.

Experiment		Mode 7		Mode 8		Mode 9		p-value	Samples
Ref	Dataset	mean	Std	mean	std	mean	std		
Atomic	"Discrete"	16.51	11.87	10.91	7.64	10.7	6.66	P > 0.01	871/900
Volume	"Discrete"	17.7	13.26	11.9	10.13	12.29	8.03	P > 0.01	870/900

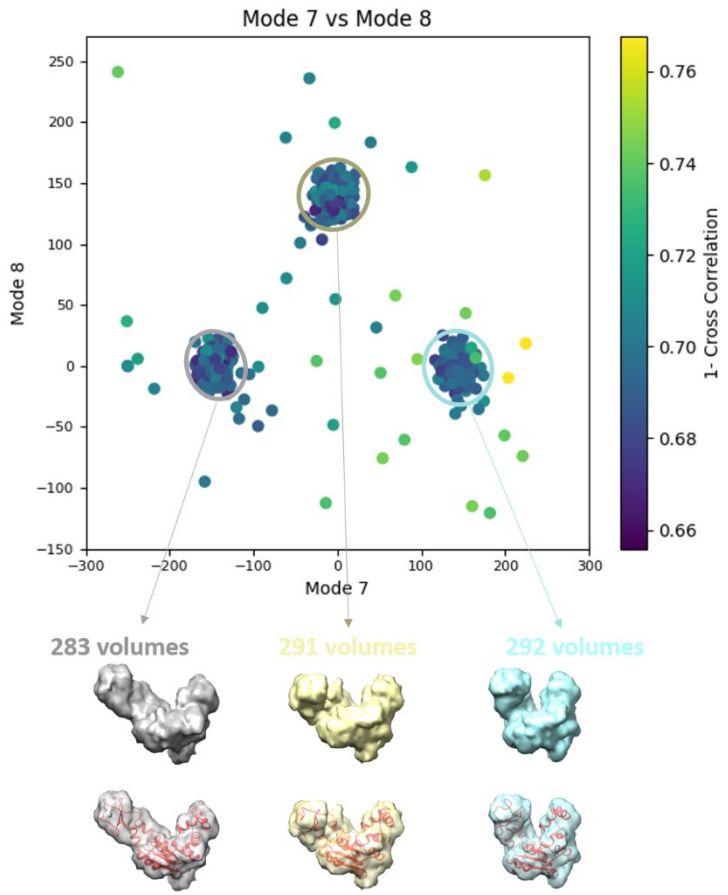


Figure 38 Averages of the three groups (enclosed by ellipses) of subtomograms identified from the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with “Discrete” dataset, using the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. Subtomograms are represented by points and close points represent similar conformations. The numbers of volumes written above the shown subtomogram averages are the numbers of synthetic subtomograms used for computing these subtomogram averages (the numbers of points enclosed by the corresponding ellipses). On the bottom, the subtomogram averages are shown at 50% transparency along with the corresponding ground-truth deformed atomic structure (in red).

Synthetic continuous-type conformational variability

Similarly to the previous experiment, our goal in this experiment is to find a solution for the inverse problem of finding the conformation of the structure in each subtomogram using the combined elastic and rigid-body alignment of a reference model with the subtomograms in the "Continuous" dataset.

We used the same two reference models as in the previous experiment to estimate the normal-mode amplitudes: an atomic structure (chain A of PDB:4AKE) and a density map from this atomic structure.

Also, as in the previous experiment, we used three modes for both tests (atomic or pseudoatomic modes 7, 8 and 9).

Figure 39 presents the estimated amplitudes of modes 7 and 8 (the estimated amplitude of mode 9 is close to 0 and is not shown in the plots). Table 4 shows the mean absolute error and the standard deviation between the estimated and ground-truth normal-mode amplitudes.

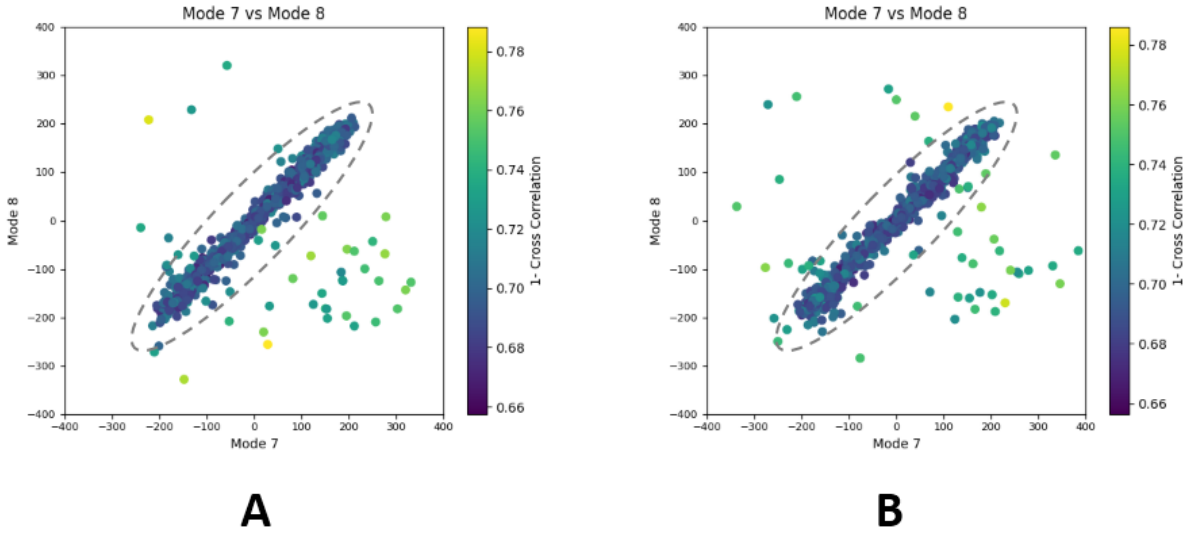


Figure 39 Plots showing the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with "Continuous" dataset (synthetic subtomograms are simulating continuous conformational heterogeneity). (A) Use of the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. (B) Use of a pseudoatomic structure (from a simulated density map) and its normal modes to estimate the conformational and rigid-body parameters of the molecules in the input synthetic subtomograms. The goal was the retrieval of the ground-truth relationship between the amplitudes along normal modes 7 and 8 (ideally linear relationship, with equal amplitudes of normal modes 7 and 8); each point in the plot represents a subtomogram and close points represent similar conformations. Note that the dashed ellipses enclose the data points where p-value > 0.001 in Table 4. See the text for more details on this experiment.

Table 4 Mean absolute error and standard deviation between the estimated and ground-truth normal-mode amplitudes along with the angular and shift distances obtained with HEMNMA-3D and “Continuous” synthetic dataset, using an atomic structure (Atomic) and simulated EM map (Volume) as input references.

Experiment		Mode 7		Mode 8		Mode 9		p-value	Samples
Ref	Dataset	mean	std	mean	std	mean	std		
Atomic	"Continuous"	20.12	11.3	12.78	11.24	12.74	7.71	P > 0.001	960/1000
Volume	"Continuous"	21.94	12.59	14.03	9.92	15.68	10.04	P > 0.001	957/1000

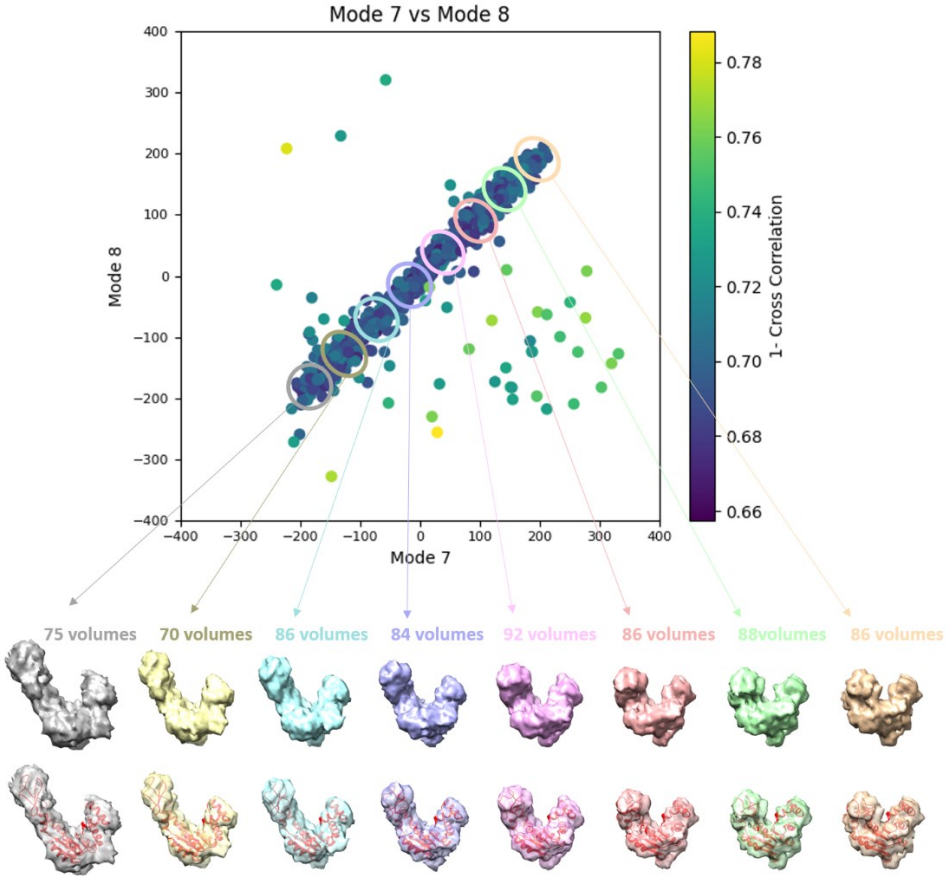


Figure 40 Averages of eight groups (enclosed by ellipses) of subtomograms identified from the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with “Continuous” dataset, using the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. Subtomograms are represented by points, and close points represent similar conformations. The numbers of volumes written above the shown subtomogram averages are the numbers of synthetic subtomograms used for computing these subtomogram averages (the numbers of points enclosed by the corresponding ellipses). On the bottom, the subtomogram averages are shown at 50% transparency along with the corresponding theoretical centroid deformed atomic structure (in red).

In both test cases, a linear relationship between the estimated amplitudes of normal modes 7 and 8 is clearly distinguishable, which is close to the identity relationship between the ground-truth amplitudes considering strong noise present in the data. As in the previous experiment, the results show a slightly less accurate alignment in the second test type (pseudoatomic reference) for the same aforementioned reason. Figure 40 shows the grouping and averaging of subtomograms in this experiment, with 8 subtomogram averages calculated along the distribution of the points for the first test type (atomic reference). The subtomogram averages show different conformations of adenylate kinase chain A. Note that the noise contained in the individual subtomograms (SNR = 0.01, Figure 36B) was reduced through subtomogram averaging (Figure 40). Additional experiments for other noise levels in input subtomograms can be found in the next section.

Additional synthetic data tests with different noise levels

Additional tests were performed on HEMNMA-3D using synthesized datasets at different noise levels of conformationally heterogeneous subtomograms that mimic continuous conformational variability. The noise levels were chosen as A) without noise, B) SNR = 0.4, C) SNR = 0.1, D) SNR = 0.04, E) SNR = 0.01 and F) SNR = 0.005.

Each dataset comprises 200 synthetic subtomograms representing a continuum of conformations of the same PDB:4AKE structure. Here, the amplitude value of normal mode 7 is chosen randomly in the range {-300, 300} and the amplitude value mode 8 is half of the value of mode 7. The goal in this experiment is to find a solution for the inverse problem of finding the conformation of the structure in each subtomogram using the combined elastic and rigid-body alignment of a reference model with the subtomograms in the different-SNR datasets.

Figure 41 presents the estimated amplitudes of modes 7 and 8 (the estimated amplitude of mode 9 is close to 0 and is not shown in the plots). Table 5 shows the mean absolute error and the standard deviation between the estimated and ground-truth normal-mode amplitudes.

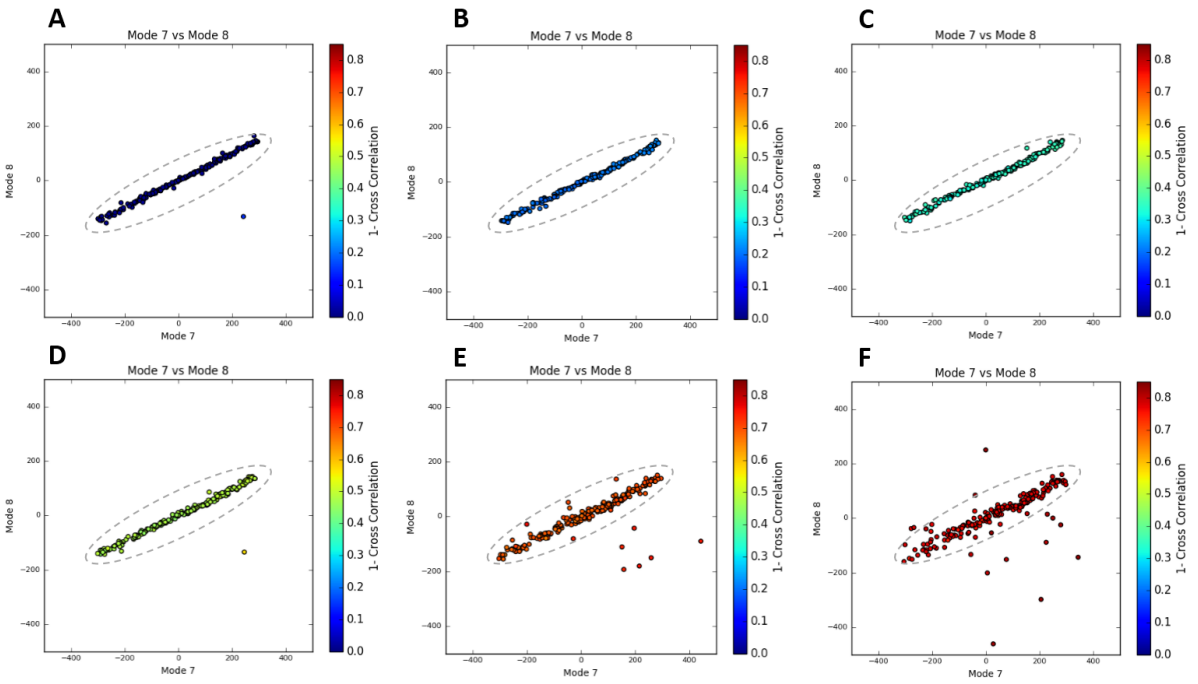


Figure 41 Plots showing the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with synthetic datasets at different noise levels (synthetic subtomograms are simulating continuous conformational heterogeneity), using the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. (A) Without noise, (B) SNR = 0.4, (C) SNR = 0.1, (D) SNR = 0.04, (E) SNR = 0.01, (F) SNR = 0.005. The goal was to retrieve the ground-truth relationship between the amplitudes along normal modes 7 and 8 (ideally a linear relationship, with the amplitude of normal modes 8 equals to half the amplitude of mode 7); each point in the plot represents a subtomogram, and close points represent similar conformations. Note that the dashed ellipses contain the data points where the p-value is specified in Table 5.

Table 5 Mean absolute error and standard deviation between the estimated and ground-truth normal mode amplitudes obtained with HEMNMA-3D synthetic datasets for different noise levels, using an atomic structure as an input reference. The corresponding region for the p-value is shown in Figure 41.

Noise and CTF		Mode 7		Mode 8		Mode 9		p-value	Samples
Defocus [µm]	SNR	mean	std	mean	std	mean	std		
No Noise		6.38	4.47	3.37	3.51	3.7	3.85	P > 10^-9	199/200
-1	0.4	8.23	6.35	7.49	5.29	5.64	4.02	P > 0	200/200
-1	0.1	8.26	6.34	8.1	5.83	5.85	4.26	P > 0	200/200
-1	0.04	11.14	7.2	8.11	5.89	6.98	4.83	P > 10^-9	199/200
-1	0.01	26.68	9	13.11	8.75	16.59	9.26	P > 0.01	190/200
-1	0.005	35.86	14.2	20.87	11.02	19.4	9.53	P > 0.01	187/200

Experimental cryo-ET data: nucleosomes *in situ*

We applied HEMNMA-3D on a dataset comprising ~650 in situ subtomograms of nucleosomes collected from a cell of a Drosophila embryonic brain (dataset presented in the previous chapter), whose conformational variability was detected but not fully explored in previous work [28]. The subtomograms had the size of 64^3 voxels and a voxel size of 4.4 $Å^3$. A density map obtained with classical subtomogram averaging (without considering conformational heterogeneity) was used as the reference density map for HEMNMA-3D (Figure 42C). The resolution of this reference density map is around 2 nm (as determined by Fourier Shell Correlation between the reference density map and the density map from the atomic nucleosome structure PDB:3w98 [137] shown in Figure 42B). For more information on how this reference density map (global initial subtomogram average) was obtained, please see Chapter 4. This reference density map was converted into pseudoatoms (1368 pseudoatoms for the pseudoatom radius of 0.5 voxels and the target approximation error of 5%) and normal mode analysis of the obtained reference pseudoatomic structure was performed. The combined elastic and rigid-body alignment was performed using the pseudoatomic structure and a set of its six low-frequency high-collectivity normal modes (selected as described above and in HEMNMA-related works [43], [124] and [123]). The normal-mode amplitudes estimated through the alignment (six normal mode amplitudes per subtomogram) were then projected

onto a 2D space of conformations using PCA. The space of conformations is presented in Figure 42. Recall that each of the points represents a subtomogram, and close points represent similar conformations. By inspecting this conformational space, we identified four densest regions with 70, 183, 74, and 64 points from left to right in Figure 42D. Following this analysis, we grouped the subtomograms in each of these four regions and averaged them. Before averaging, we filled in the missing-wedge Fourier space region of the individual subtomograms with the corresponding region of the global average computed from all subtomograms (please note that this global average was computed after aligning subtomograms using the rigid-body alignment parameters found along with the 3D-to-3D elastic alignment by HEMNMA-3D, which is a similar density map to the initial global average map shown in Figure 42C as both density maps result from averaging conformational heterogeneous subtomograms). The displacement of the reference pseudoatomic structure (converted into a density map) along two directions D1 and D2 in the space of conformations is shown in Figure 43 and in the supplementary material of the published article [120] Movie S1 and Movie S2. The significant difference between the four group averages (Figure 42D) and the reference density map (Figure 42C) as well as the motion observed along the two directions D1 and D2 (Figure 43, Movie S1 and Movie S2) can be described, mainly in terms of opening the nucleosome by increasing the distance between the two gyres of the DNA superhelix. This result consents the previous findings, observed but not fully explored in a previous study (manual analysis) of the nucleosome conformational variability [28]. The group averages are also compared with the atomic nucleosome structure PDB:3w98 in Figure 44.

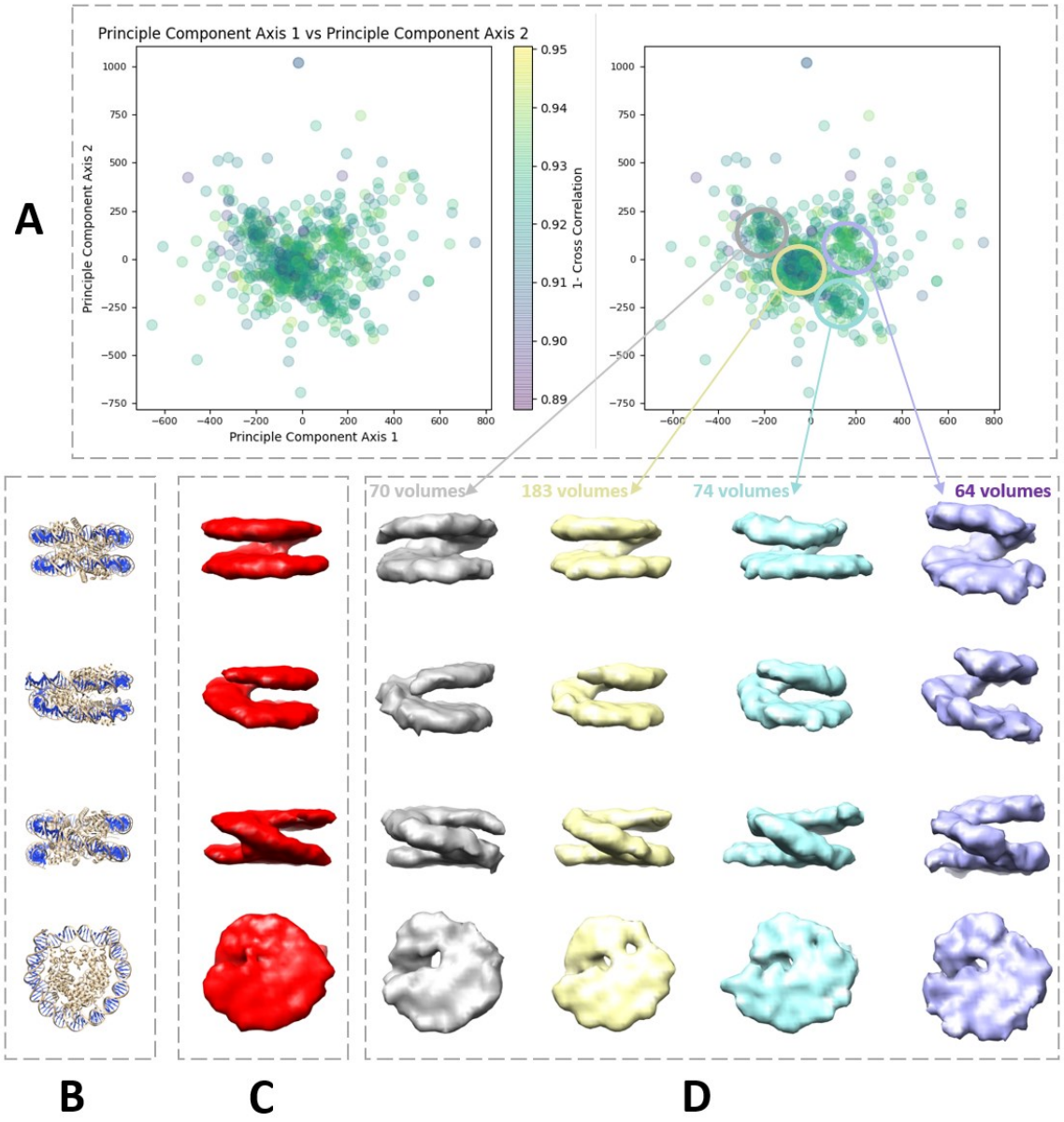


Figure 42 Illustration of HEMNMA-3D use with in situ cryo-ET nucleosome dataset. (A) Space of conformations resulting from projecting the estimated amplitudes of six normal modes onto a two-dimensional space using PCA. (B) Nucleosome atomic structure PDB:3w98, for comparison purposes. (C) Nucleosome subtomogram average (around 2 nm resolution) used as the input reference density map for HEMNMA-3D, obtained by classical subtomogram averaging, without considering conformational heterogeneity [for more information on how this global initial subtomogram average was obtained, see Chapter 5 (Nucleosome data preparation and acquisition)]. (D) Four subtomogram averages from four densest regions in the space of conformations (regions encircled with ellipses) showing different nucleosome conformations, mainly different gap distances between the nucleosome gyres. The numbers of volumes written above the subtomogram averages shown in (D) are the numbers of in situ cryo-ET subtomograms used for computing these subtomogram averages (the numbers of points enclosed by the corresponding ellipses).

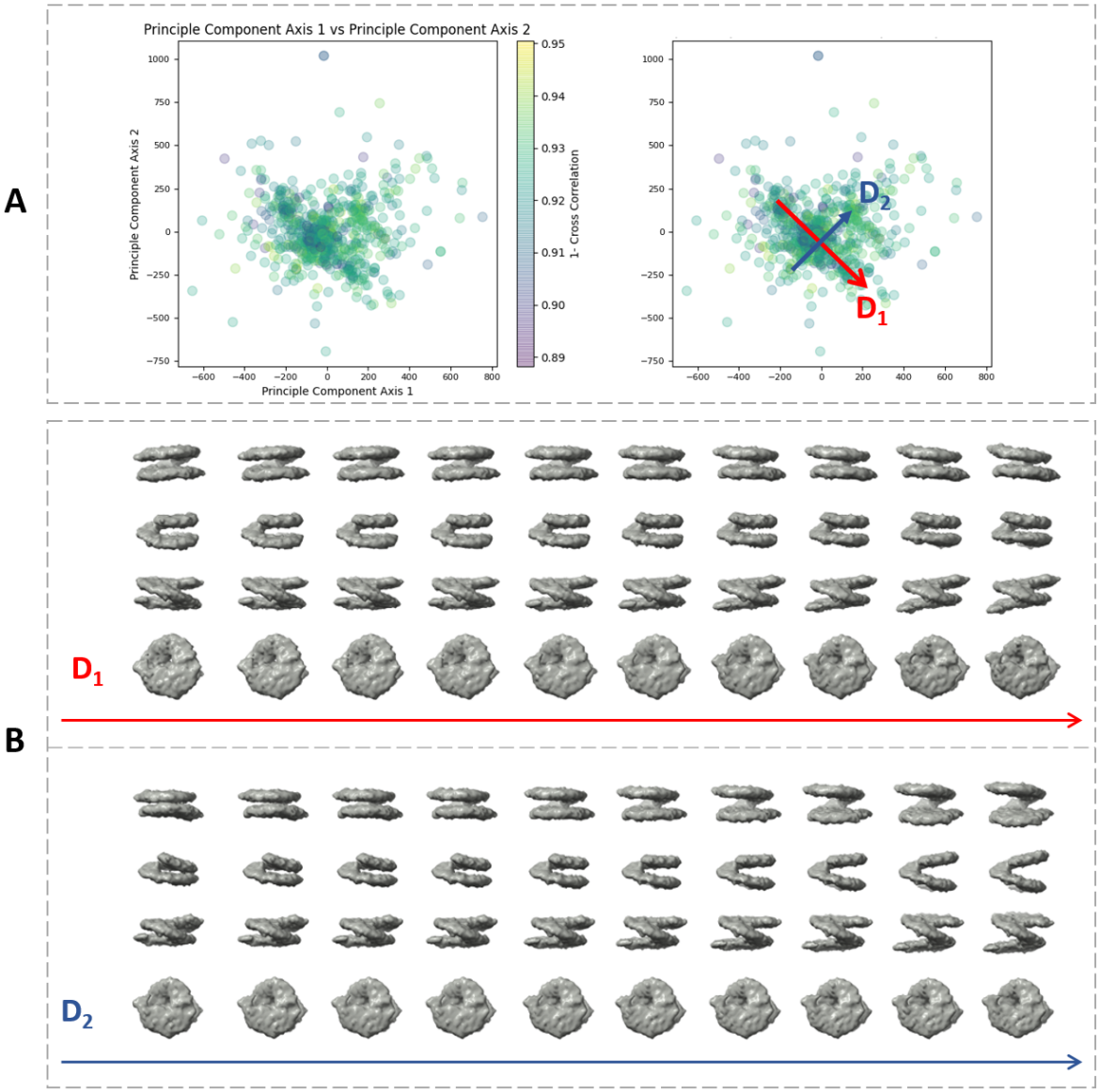


Figure 43 Displacement of the reference density map along two directions D1 and D2 in the space of conformations obtained (Figure 42) with HEMNMA-3D with in situ cryo-ET nucleosome dataset. (A) Space of conformations (left) as shown in Figure 42 and two directions D1 and D2 used to displace the reference density map (Figure 42C) in this space (right). (B) Displacement of the reference density map along the D1 and D2 directions (10 frames of the corresponding trajectory are shown row-wise).

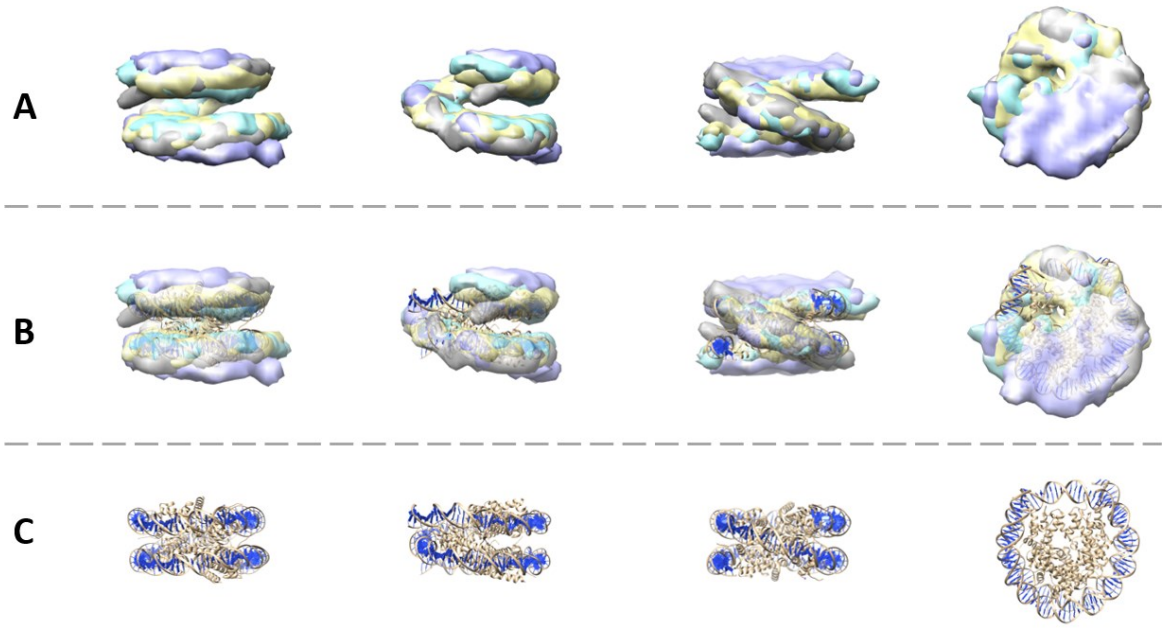


Figure 44 Comparison of the atomic nucleosome structure PDB:3w98 with the four in situ nucleosome subtomogram averages obtained with HEMNMA-3D (the experiment shown in Figure 42, which used a preliminary nucleosome subtomogram average as input reference density map for HEMNMA-3D). (A) Four views of the four subtomogram averages overlapped. (B) Four views of the four averages overlapped at 50% transparency with PDB:3w98. (C) Four views of PDB:3w98.

Comparing HEMNMA-3D to traditional STA and classification

This section compares HEMNMA-3D performance with existing literature on a synthesized dataset of nucleosomes with a synthetic continuous shape variability. First, it presents how a dataset was synthesized; then, it presents the methods applied and results obtained on this dataset using i) the traditional methods of subtomogram averaging and classification, and ii) HEMNMA-3D.

Simulating a dataset of nucleosome conformational variability

We synthesized a dataset comprising 1000 subtomograms with a continuous shape variability of the nucleosome by generating a linear combination of two reported motions for the nucleosome, breathing and gaping [90], with a linear dependence between the amplitudes of normal modes corresponding to the two motions.

First, we performed NMA of the nucleosome atomic structure available in the PDB database under the code 3w98 and we visualized the motions carried by the different computed

normal modes. Among these modes, we identified the modes describing breathing and gaping motions as normal modes 9 and 13, respectively. We generated a dataset using a linear relationship between the amplitudes of normal modes 9 and 13 so that the nucleosome is simultaneously breathing and gaping. Precisely, at one end of the generated ground-truth conformational distribution, the nucleosome's two DNA ends (arms) are moving away from each other, and at the same time, the gap between the two DNA gyres increases. At the other end of the generated conformational distribution, the DNA arms approach each other and the gap between the two DNA gyres decreases. We simulated a gradual transition between the two ends, representing a continuum of nucleosome shapes, combining breathing and gaping. Equal random amplitudes uniformly distributed in the range [-150, 150] were used for the two normal modes 9 and 13. An illustration of the simulated movements is provided in Figure 45.

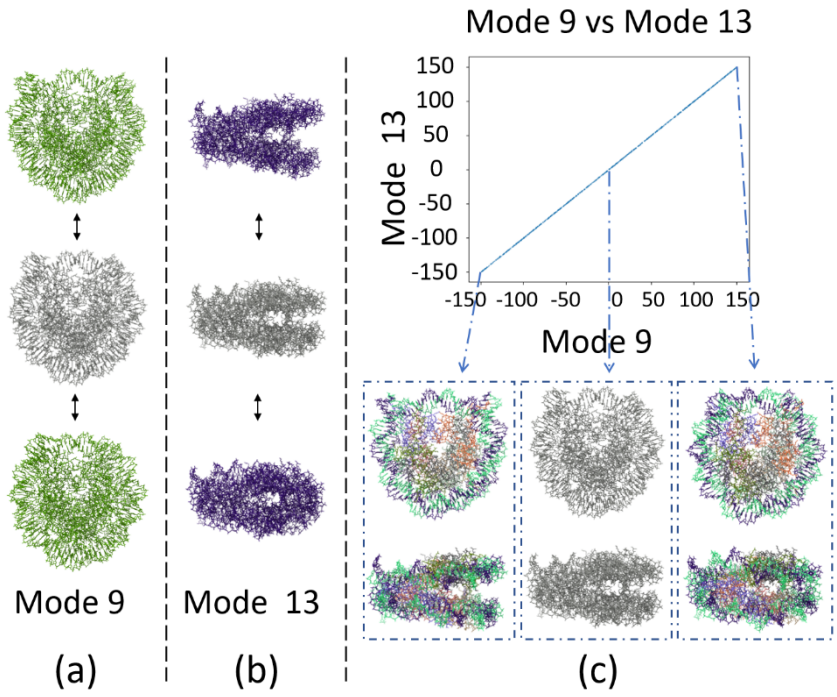


Figure 45 Synthesized combined breathing and gaping motions of the nucleosome (PDB 3w98 structure): (a) nucleosome breathing motion, (b) nucleosome gaping motion, (c) generated ground-truth conformational distribution (top) comprising 1000 synthetic nucleosome shape variants obtained by a linear combination of modes 9 and 13, with a linear dependence between the normal-mode amplitudes (blue points in the plot), and 3 representative shapes (bottom) corresponding to the two ends and the middle of the conformational distribution.

To generate this dataset, for each subtomogram, we performed the following steps:

- Elastically deform the atomic structure (PDB:3w98) using equal random amplitudes for the two normal modes 9 and 13 in the range [-150, 150].
- Convert the elastically deformed structure to a density map of size 64 x 64 x 64 voxels (voxel size: 3.45 Å x 3.45 Å x 3.45 Å), using [136].
- Rotate and shift the volume in 3D space using random Euler angles and random x, y, z shifts (the random shift range is ±5 pixels from the center).
- Tilt and project the randomly rotated and shifted volume, using the tilt angle from -60˚ to +60˚ with 1˚ step, to obtain a collection of 2D projection images (i.e. tilt series).
- Simulate microscope conditions by adding noise, modulating the images with the contrast transfer function (CTF) of the microscope (using the defocus of -0.5 μm), then adding noise again in such a way that a part of the noise will be modulated by the CTF, and the other part will not, with a total SNR = 0.01.
- Invert the CTF phase (a common CTF correction).
- Reconstruct a volume (our synthetic subtomogram) from the obtained tilt series using a Fourier reconstruction method [20]

Figure 46 shows an example subtomogram from the synthesized dataset and the corresponding ideal volume, for comparison in real space and in Fourier space.

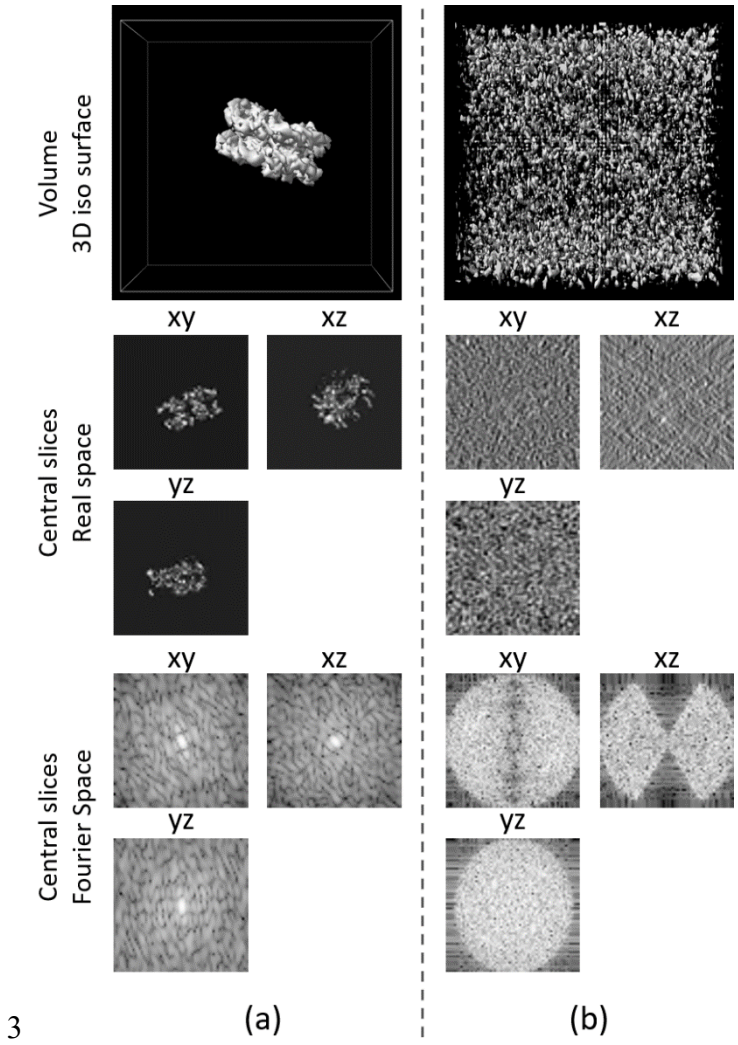


Figure 46 Example of a noisy and missing-wedge affected synthetic subtomogram compared with the corresponding ideal volume of the nucleosome: (a) ideal volume (without noise and without missing wedge artifacts), (b) noisy and missing-wedge affected synthetic subtomogram.

Traditional subtomogram averaging and post alignment classification

StA provides a global average without considering the shape variability, and it provides a basis for performing classification of subtomograms (classification of the subtomograms aligned through StA).

We applied StA on the synthetic nucleosome dataset, using the protocol based on the rigid-body alignment approach of [48] (recall that this rigid-body alignment approach is also used in the elastic and rigid-body alignment of HEMNMA-3D). This StA protocol uses an exhaustive angular search (with FRM method) and a shifts search within a region of interest, and compensates for the missing wedge by using the CCC (evaluation of the correlation

between the subtomogram average of each iteration and the given subtomogram density maps, but excluding the evaluation in the missing-wedge region of the given subtomogram).

We followed the procedure in [48] and set the shifts search region to 10 voxels from the image center. We started iterations using an average of the unaligned subtomograms (this StA procedure is referred to as reference-free alignment). After six iterations, StA converged (further iterations gave the same results). The StA averages are shown in Figure 47.

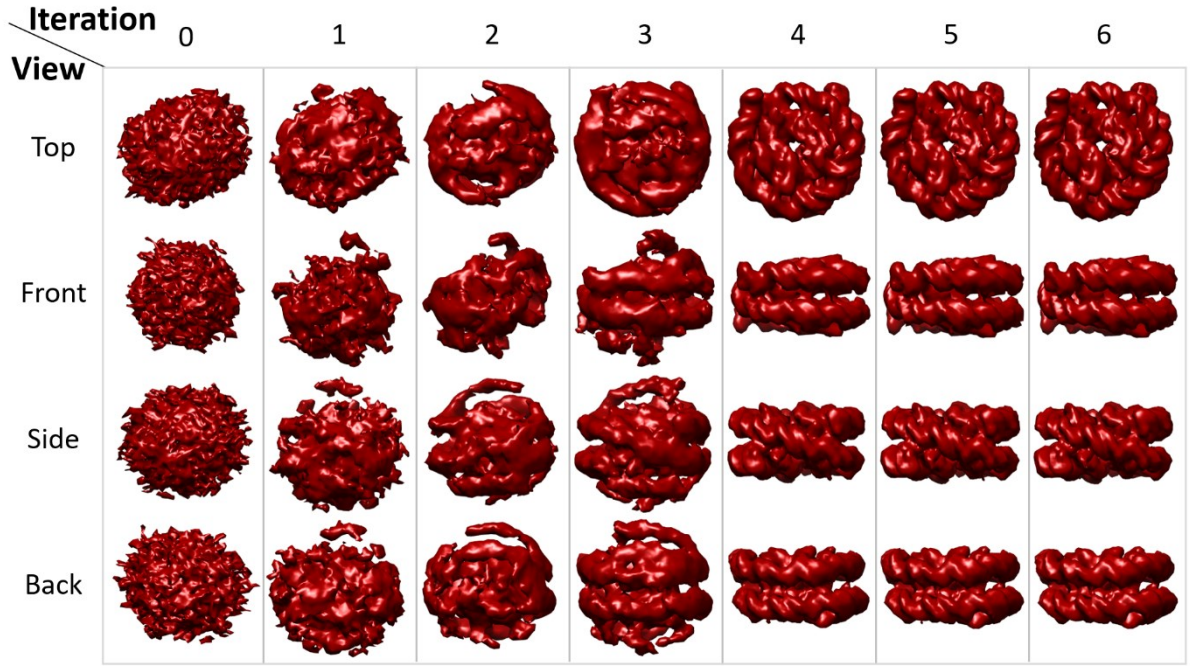


Figure 47 Subtomogram averaging applied to the synthetic nucleosome subtomograms. A reference-free alignment was performed using Fast Rotational Matching

After StA, we applied the obtained rigid-body alignment parameters (found through StA) on the subtomograms, and we evaluated the covariance matrix CCC_{ij} of pairwise constrained cross-correlation (see Chapter 2 for more details). We performed the two most common post-alignment classification techniques on the CCC_{ij} matrix, namely hierarchical clustering and PCA followed by k-means [47, 51].

The hierarchical clustering on 1-CCC_{ij} matrix was performed to 10 classes using the Agglomerative Clustering module of Python Scikit-Learn package (version 0.22.1 and default parameters were used) [138]. We note that applying the clustering algorithm directly on the CCC_{ij} matrix gives identical results, and we used the convention proposed in the literature [47, 51]. The clustering tree (dendrogram) and class averages are shown in Figure 48.

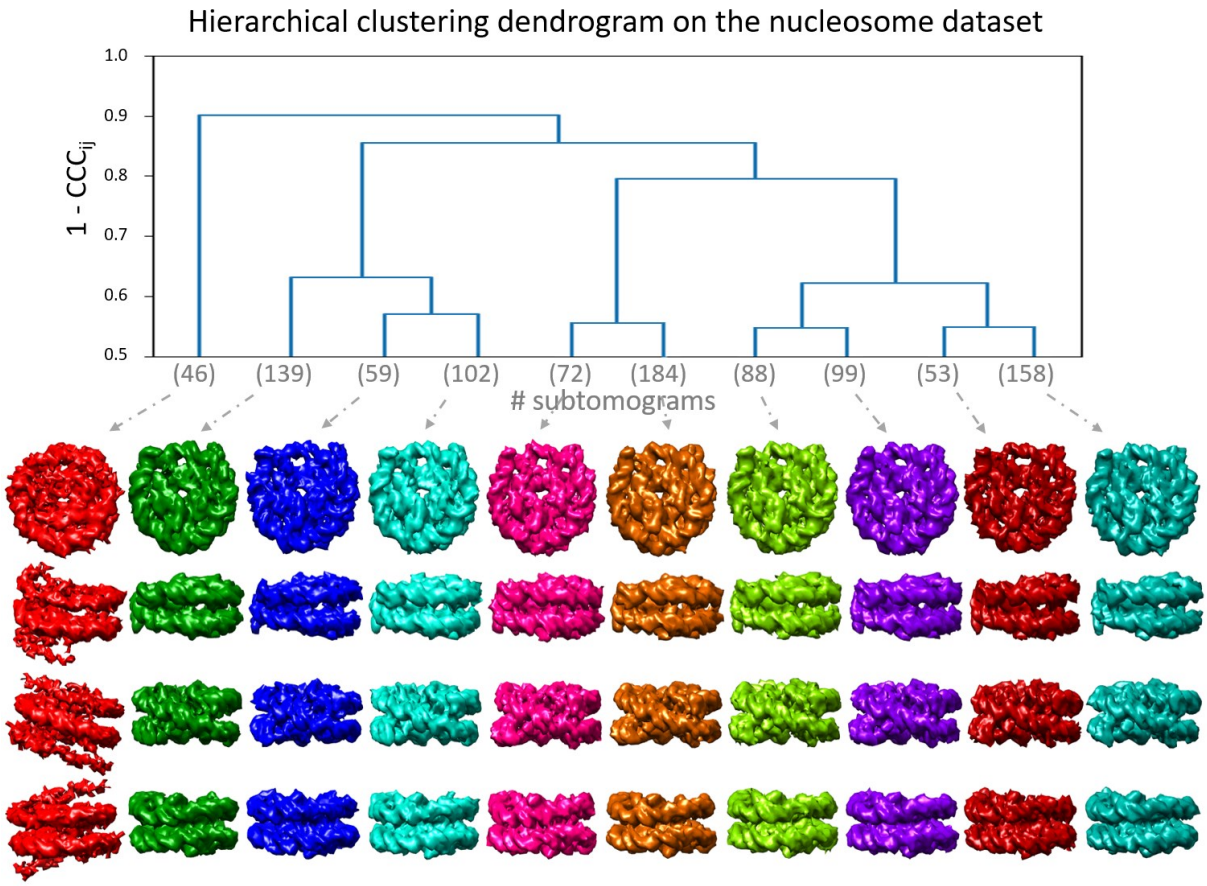


Figure 48 Hierarchical clustering applied to the synthesized nucleosome subtomograms. Top: hierarchical tree for 1-CCC_{ij} matrix. Bottom: views (vertically in the same color) of different subtomogram class averages (horizontally in different colors).

The k-means clustering was performed following PCA on the CCC_{ij} matrix. The clustering was done into 10 classes (k=10) based on the first two principal axes, using the k-means module of Scikit-Learn. In general, the choice of the number of principal axes to perform classification is arbitrary, as explained in [51]. Since the dataset was synthesized with two degrees of freedom (nucleosome breathing and gaping), we set the number of principle axes to 2, to obtain the best results. Figure 49 shows the classification of the PCA space and the resultant class averages.

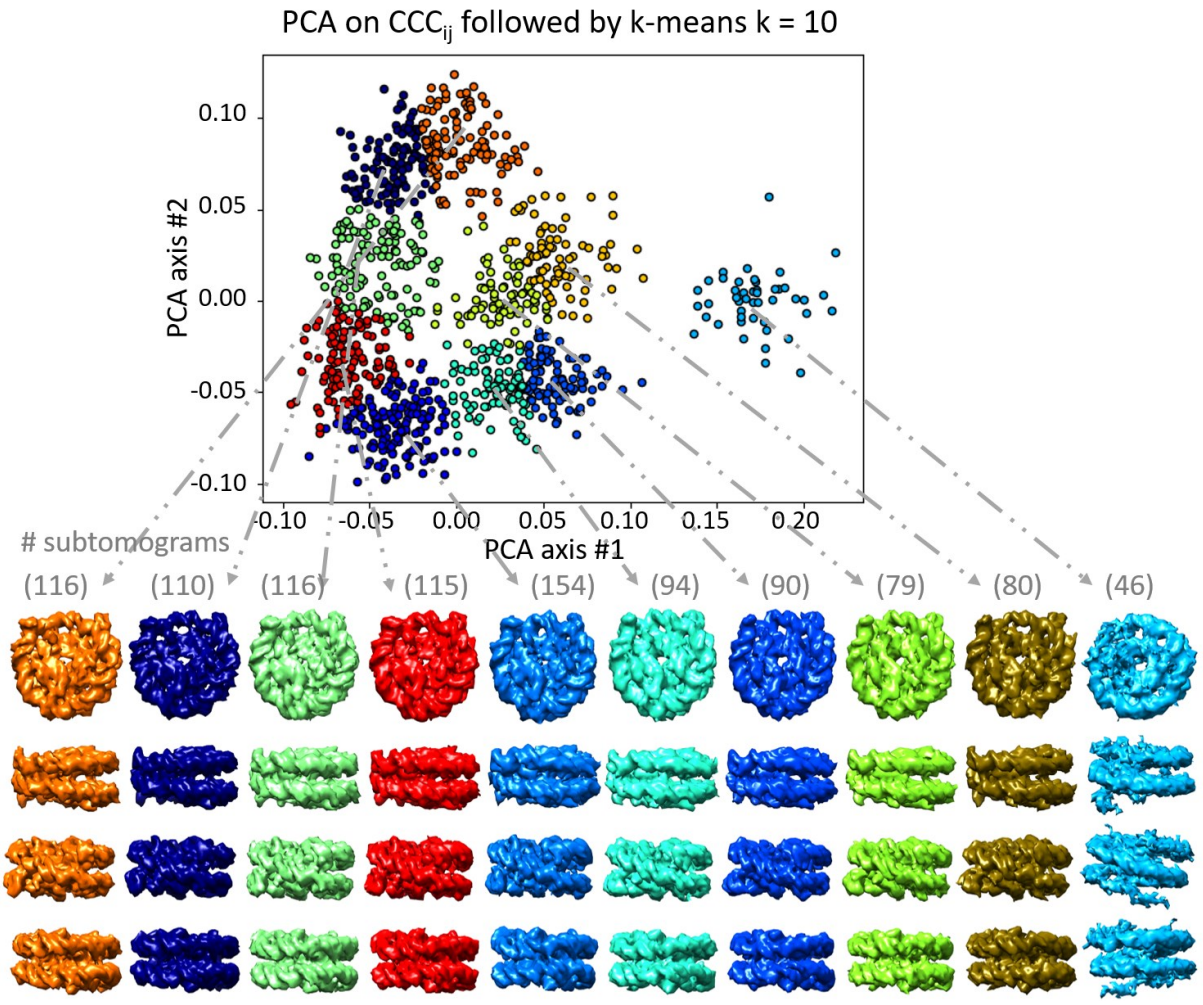


Figure 49 K-means clustering applied to the synthesized nucleosome subtomograms. Top: k-means clustering in the space of the first two PCA axes of CCC_{ij} matrix. Bottom: views (vertically in the same color corresponding to the color in the PCA space) of different subtomogram averages (horizontally in different colors).

We note that the two tested classification techniques give similar outputs, showing different discrete class averages of the nucleosome, at different breathing and gaping magnitudes. However, these outputs do not allow an unambiguous interpretation of the results in terms of the synthesized ground-truth conformational transitions of the nucleosome (from the smallest magnitudes to the largest magnitudes of breathing and gaping and vice versa).

HEMNMA-3D

Applying HEMNMA-3D to the synthesized nucleosome dataset aims at solving the inverse problem of finding the nucleosome shape variant in each subtomogram, i.e. estimating the amplitudes of normal modes 9 and 13 of the PDB structure 3w98 as close as possible to the generated ground-truth amplitudes.

We set the method parameters as follows:

- NMA settings: To make the elastic and rigid-body 3D registration task more realistic and challenging, we used three normal modes (modes 9, 10 and 13) instead of only two modes (modes 9 and 13 used to generate the dataset).
- FRM settings: The shift range for the rigid-body registration (FRM method) is set to 10 pixels.

The amplitudes estimated for modes 9 and 13 using HEMNMA-3D are shown in Figure 50a. It is graphically intelligible that the linear relationship is retrieved between the estimated amplitudes of the two modes. Figure 50b shows the histogram of the amplitudes estimated for mode 10 and confirms that they are globally near zero.

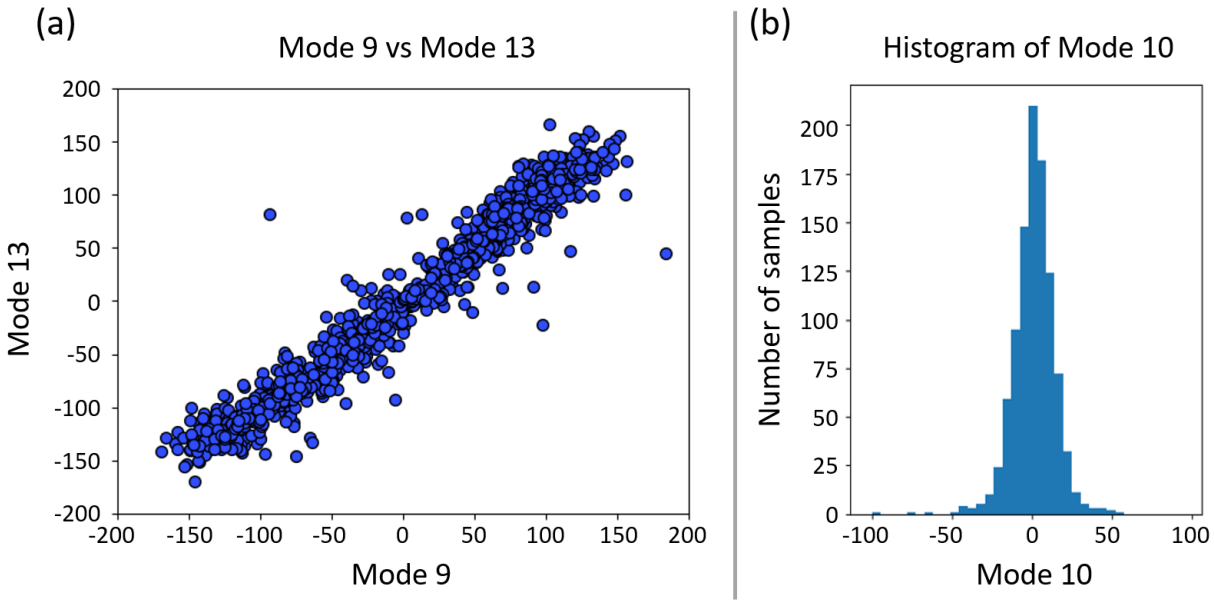


Figure 50 Output of the elastic and rigid-body 3D registration module of HEMNMA-3D using synthesized nucleosome subtomograms. The goal was the retrieval of the ground-truth amplitudes of normal modes 9, 10 and 13. Ideally, the amplitudes of mode 10 are equal to zero and there is a linear relationship between the amplitudes of modes 9 and 13 in the range [-150, 150]: (a) amplitudes of mode 9 vs amplitudes of mode 13, (b) histogram of amplitudes of mode 10.

Table 6 presents the mean absolute error between the estimated and ground-truth normal-mode amplitudes and the standard deviation of the error. It should be noted that 14/1000 points were excluded from the statistics as found to differ significantly (outlier points) from the remaining observations. These points were excluded for having a p-value below 10^{-4} based on the Mahalanobis distance [133].

Table 6 Mean absolute error between the estimated and ground- truth normal-mode amplitudes and the standard deviation of the error, for HEMNMA-3D using synthesized nucleosome subtomograms. Points below the p-value of 10^{-4} were excluded (14/1000 points) from the error evaluation based on the Mahalanobis distance.

Normal mode amplitude	Mode 9		Mode 10		Mode 13	
Actual range	[-150, 150]		0		[-150, 150]	
Measure	Mean	Std	Mean	Std	Mean	Std
Absolute error	10.86	8.66	9.98	7.82	10.81	8.83

Normal-mode amplitudes do not have a physical unit. Nonetheless, the Root Mean Square Deviation (RMSD) [139] between the reference atomic coordinates and these coordinates displaced using the calculated errors as the normal-mode amplitudes can transform these errors in physical units. The nucleosome core complex comprises eight histone proteins surrounded by 146 DNA base pairs. The synthesized movements (breathing and gaping) mainly impacted the DNA loops. Evaluating the RMSD without excluding the core histones can give a false sense of achieving higher accuracy by pulling the RMSD value towards zero. Therefore, the reported RMSD hereafter is based on the nucleosome's DNA loops only (chain I and J of the PDB structure 3w98).

We found an RMSD of 0.44 Å corresponding to the mean absolute errors in Table 6 (for a combined displacement along modes 9, 10 and 13). Also, we found an RMSD of 0.79 Å corresponding to the sum of the mean and standard deviation of the errors in Table 6. Hence, the error range is significantly inferior to the pixel size used to create the data (3.45 Å).

Figure 51a shows grouping and averaging of subtomograms through the point distribution in the conformational space (ten equally distanced groups). The corresponding subtomogram averages show the expected combination of continuous motions of breathing and gaping, which can be compared with the ground-truth motion in Figure 45.

Figure 51b shows the displacement of the reference structure along 10 points in the direction of the point distribution in the conformational space.

The obtained subtomogram averages and animation show that the ground-truth nucleosome motion (a combination of breathing and gaping) was retrieved.

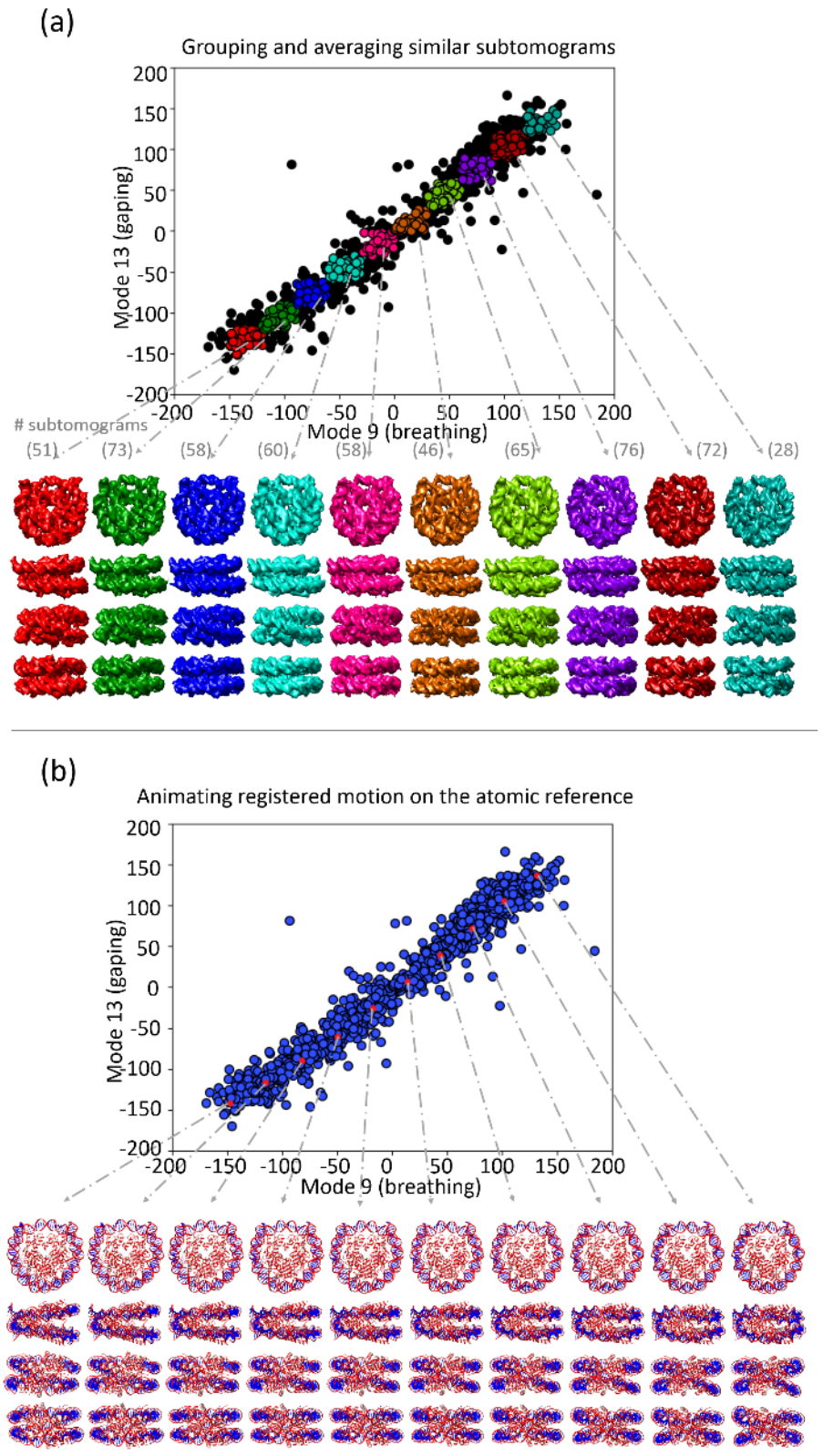


Figure 51 HEMNMA-3D applied to the synthesized nucleosome subtomograms. (a) group averages for ten equally distanced groups along the subtomogram (point) distribution in the conformational space, (b) displacement of the reference PDB structure 3w98 along the direction of the data distribution in the conformational space, 10 frames represented by red dots. Note: each column represents four different views of the same structure.

Discussion

This chapter presented HEMNMA-3D, the first cryo-ET subtomogram data analysis approach to study continuous conformational variability of biomolecular complexes, which maps a set of subtomograms into a space of conformations using a reference model and its normal modes. The conformational space permits i) grouping (and averaging) subtomograms with similar conformations and revealing hidden conformations and ii) recording animated displacements of the reference model along the densest regions of the space, along trajectories identified by curve fitting of the data in these regions. These HEMNMA-3D outputs could be valuable to cryo-ET studies of molecular mechanisms involved in conformational changes of complexes *in vitro* and *in situ*. HEMNMA-3D is thoroughly tested using synthetic subtomograms and applied to a cryo-ET experimental dataset (nucleosome subtomograms recorded *in situ* in Drosophila interphase nucleus). It provides promising results coherent with previous findings.

Additionally, we compared HEMNMA-3D to two state-of-the-art methods for cryo-ET classification following STA, on a synthetic dataset of nucleosome shape variability. Both methods gave similar outputs, showing different discrete class averages of the nucleosome at different breathing and gaping magnitudes. However, the choice of the number of classes is arbitrary in these methods and the shape transitions between the obtained class averages are ambiguous, probably because of the continuous nature of the shape variability.

However, unlike the classification methods, HEMNMA-3D is limited to macromolecular elastic shape variability that can be explained with NMA. It is not suitable for analyzing other structural variabilities such as macromolecular disassembly or binding and unbinding of ligands. Future work can involve combining this method with classification to first disentangle such discrete structural variabilities and then analyze continuous intraclass variability.

An open-source software with a graphical user interface is provided for this method with a C++ backend, and a Message Passing Interface parallelization scheme, explained in Chapter 7.

Chapter 6. TomoFlow: Cryo-ET data processing method based on optical flow to analyze continuous conformational variability of biomolecular complexes

This chapter presents TomoFlow, a method for analyzing macromolecular continuous conformational variability in cryo-ET subtomograms based on a three-dimensional dense optical flow (OF) approach. The resultant lower-dimensional conformational space allows generating movies of macromolecular motion and obtaining subtomogram averages by grouping conformationally similar subtomograms. The animations and the subtomogram group averages reveal accurate trajectories of macromolecular motion based on a novel mathematical model that makes use of OF properties.

TomoFlow was published in early 2022 [140]. This chapter describes the method, results, and conclusions of TomoFlow as presented in the published manuscript. It mainly shows the tests on simulated datasets generated using different techniques, namely Normal Mode Analysis and Molecular Dynamics Simulation, and an application of TomoFlow on a dataset of nucleosomes *in situ* (the dataset presented in Chapter 4 and previously analyzed using HEMNMA-3D in Chapter 5), which provided promising results coherent with previous findings using the same dataset but without imposing any prior knowledge on the analysis of the conformational variability.

TomoFlow method

This section first introduces TomoFlow's general scheme and objectives and then walks the reader through its building blocks with the necessary mathematical derivations and theoretical background.

TomoFlow (shown in Figure 52) analyzes the conformational variability in subtomograms after MW-correction and rigid-body alignment. It performs OF-based matching of the subtomograms with an input reference (e.g., global subtomogram average) in the presence of a mask of the region of interest. Then, it collectively analyzes the resultant OFs between the input reference and each of the subtomograms by finding their Gram matrix and mapping it to a lower-dimensional space called the space of conformations (e.g., via PCA). In

the conformational space, each point corresponds to a subtomogram, and close points correspond to subtomograms containing similar conformations. Accordingly, the conformational space is interactively processed by i) grouping close points in dense regions and averaging the corresponding subtomograms to obtain subtomogram averages at different conformations, and ii) generating movie animations on the input reference while it fits curves in the conformational space following data distribution manifolds, i.e., animating the input reference to show the motion following the dense regions in the space.

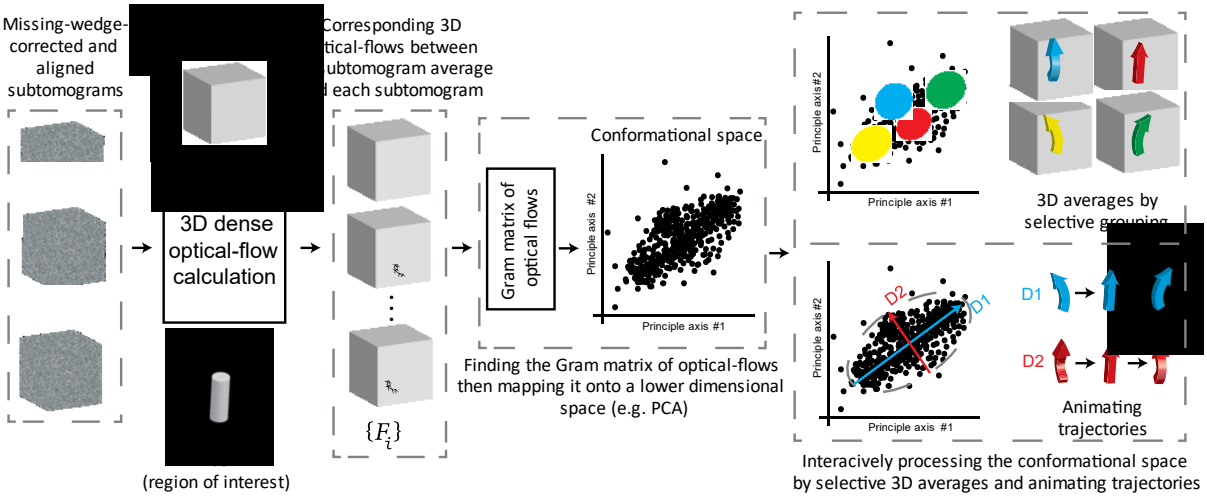


Figure 52 Proposed pipeline for analyzing conformational variability in a set of subtomograms using 3D dense optical flows between a reference (here, subtomogram average) and each of the subtomograms.

Employment of 3D dense OF for elastic and rigid-body matching of subtomograms

Dense optical flow has been explained in Chapter 4 in detail. This subsection reminds the reader of the working principles of 3D dense OF. It then explains how OF can be employed for 3D elastic and rigid-body matching, allowing subtomogram rigid-body alignment refinement and continuous conformational variability analysis.

3D dense OF is an algorithm that aims at finding the voxel-to-voxel correspondence between two volumetric images. OF calculations depend on two principles; the first principle is brightness consistency, which means that the gray-level values (i.e., the brightness) of the corresponding voxels in the two input volumes are similar. To find the relationship between the voxels of two volumes I and H, we assume that for a voxel (x, y, z) in H, a voxel in I with similar brightness can be found at some distance (u, v, w) on x, y, and z axis respectively:

$$H(x, y, z) \approx I(x + u, y + v, z + w) \tag{6.1}$$

The second principle is that the distance (u, v, w) is small and that a limited number of terms (e.g., one term) of a Taylor expansion of the right-hand term of eq (6.1) is enough to describe the motion:

$$I(x + u, y + v, z + w) \approx I(x, y, z) + \frac{dI}{dx}u + \frac{dI}{dy}v + \frac{dI}{dz}w \tag{6.2}$$

Dense OF between I and H can be defined as the set of magnitudes (u, v, w) for all (x, y, z) between I and H to satisfy (6.1 - 6.2).

The two principles above established a practical computational background for OF over the years. Still, OF has suffered limitations in its functionality when the corresponding pixels (or voxels) in the two input images (or volumes) do not have the same brightness or are significantly distanced, which rendered OF sensitive to noise and only accounting for small displacements [115]. The 2D OF method of Farnebäck [119], which is a more recent approach, deals with these issues by combining two features. The first is enforcing local smoothness of the OF (close pixels move in the same direction) by approximating a neighborhood of each pixel in each of two given images with a polynomial (the coefficients of the local polynomial are estimated from a weighted least squares fit to the signal values in the neighborhood) and by integrating information about the displacement field between the two images over a neighborhood of each pixel. The data approximation by local polynomials is similar to local data smoothing and the displacement field integration over a pixel neighborhood is similar to local OF smoothing. It should also be noted that this method is not based on calculating image gradients (in eq (6.2)), but it finds a solution of a set of linear algebraic equations (the displacement of a pixel is calculated by directly evaluating matrices expressed in terms of the polynomial coefficients over a neighborhood of the pixel) and this solution is generally unique except in the case when the neighborhood is exposed to the aperture problem [119]). The aperture problem refers to the fact that when a moving object is viewed through a limited-size aperture, the direction of motion of a local feature or a region of the object may be ambiguous. In general, this problem is relevant to rigid objects with straight-line edges or flat regions (e.g., for a moving rectangle, motion of an edge in the direction perpendicular to that edge can be determined unambiguously, but motion of the edge along itself and motion of the inner flat region of the rectangle cannot be determined unambiguously). This problem is less

relevant to cryo-EM and cryo-ET of biological macromolecules, which are generally flexible with curved edges and without flat regions. The second important feature of the method of Farnebäck [119] is calculating OF iteratively and over multiple scales of the input images (image pyramids) [141, 142], which involves refining an OF estimation from a previous iteration or from a coarser image scale, i.e., propagating a refined OF from a coarser to a finer image scale and iterating on each image scale. The two features mentioned above increase robustness to noise, brightness differences, and larger displacements, leading to improved accuracy of the OF calculation.

An extension of the 2D OF method of Farnebäck [119] to deal with volumetric data (Farneback-3D) has been recently implemented (https://pypi.org/project/farneback3d) and this 3D OF calculation method was used in TomoFlow that is presented here. For more information on the iterative multiscale (pyramidal) approach for 3D OF calculation used in TomoFlow, the reader is referred to Chapter 4.

The concept of OF can be employed for 3D elastic and rigid-body matching, as explained hereafter. Let V be a reference volume with a high SNR (e.g., a subtomogram average), and let r be the (x, y, z) coordinates of V. Moreover, let S be an MW-corrected and rigid-body aligned subtomogram. Then, the following relationship between V and S is valid:

$$S = V(r + \delta_O(r) + \delta_C(r) + \delta_A(r)) + N \tag{6.3}$$

Where δ_O represents the voxel-to-voxel relationship between V and S to have an ideal rigid-body alignment, i.e., it stands for the rigid-body alignment imperfections of S; δ_C represents the relationship between the voxels of V and S to have an ideal elastic matching, i.e., it stands for the conformational variability of the subtomograms with respect to the reference; δ_A represents the residual anisotropies of the subtomogram after MW and Contrast Transfer Function (CTF) correction; N is the subtomogram background noise.

3D dense OF between V and S can provide an estimate of the three voxel relationships combined, i.e., $\delta_O + \delta_C + \delta_A$, challenged by the noise N. Luckily, recent 3D dense OF implementations are loyal to the signal and can operate under very low SNR, especially when helped by a mask that eliminates the background. We can apply algorithms that can minimize the data anisotropies, i.e., δ_A, mainly in terms of MW correction [40, 41] and 3D CTF correction

[101, 143]. Hence, 3D dense OF between V and S in the aforementioned conditions is an estimate of their rigid-body and elastic relationships combined:

$$OF(V,S) \approx \delta_O(r) + \delta_C(r) \tag{6.4}$$

When OF is calculated, it can be applied to the voxels of V to estimate S; this operation is called warping, and the result $\hat{S}$ will be an estimate of S with high SNR that we will refer to as a "matched" subtomogram. An illustration of OF calculation and its usage in matching subtomograms is shown in Figure 53.

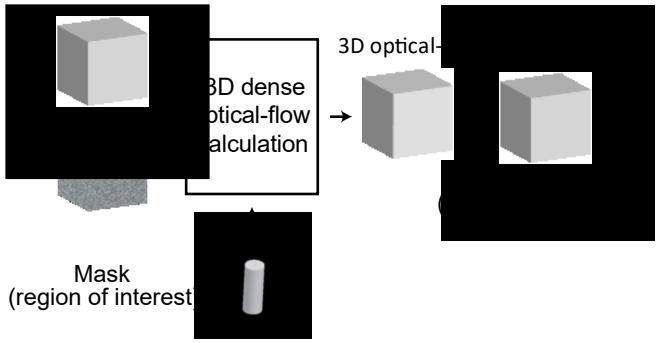


Figure 53 Illustration of the employment of 3D dense OF for elastic and rigid-body matching of subtomograms: V is a volume with a high SNR (e.g., a subtomogram average). S is a volume with a low SNR and contains a similar object as V but at a different conformation and a slightly different orientation and position (e.g., a MW-corrected subtomogram that was rigid-body aligned but not perfectly). $\hat{S}$ is an estimation of S found by warping V using 3D OF, i.e., $\hat{S}$ is a matched version of S using V and the OF.

MW correction and refining the rigid-body alignment

In conventional StA and classification, a compensation for the MW is commonly performed using a scoring function that operates in the Fourier space region excluding the MW [144]. However, for analyzing individual cryo-ET subtomograms in real space, the MW artifacts should be corrected.

In the previous subsection, we have shown that MW correction is needed to minimize the data anisotropies of an analyzed subtomogram, i.e., δ_A in eq (6.3). Also, we have shown that the 3D dense OF can match a subtomogram with a reference in terms of the rigid-body and elastic relationships combined, i.e., $\delta_O(r) + \delta_C(r)$ in eq (6.4). Therefore, to analyze the conformational variability of subtomograms, it is essential to correct the MW and disentangle between OF's rigid-body and elastic matching, which we will discuss in this subsection.

Several methods for MW correction were proposed in the literature [5, 145] and any of them could be used in conjunction with the proposed method. Here, we use a simple method to fill the MW of each subtomogram in Fourier space by the corresponding section from the aligned average, which was initially implemented in Eman2 [146]. We incorporate the MW correction in an iterative rigid-body refinement procedure based on OF subtomogram matching. The procedure is shown in Figure 54 has the following steps:

Step 1. Rigid-body alignment: this can be achieved using StA methods [48, 50, 58, 146, 147] to obtain a table of rigid-body parameters (angles and shifts) that can align the subtomograms to a global subtomogram average. Here, we use reference-free rigid-body alignment using the StA protocol in [48].

Step 2. MW correction of subtomograms: this can be done using any MW correction algorithm (e.g., LoTToR [40]). Here, we fill the subtomogram MW region in Fourier space by the corresponding region of an aligned subtomogram average, as initially implemented in Eman2 [146].

Step 3. Alignment of the MW-filled subtomograms with the average using the StA table: after filling the MW of the subtomograms, we apply the rigid-body alignment of the latest StA table and obtain MW-corrected and aligned subtomograms.

Step 4. Calculation of 3D OF between the subtomogram average and each MW-filled and aligned subtomogram: this should be done in the presence of a mask that determines the region of interest, which can be obtained by thresholding the subtomogram average and applying morphological operations such as dilating and closing. At this step, we also calculate warped versions of the subtomogram average using 3D OF calculated for each subtomogram. These warped versions of the subtomogram average are referred to as "matched subtomograms".

Step 5. Rigid-body alignment of matched subtomograms against the subtomogram average: this step disentangles the rigid-body and elastic matchings of OF by searching for rigid-body alignment of matched subtomograms against the subtomogram average. We perform this step using Fast Rotational Matching (FRM) [48].

Step 6. Updating the table of rigid-body alignment and calculating a new subtomogram average: this is done by combining the initial rigid-body alignment parameters (from the StA table) with the rigid-body refinement parameters obtained in the previous step, which is done by multiplying the corresponding rotational matrices and finding the rigid-body parameters for the resultant matrix [148].

This process can be repeated (1-3 times is usually enough), restarting at **Step 2**, which results in subtomograms that are MW-corrected and whose rigid-body alignment is refined.

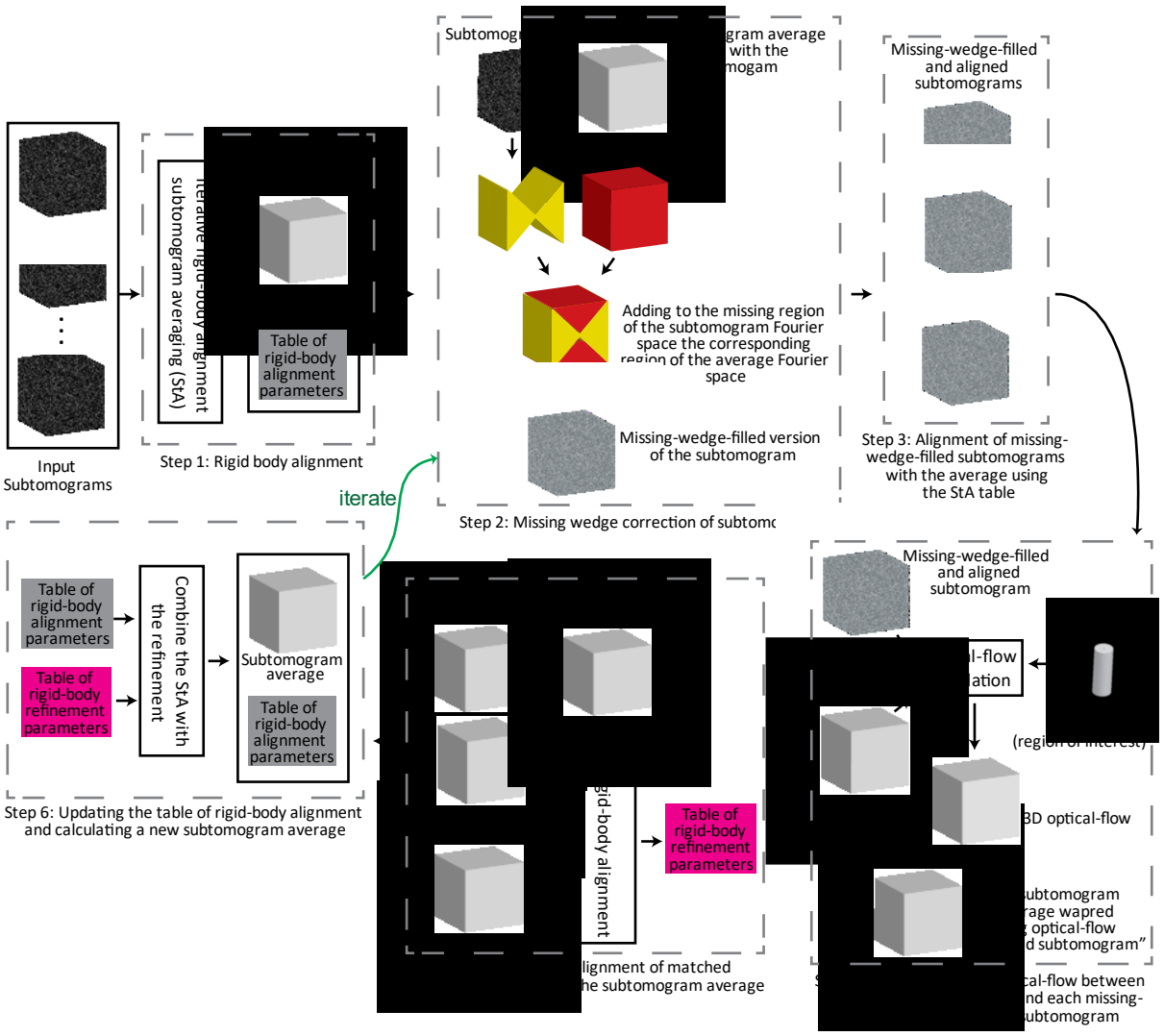


Figure 54 Pipeline for rigid-body alignment of subtomograms, MW correction, and refinement of the rigid-body alignment based on OF subtomograms matching.

Analyzing the continuous conformational variability based on OF

Assume that a set of subtomograms $\{S_i\}$ has undergone MW correction and rigid-body alignment and refinement, following the procedure presented in the previous subsection and Figure 54. Then let V be a reference for the target macromolecule contained in $\{S_i\}$, e.g., the corresponding subtomogram average of $\{S_i\}$. Concurrently, OF calculation in the presence of the mask of the region of interest, between V and each subtomogram in $\{S_i\}$, will mainly stand for the term $\delta_C(r)$ in equation (4) since $\delta_O(r)$ was minimized as a result of the rigid-body

refinement. In other words, the OF after MW correction and rigid-body alignment refinement represents the elastic matching relationship between the voxels of the V and S_i.

Now let set $\{F_i\}$ be the set of OFs between V and $\{S_i\}$:

$$F_i = OF(V, S_i) \tag{6.5}$$

Let $S_i \in R^{l*m*n}$, then $F_i \in R^{3*l*m*n}$ since OF gives a 3D vector for each voxel in V to its matching voxel in S_i. We note here that cryo-ET subtomograms are usually cubic volumes, therefore $l = m = n$.

The corresponding Gram matrix G of $\{F_i\}$ can be defined as:

$$G_{i,j} = vec(F_i)^T * vec(F_j) \tag{6.6}$$

Where $vec(.)$ is the vectorization operation, i.e., it reshapes the matrix to a single column.

Once the Gram Matrix is found, a dimensionality reduction technique can be applied (e.g., PCA) to obtain an essential conformational space.

Interactively processing the conformational space by selective 3D averages and animating trajectories

In the conformational space, each point assigns an OF, which in turn assigns a subtomogram. Close points represent subtomograms of similar conformations and vice versa. Dense regions in the conformational space can be grouped interactively, and the corresponding subtomograms can be averaged. Comparing the subtomogram averages from different groups can help understand the conformational changes of the complex in the given set of subtomograms.

Data distribution paths (trajectories) can be interactively determined in the conformational space, by choosing a set of points $\{P_i\}$ across the data distribution. The motion of the macromolecule can be obtained by displacing the reference (e.g., the subtomogram average) along the trajectory determined by points $\{P_i\}$, as explained below.

Once a trajectory is determined in the conformational space (points $\{P_i\}$ are chosen), the inverse mapping should be applied (e.g., using inverse PCA) on $\{P_i\}$, and the result will be a set of vectors $\{\hat{G}_i\}$ of the same length as the columns of the Gram matrix G given in equation (6). To proceed on how animations can be obtained, we need to rewrite eq (6.6) alternatively. Let O be the matrix of vectorized OFs in its columns as follows:

$$O_i = vec(F_i) \tag{6.7}$$

Then, G can be written as:

$$G = O^T * O \tag{6.8}$$

Hence, any column of G can be expressed as:

$$G_i = O^T * vec(F_i) \tag{6.9}$$

We take advantage of the representation of G in eq (6.9) to approximate a set of OFs $\{\hat{F}_i\}$ that correspond to $\{\hat{G}_i\}$ as follows:

$$vec(\hat{F}_i) \sim (O^T)^+ * \hat{G}_i \tag{6.10}$$

Where the $(.)^+$ is the Moore-Penrose matrix pseudoinverse operation

The retrieved set of $\{vec(\hat{F}_i)\}$ can be reshaped to OFs:

$$\hat{F}_i = vec^{-1}_{3*l*m*n}\big(vec(\hat{F}_i)\big) \tag{6.11}$$

The set of retrieved OFs, i.e., $\{\hat{F}_i\}$, can be used to warp the input reference, which will generate a set of Trajectory Volumes $\{\widehat{TV}_i\}$ that represent the set of trajectory points $\{P_i\}$. Finally, displaying $\{\widehat{TV}_i\}$ shows a movie-like animation of the reference while traversing the selected trajectory.

Results

This section first provides a step-by-step showcase and evaluation of the proposed method, TomoFlow, on simulated datasets. Then, it shows an application of TomoFlow on an experimentally obtained dataset for nucleosomes *in situ*.

Tests on simulated datasets with continuous and discrete conformational variability

The experiments presented in this section were carefully designed to demonstrate the ability of the method to retrieve continuous and discrete conformational variabilities under simulated microscope conditions. These experiments are not claimed to be realistic in terms of their biological significance; rather, they are as realistic as possible in terms of the sophistication of simulating noise, MW artifacts, CTF, and radiation damage compared to other works [47, 51, 54]. For a quantitative assessment of the algorithm for mapping conformations while disentangling it from the subtomographic-approach limitations such as MW and rigid-body (angular and shift) variability, the reader is referred to Chapter 4.

Simulating datasets with discrete and continuous macromolecular conformational variability

In order to test the proposed method, we synthesized two conformationally different datasets, each with different noise intensities.

The first dataset simulates discrete conformational variability. It was created using Normal Mode Analysis (NMA) [149]. We will call this dataset the "NMA-dataset". This dataset comprises 999 subtomograms at three simulated conformations of chain A of the atomic PDB:4AKE structure of adenylate kinase. More precisely, we synthesized 333 subtomograms for each of the three conformations simulated using normal modes 7 and 8 of chain A of the atomic PDB:4AKE structure The three conformations in this dataset correspond to the following normal mode amplitudes $(\text{mode } 7, \text{mode } 8) \in \{(-100, 0), (100, 0), (0, 100)\}$. The ground-truth conformational space for NMA-dataset is determined by the amplitudes along normal modes 7 and 8. A visual representation of this space and the conformations it contains is presented in Figure 55A.

The second dataset simulates continuous conformational variability. It was created using Molecular Dynamics (MD) [150]. We will call this dataset the "MD-dataset". MD is a simulation approach for exploring conformational dynamics by generating trajectories describing a structure evolving over time. This dataset comprises 1000 subtomograms representing a continuum of conformations generated using an MD trajectory between two conformations of adenylate kinase chain A from the PDB structures PDB:4AKE (most open conformation) and PDB:1AKE (most closed conformation). The MD trajectory was simulated

using GENESIS [151] by Rémi Vuillemot, another Ph.D. student in the team at IMPMC. The ground-truth conformational space of the MD-dataset will be presented here by its first two principal axes (PCA). A visual representation of this space and the conformations it contains is presented in Figure 55B.

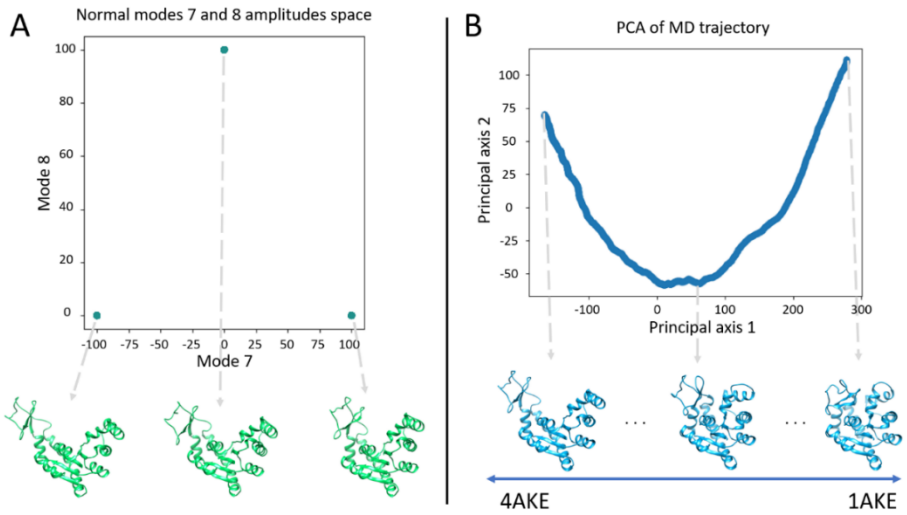


Figure 55 Ground-truth conformational spaces for the two simulated datasets. (A) NMA-dataset: mode 7 and 8 amplitude space with the corresponding three conformations that it contains. (B) MD-dataset: principal axes 1 and 2 showing a continuum of conformations (MD trajectory) between the PDB structures 4AKE (most open conformation) and 1AKE (most closed conformation).

While simulating data, we followed the best practices presented in the literature [47, 51, 54, 120] to make the data challenging while keeping the objectives clear. For each subtomogram, we convert the PDB structure that represents the desired conformation (i.e., either one of the three conformations in the NMA-dataset or one of the continuum of conformations in the MD-dataset) to a volume of size 64^3 voxels and voxel size of 2.2 $Å^3$ [136]. Then, we low-pass filter the volume to 6 Å resolution in order to simulate radiation damage and other effects such as data misalignments incorporated at the tomogram reconstruction step (skipping low-pass filtering would result in better retrieval of conformational variability but is less realistic). Afterward, we rotate and shift this volume in 3D space using random Euler angles and random x, y, z shifts in the radius of 5 voxels from the center. To obtain a tilt series, we project the rotated and shifted volume using tilt values -60° to +60° with 2° step. We simulate microscope conditions by adding noise and modulating the tilt series with a CTF of defocus -1 µm. Then we add noise again (a part of the noise will be modulated by the CTF, and the other part will not). The same procedure is repeated for three different SNR values (0.1, 0.03, 0.01)

and without noise. Then, we invert the CTF phase. Finally, we reconstruct volumes (synthetic subtomograms) using a Fourier reconstruction method [20]. Figure 56 shows four examples of the simulated subtomograms (without noise and at the three different SNR values) for the same conformation, orientation, and position of the macromolecule, along with the corresponding ideal (ground-truth) density volume (the volume with no noise and no missing wedge artifacts, which is not a result of the reconstruction but obtained by converting the atomic structure of that conformation). To give the reader an idea of the resolutions of these subtomograms, we compared them with the ground-truth volume of the same conformation and found the resolutions of 6.4 Å, 13.9 Å, 19.9 Å and 23.6 Å for the simulated subtomogram without noise and with SNR of 0.1, 0.03, and 0.01, respectively (volumes in Figure 56), based on the Fourier Shell Correlation (FSC) between the non-masked volumes and the FSC threshold of 0.5.

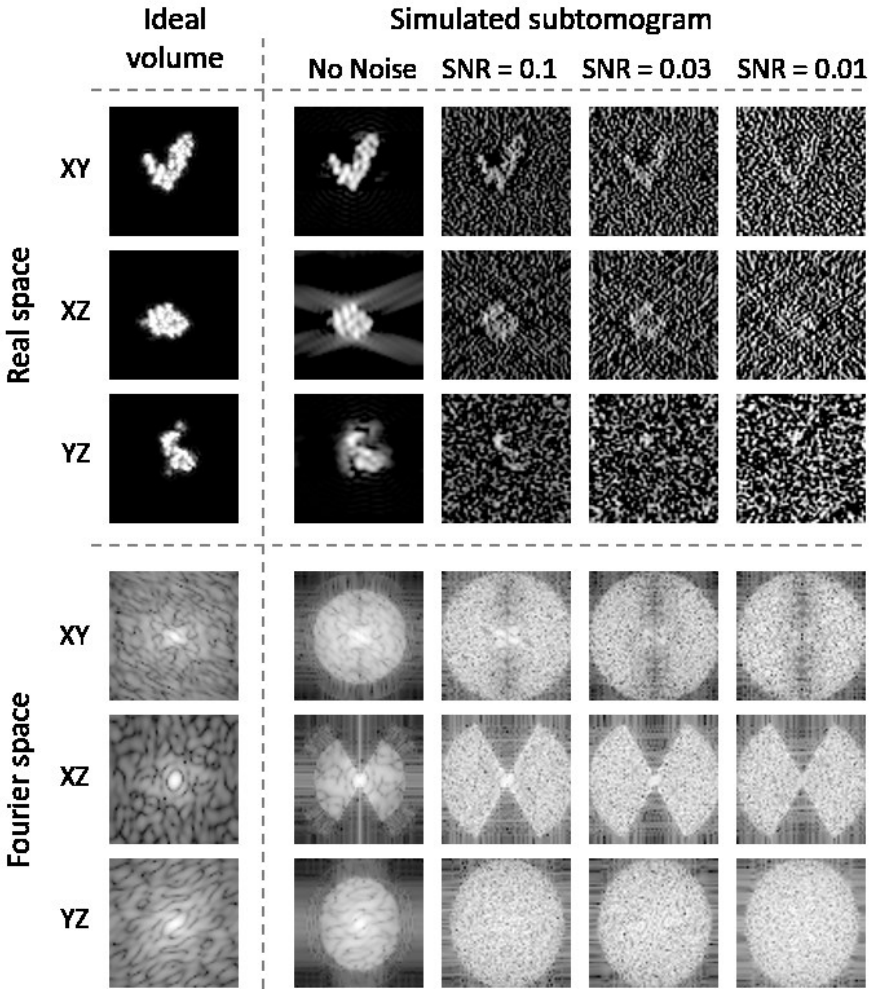


Figure 56 Central slices in real and Fourier spaces of a simulated subtomogram without noise and at different SNRs compared to the corresponding ideal volume.

Rigid-body alignment and refinement with MW correction

We applied the proposed in Figure 54 for rigid-body alignment and refinement with MW correction on the simulated datasets as follows:

Step 1. Performed reference-free rigid-body alignment using the StA protocol in [48]. This StA protocol uses an exhaustive angular search (with FRM), a shifting search within a region of interest, and MW compensation. The shifting search was set to the range of 10 voxels from the center, and the maximum searched normalized frequency to 0.25. The iterative alignment was performed for 15 iterations (complete stability was achieved in the rigid-body alignment parameters and the resulting average).

Step 2. Filled the MW region in Fourier space for each subtomogram by the corresponding region of the aligned subtomogram average.

Step 3. Rigid-body aligned the MW-corrected subtomograms.

Step 4. Calculated the 3D OFs between the subtomogram average and each MW-filled and aligned subtomogram using Farneback-3D after multiplying both volumes with a mask. We generated this mask by binarizing the subtomogram average, dilating it by a structural element of size three, keeping its largest connected component, and smoothing its boundaries with a Gaussian filter of standard deviation equal to two. This step results in "matched subtomograms". In all experiments in this article, Farneback-3D was run with i) a 2-level volume pyramid of scaling factor of 0.5 (meaning a pyramid with the levels of 64^3 and 32^3 voxels in the case of these two test datasets), ii) a window size of $10 \times 10 \times 10$ voxels for integrating the displacement field over a neighborhood of each voxel, iii) 10 iterations of the algorithm at each pyramid level, and with default values of all other parameters. It should be noted that 32^3 voxels is the coarsest pyramid level allowed by Farneback-3D.

Step 5. Performed rigid-body alignment of matched subtomograms against the subtomogram average using FRM (the same method used for StA but here without MW compensation) in the range of 4 voxels from the center.

Step 6. Combined the StA table with the rigid-body refinement parameters obtained in the previous step and calculated a new subtomogram average (refined reference).

We iterated the MW correction and rigid-body refinement process by first performing Step 1 once and Steps 2-6 three times. Table 7 shows the obtained rigid-body alignment results before and after applying this refinement algorithm. They show that the refinement globally

reduces the distances between the estimated and ground-truth rigid-body parameters (angles and shifts) in the presence of noise and conformational variability (Table 7).

Conformational variability analysis

The datasets are ready for continuous conformational variability analysis after applying the MW correction and rigid-body refinement algorithm. Subsequently, we calculated the OFs between the refined reference and the MW corrected and rigid-body aligned subtomograms for each dataset. Then, we found the Gram matrix of OFs based on equations (6.5 - 6.6) and applied PCA. The conformational space, represented by the space of the first two principal vectors, for each dataset, at different noise intensities, is shown in Figure 57. A comparison between the ground-truth conformational spaces in Figure 55 and the retrieved conformational spaces in Figure 57 shows that i) for the NMA-dataset, the separation between the three conformations in the retrieved conformational space is evident for all the tested noise intensities, and ii) for the MD-dataset, the trajectory is more evident for lower noise intensity (higher SNR), and it is the least evident when the SNR is 0.01.

Table 7 Mean and standard deviation (STD) of the absolute distance between ground-truth and estimated rigid-body parameters via StA before and after the proposed rigid-body refinement algorithm applied to NMA-dataset and MD-dataset.

Dataset	Noise	Before/After Refinement	Angular distance [deg]		Shifting distance [vox]	
			Mean	STD	Mean	STD
NMA-dataset	No Noise	Before	2.8	1.5	1.9	0.2
		After	**2.5**	**1.4**	**0.9**	**0.1**
	SNR = 0.1	Before	2.8	1.5	1.2	0.2
		After	**2.5**	**1.3**	**1.2**	**0.2**
	SNR = 0.03	Before	3.1	3.1	1.3	0.4
		After	**2.4**	**2.3**	**1.2**	**0.4**
	SNR = 0.01	Before	17.0	39.3	2.7	2.5
		After	**16.5**	**39.7**	**2.6**	**2.4**
MD-dataset	No Noise	Before	3.3	2.5	1.4	0.2
		After	**3.1**	**2.4**	**1.0**	**0.2**
	SNR = 0.1	Before	2.6	1.6	1.5	0.2
		After	**2.5**	**1.7**	**1.5**	**0.2**
	SNR = 0.03	Before	4.0	2.9	1.6	0.2
		After	**3.9**	**2.7**	**1.4**	**0.2**
	SNR = 0.01	Before	4.8	3.4	1.2	0.4
		After	**4.5**	**3.1**	**1.0**	**0.3**

To give the reader a sense of what the method can achieve when applied to challenging datasets expected in experimental studies, we will base our evaluation of the retrieved

conformations for both datasets on the most challenging noise case (SNR = 0.01), in which the molecule is barely visible in the subtomograms (Figure 56).

Conformational variability analysis for the NMA-dataset

This subsection presents the results of analyzing the conformational variability in the NMA-dataset at SNR = 0.01 using TomoFlow. Figure 58 presents the retrieved conformational space of this dataset, highlighting three distinct groups of points in this space and their corresponding subtomogram averages. Moreover, each group average is compared with its ground-truth atomic structure at the corresponding conformation by docking this atomic structure into the average volume and displaying the volume at 40% opacity.

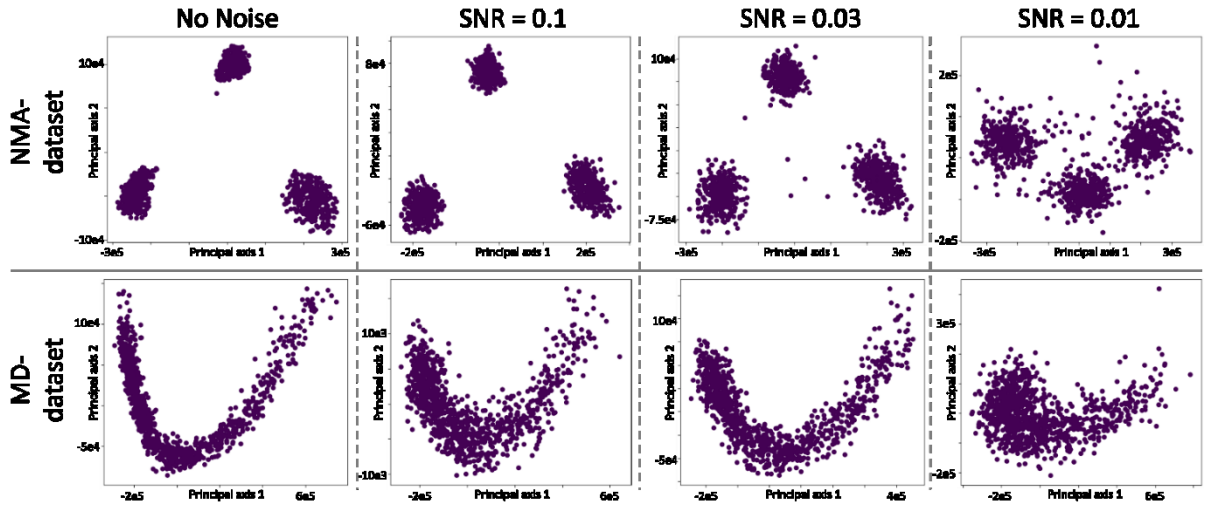


Figure 57 Plots showing the output conformational spaces found by TomoFlow on NMA-dataset and MD-dataset for different noise intensities. The ground-truth conformational spaces for these datasets are shown in Figure 55. We note here that only the distribution should be compared with the ground-truth, which indicates that the inter-relationship between the conformations was retrieved correctly (i.e., similar conformations were mapped to close points and vice versa); the limits of the horizontal and vertical axes do not correspond to those of the ground-truth since the ground-truth conformational space relates atomic structures (NM amplitudes or PCA of MD trajectory) and retrieved conformational space relates OFs.

We compared the three obtained subtomogram averages with the ground-truth volumes of the corresponding conformations (the atomic structures of these conformations converted into volumes), based on the FSC between non-masked volumes and the FSC threshold of 0.5. The obtained resolutions of the three volumes from left to right in Figure 58 are 9.6 Å, 9.5 Å, and 9.5 Å, respectively. Note here that each of the three volumes was obtained by averaging around 300 subtomograms (Figure 58) and recall that the resolution of an individual

subtomogram is around 24 Å for SNR = 0.01. Thus, we observe that the resolution was improved by more than 50 % by averaging only 300 subtomograms, which were aligned in terms of molecular conformation, orientation and position using TomoFlow.

Since the conformational variability in this dataset is discrete, animations are not presented for this dataset.

Conformational variability analysis for the MD-dataset

This subsection presents the results of analyzing the conformational variability in the MD-dataset at SNR = 0.01. Figure 59A presents the retrieved conformational space of this dataset, highlighting six selected groups of subtomograms in this space and their corresponding averages. Moreover, each group average is compared with the ground-truth atomic structure found as the group's centroid at the corresponding conformation by docking this atomic structure in the average volume and displaying the volume at 40% opacity (Figure 59A). Regarding the number of groups and their locations, it is encouraged to try more and fewer groups, which may help to better understand the conformational variability. TomoFlow software (so is the software of HEMNMA-3D in Chapter 4) provides a graphical interface for an interactive selection of the regions in the low-dimensional conformational space in which subtomograms will be summed and their averages computed. Usually, the averages will be calculated from the densest regions (the regions with the largest numbers of points, i.e., subtomograms). The density of points can be visualized using different shades of coloring the points (from the lowest density indicated by the lightest color to the highest density indicated by the darkest color). The size (radius) of each subtomogram averaging region in the conformational space should be selected carefully. Indeed, small-radius regions should still contain enough subtomograms to produce subtomogram averages with sufficiently attenuated noise and MW-induced deformations. Also, large-radius regions should not result in smooth subtomogram averages because the conformational differences between such smooth averages from different regions may not be distinguishable. In general, the higher the resolution and the number of subtomograms, the more it is possible to select denser regions of smaller radii and reveal the conformational variability of the targeted macromolecule, and vice versa.

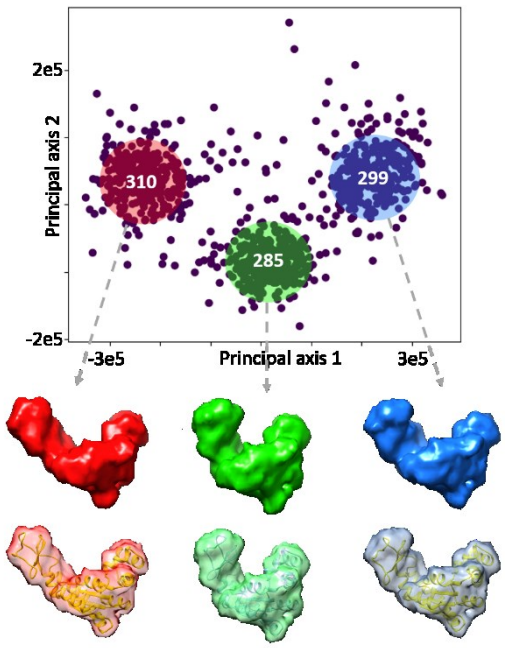


Figure 58 The conformational space found using TomoFlow on the NMA-dataset with SNR = 0.01 (the ground-truth conformational space is shown in Figure 55). The shown volumes are averages of three groups of subtomograms identified by the highlighted ellipses. The number shown inside an ellipse corresponds to the number of points it encloses. The bottom row displays the averages at 40% opacity with their corresponding ground-truth atomic structure docked inside for comparison.

The averages obtained from the six selected groups of subtomograms in Figure 59A (the selected regions of the conformational space of the MD-dataset at SNR = 0.01) were compared with the corresponding ground-truth volumes (the volumes obtained by converting the atomic structure of the group's centroid at the corresponding conformation), based on the FSC between non-masked volumes and the FSC threshold of 0.5. The obtained resolutions of the six volumes from left to right in Figure 59 are 9.8 Å, 9.7 Å, 9.7 Å, 9.8 Å, 11.6 Å, and 15.8 Å, respectively. It can be noted that these resolutions are correlated with the numbers of subtomograms averaged in each group (the numbers shown in Figure 59A). For instance, the subtomogram average of the lowest resolution (15.8 Å) was obtained from the lowest number of subtomograms (80). Also, note that the resolution is 11.6 Å for averaging 101 subtomograms and it is below 10 Å for averaging 112 subtomograms (the averaging of 112-185 subtomograms resulted in the resolution of 9.7-9.8 Å). Thus, we observe that the resolution can improve by more than 50 % with respect to the resolution of an individual subtomogram (24 Å for SNR = 0.01) by averaging as little as around 100 subtomograms, if these subtomograms are aligned in terms of molecular conformation, orientation and position using TomoFlow.

Figure 59B presents an animation following the data distribution manifold. This animation is generated by applying the inverse PCA mapping on the identified points (the ten numbered red points shown in the space), then using equations (6.10 - 6.11) to generate the corresponding OF for each point. These generated OFs are then used to warp the subtomogram average (the global average found after refinement) to generate volumes. The latter volumes correspond to the animation frames, shown as the numbered volumes at the bottom row. When these volumes are displayed sequentially, they show an animation that reveals the MD trajectory used to create the data. This animation is provided in the supplementary material of the published article [140] (Supplementary Movie 1).

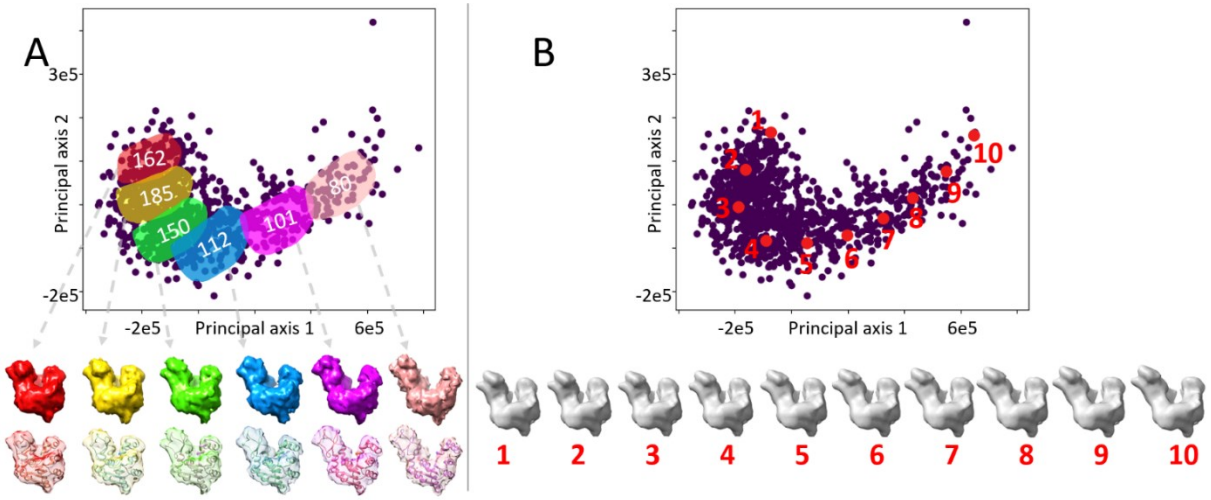


Figure 59 Continuous conformational variability analysis via selective subtomogram averages and animation using TomoFlow conformational space of the MD-dataset with SNR = 0.01. (A) Subtomogram averages of six groups of subtomograms identified by the highlighted areas of the conformational space. The number shown inside a highlighted area corresponds to the number of points it encloses. The bottom row displays the averages at 40% opacity with their corresponding ground-truth atomic structure (group centroid) docked inside for comparison. (B) Displacement of the global subtomogram average along the direction of the data distribution in the conformational space (molecular motion along a trajectory); animation consisting of ten frames represented by a sequence of red dots (from 1 to 10, see also Supplementary Movie 1 in the Supplementary Material of the published article [140]). The ground-truth conformational space is shown in Figure 55.

Conformational variability of nucleosomes *in situ*

This section describes the application and results of TomoFlow on nucleosomes *in situ*, in their interphase nucleus context. We use a dataset containing 666 subtomograms (EMPIAR-10679) of nucleosomes extracted from cryo tomographic reconstruction of a vitreous section of

a high-pressure frozen *Drosophila* embryonic brain [28, 120], with a subtomogram volume size of 64^3 and voxel size of 4.4 $Å^3$ (dataset described in Chapter 4)

First, we used the StA parameters to reproduce the global subtomogram average. We used the average to generate a mask of the region of interest, which was then used to refine rigid-body alignment and analyze the conformations in the data. We generated the mask by binarizing the subtomogram average, dilating it by a structuring element of size three, keeping its largest connected component, and smoothing its boundaries with a Gaussian filter of standard deviation equal to two. Second, we performed seven MW correction and rigid-body refinement iterations following the procedure shown in Figure 54. Third, we analyzed the conformational variability after rigid-body alignment and MW correction of subtomograms. We applied PCA on the Gram matrix of OFs, and the conformational space determined by the first two principal axes is shown in Figure 60.

By inspecting the conformational space, we notice that the first principal axis has significantly larger variability than the second principal axis. We analyzed the variability carried along the first principal axis by analyzing two subtomogram averages and an animation generated along this principal axis. We selectively generated subtomogram averages from groups of points at the beginning and the end of the data distribution. The regions for the groups of points are shown as highlighted areas in Figure 60A, along with their corresponding subtomogram averages. The averages are generated for the MW-corrected and rigid-body aligned subtomograms. Also, we generated animation for the variability along the first principal axis, within the limits of the data distribution manifold represented by the line D in Figure 60B. This animation is generated by estimating the OFs for ten points along line D, then warping the global subtomogram average using these estimated OFs. The resulting volumes are displayed sequentially to generate the animation (Supplementary Movie 2 in Supplementary Material of the published article [140] and Figure 60B).

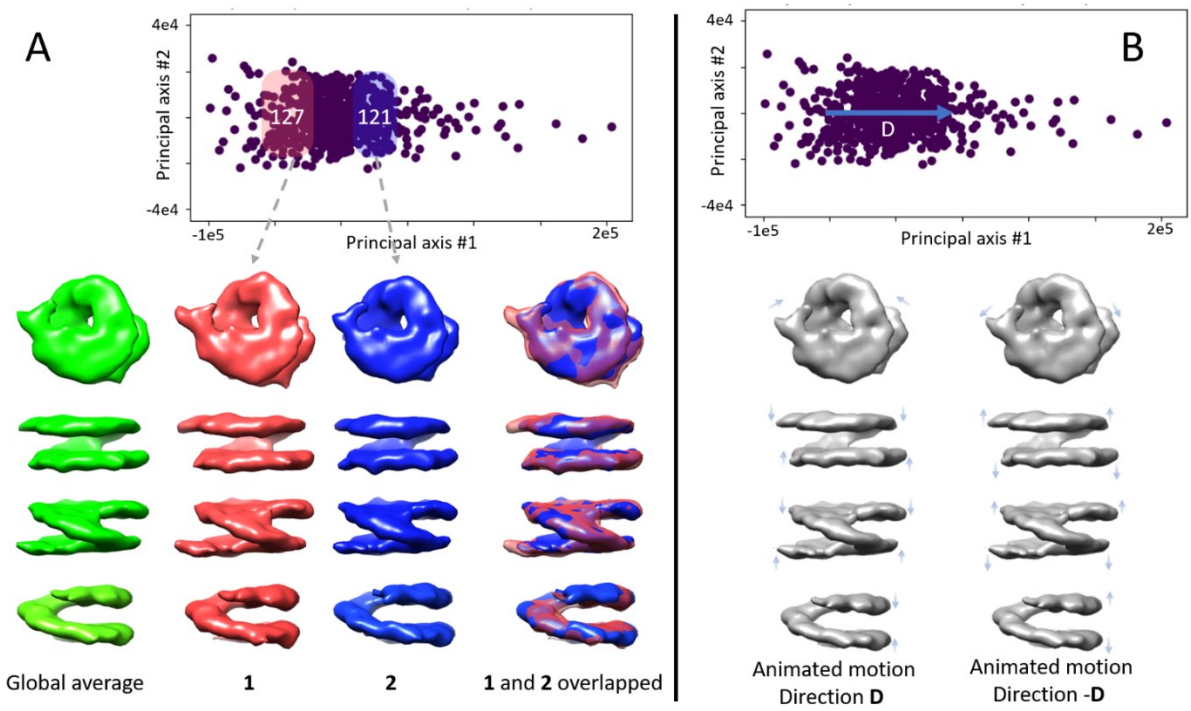


Figure 60 TomoFlow applied to cryo-ET dataset of nucleosomes in situ. (A) group averages for two regions specified by highlighted areas in the conformational space and the corresponding averages. The number inside a highlighted area indicates the number of subtomograms the area encloses. (B) an illustration showing the displacement of the global subtomogram average along the first axis in the limits of the data distribution shown by line D in the conformational space. The arrows on the different views of the global average show the direction of the movement in this animation. The animation is provided in the supplementary material (Supplementary Movie 2 in Supplementary Material of the published article [140]).

The main differences between the two group averages (Figure 60A) and the motion observed along line D (Figure 60B) indicate that TomoFlow detected a combined breathing and gapping motion of the nucleosome, with the breathing motion more expressed. These results are consistent with those of HEMNMA-3D using the same dataset [120] (Chapter 5) but without imposing any prior knowledge (like simulated motion directions by NMA) on the analysis of the conformational variability.

Discussion

This chapter presented TomoFlow method that addresses continuous macromolecular conformational variability captured in cryo-ET subtomograms. TomoFlow employs dense 3D optical flow, posterior to conventional StA, to refine the rigid-body alignment and analyze the conformational variability. The method maps the subtomograms to a space of conformations and allows i) interactively generating subtomogram averages of different conformations, and

ii) navigating the conformational space of the macromolecule via animations based on a reference (e.g., the global subtomogram average).

We presented the method with simplified yet sufficient mathematical derivations that are necessary for understanding and implementing it. We tested the method by synthesizing two datasets under challenging conditions and testing its capability in retrieving their ground-truth conformational spaces. The results of the tests with simulated data indicate that i) the proposed rigid-body refinement can improve the alignment quality in the presence of conformational variability, and ii) the proposed conformational variability analysis can accurately recover hidden conformations. Additionally, we tested TomoFlow using a cryo-ET dataset of nucleosomes *in situ,* which provided promising results coherent with previous findings using the same dataset [120] but without imposing any prior knowledge on the analysis of the conformational variability.

TomoFlow performs the analysis in real space. Hence, it requires MW correction instead of the MW compensation in reciprocal space that is used in HEMNMA-3D. MW correction algorithms exist and TomoFlow can work in conjunction with any of them. Here, we used a method for MW correction based on filling the MW region of subtomograms with the corresponding region of the global subtomogram average. A more advanced MW correction method can be used in the future and might lead to better results.

OF is a powerful and robust image analysis algorithm. Earlier OF approaches were detecting small changes, typically in a few pixels/voxels range. Recent OF methods, such as Farneback-3D used in TomoFlow, cope with this limitation by combining OFs from multiple scales (pyramid-scheme processing). Nevertheless, TomoFlow will be more efficient for smaller conformational variability in the data (for a systematic test of TomoFlow matching different conformational variability magnitudes, the reader can see Chapter 4). Additionally, Farneback-3D method enforces the smoothness of the motion field between the two given volumes (the volumes between which the OF should be calculated), which allows a correct calculation of the OF under very heavy noise and resisting against MW artifacts. However, the OF smoothness enforcement induces smoothness of the generated animation (a smooth version of the warped reference in each frame, e.g., Figure 60B). Finally, obtaining animations requires Moore-Penrose pseudoinverse to be found for a large matrix given in equation (10). This matrix is defined as the matrix of column-wise vectorized OFs that can be three times the dimensions of the input subtomogram dataset; for instance, generating animations after processing a dataset

that comprises 1000 subtomograms of volume size 64^3 will require the inversion of $3*64^3$ rows by 1000 columns, which becomes computationally challenging for large datasets, mainly in terms of the required memory. However, a downsampling (e.g., by 2) of the OFs before reconstructing the matrix significantly reduces the computational requirements, resulting in less detailed animations.

Despite the limitations discussed above, the presented TomoFlow method provides a promising new insight into what can be achieved in cryo-ET studies of macromolecular conformational variability. The advancement in OF development might allow even better TomoFlow performance in the future. TomoFlow is not directly applicable to analyzing 2D images. However, it can analyze 3D volumes reconstructed from 2D images, potentially coming from other cryo-EM modalities, such as single-particle images.

Chapter 7. Software contributions

The methods developed in this thesis were integrated into a software package called ContinuousFlex, along with other methods developed in Dr. Jonic's team for processing cryo-EM/ET data, mainly in terms of biomolecular continuous conformational variability. ContinuousFlex is a plugin of open-source software called Scipion, which is a software that was initially developed for SPA cryo-EM data processing and was recently extended to process cryo-ET data. The backend of Scipion, which is the backend of ContinuousFlex, is implemented in a software package called Xmipp.

During my thesis, I was the principal developer of ContinuousFlex, which we partially described in a journal manuscript during its early development in late 2019 [123], and in another journal manuscript (to date, it is under review) that reviews all the methods of ContinuousFlex. For three years, I was in charge of maintaining ContinuousFlex and assisting other members of Dr. Jonic's team in adding new methods. I also took part in the development of Scipion, particularly its extension to tomography [152], and in the development of Xmipp [153].

This chapter introduces ContinuousFlex, and describes the practical aspects and software of HEMNMA-3D and TomoFlow.

ContinuousFlex

ContinuousFlex is a user-friendly open-source software package primarily developed for obtaining conformational landscapes of macromolecules by an exhaustive analysis of their continuous conformational variability in cryo-EM/ET data (Figure 61a). Additionally, it provides methods for flexible fitting of cryo-EM maps with atomic models (Figure 61b). ContinuousFlex branched from Scipion [154] and its backend software Xmipp [153] in 2019 [123]. It is currently available as a plugin of Scipion, with Xmipp hosting several of its backend data processing steps. This pluginization allowed better maintenance, faster development, and more frequent releases of bug fixes and developed methods. As a plugin of Scipion, ContinuousFlex allows reproducible research, as all the data processing steps used in experiments are automatically stored on the disk (together with their parameters) and can be reproduced at any moment using the same or modified parameters. Additionally, the project

containing all the data processing steps can be directly uploaded to EMPIAR, as allowed by Scipion.

Currently, ContinuousFlex contains software for running (1) recently published methods HEMNMA-3D [120], TomoFlow [140], and NMMD [155]; (2) earlier published methods HEMNMA [43, 123] and StructMap [131]; and (3) methods for simulating cryo-EM and cryo-ET data with conformational variability and methods for data preprocessing [120, 140, 155, 156]. It also includes external software for molecular dynamics simulation (GENESIS [157]) and normal mode analysis (ElNemo [158]), used in some of the mentioned methods. Besides, ContinuousFlex currently also offers a deep learning extension of HEMNMA, named DeepHEMNMA (manuscript under review).

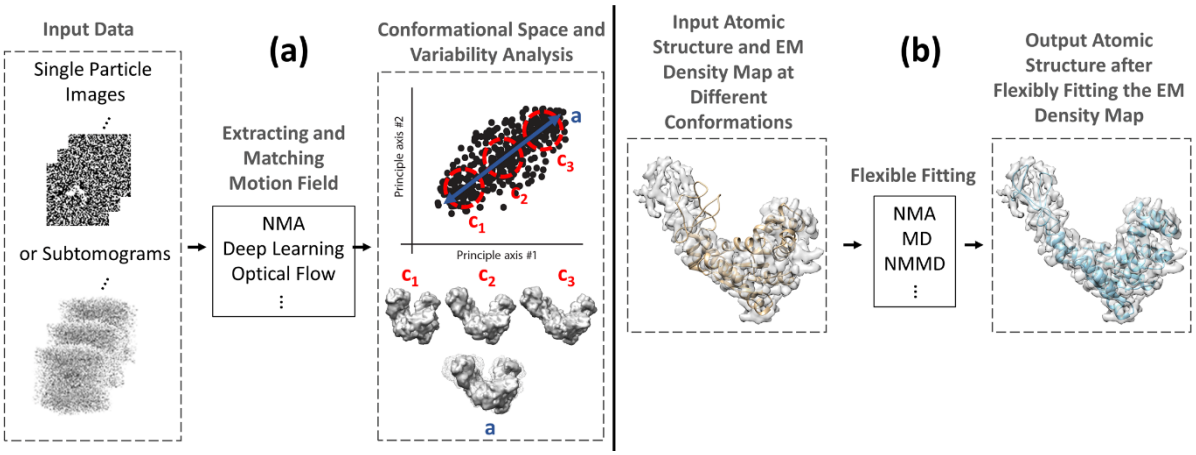


Figure 61 Illustration of the methods in ContinuousFlex: (a) methods used for obtaining conformational landscapes of macromolecules by an exhaustive analysis of their continuous conformational variability in cryo-EM/ET data; (b) methods for flexible fitting of cryo-EM maps with atomic models. The methods in ContinuousFlex are based on normal mode analysis (NMA), molecular dynamics simulation (MD), combination of NMA and MD (NMMD), deep learning, or optical flow

HEMNMA-3D software: Analysis of a set of subtomograms using normal modes

ContinuousFlex allows performing all the steps of HEMNMA-3D, as follows (Figure 62):

1- Importing an atomic structure (1.a) or an EM map (1.b1). If an EM map is imported, it is converted to a pseudoatomic structure (1.b2).
2- Performing NMA.
3- Importing volumes (subtomograms), synthesizing volumes (optional), resizing volumes (optional), masking volumes (optional), and performing traditional subtomogram averaging

(optional). We note here that the protocol "Synthesize volumes" can be used to create data for testing the method and is not a part of data processing, whereas the protocol "Subtomogram averaging" can be used to find the global subtomogram average of the input set (which can serve as a starting reference for HEMNMA-3D). The protocols "Resize volumes" and "Apply mask" are optional data preprocessing techniques.

4- Performing rigid-body and normal-mode-based elastic alignment of the reference model with each subtomogram. At this step, HEMNMA-3D calculates normal-mode amplitudes, three Euler angles, and a 3D shift for each subtomogram.

5- Obtaining and analyzing a low-dimensional conformational space using the normal-mode amplitudes, Euler angles, and shifts estimated for all subtomograms at step 4. This step allows obtaining movies of conformational transitions and subtomogram averages of similar conformations from the conformational space.

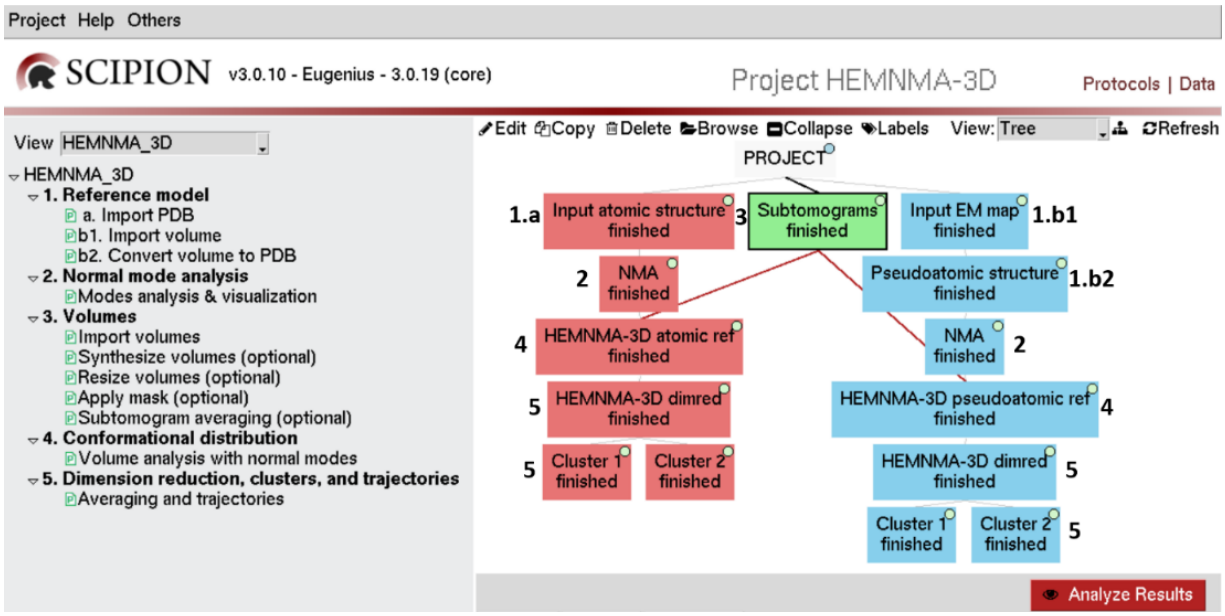


Figure 62 Graphical interface of HEMNMA-3D in ContinuousFlex. Green box: input subtomograms. Red branch: HEMNMA-3D processing with an atomic structure as the reference. Blue branch: HEMNMA-3D processing with an EM map as the reference. None of the tools marked as optional (menu on the left) were used in this figure. The numbers of the steps in the menu on the left are indicated in the tree on the right.

Step 4 of HEMNMA-3D software is MPI parallelized, allowing multiple volumes to be processed simultaneously.

TomoFlow: Analysis of a set of subtomograms using optical flow

ContinuousFlex allows performing all the steps of TomoFlow, as follows (Figure 63):

1- Importing volumes (subtomograms) and some optional tools for creating and applying masks, denoising volumes, missing wedge correction, and synthesizing volumes. The mask tools are useful for creating a mask based on the global subtomogram average and applying it to aligned subtomograms to eliminate the background noise.
2- Importing an external reference (optional) as an EM map or an atomic structure converted to an EM map to perform subtomogram alignment. This is only recommended when reference-free subtomogram alignment fails.
3- Performing or importing subtomogram alignment and averaging, generating a mask for the region of interest (used from Optional tools in Step 1), refining the rigid-body alignment, and filling the missing wedge.
4- Finding the OF between the refined subtomogram average and the missing wedge corrected and aligned subtomograms, then using the OFs to construct a Gram matrix.
5- Obtaining the conformational space (PCA on the Gram matrix) and its interactive analysis by generating movies of conformational transitions along different directions in the space and calculating subtomogram averages of similar conformations.

TomoFlow software is MPI parallelized for rigid-body alignment, and it uses GPU processing for OF calculation.

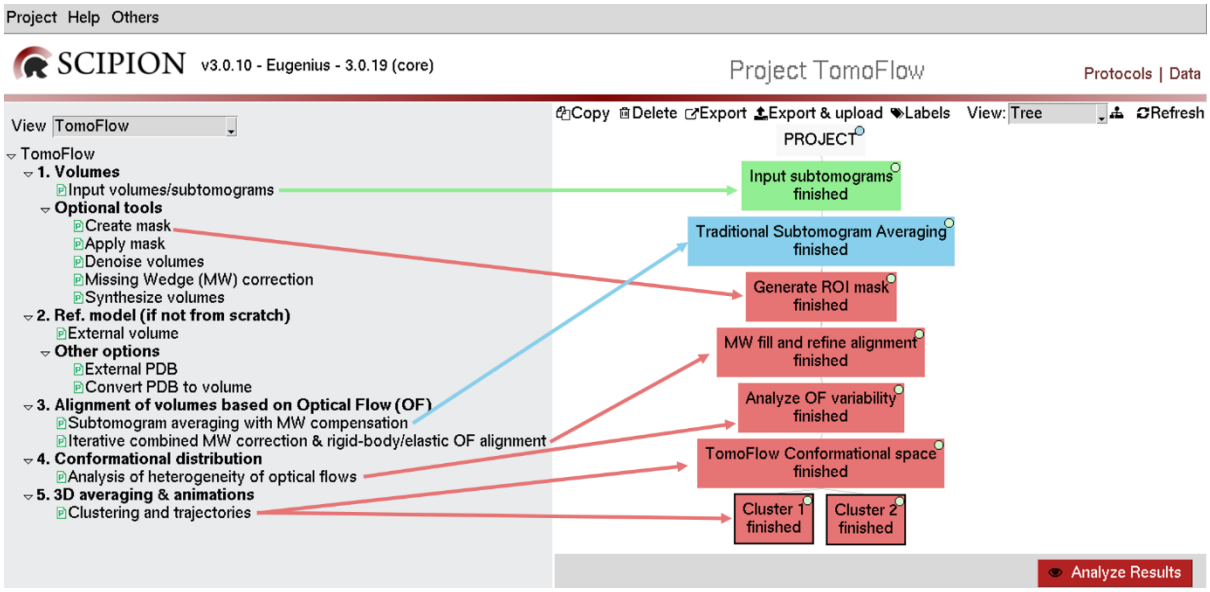


Figure 63 Graphical interface of TomoFlow in ContinuousFlex. Green box: input subtomograms. Blue box: traditional subtomogram averaging (which can also be done using other software packages and imported to Scipion). Red branch: other steps specific to TomoFlow.

Chapter 8. Discussion, conclusion and future work

The future cryo-ET research could be more inspired by the current trend in SPA. The research in SPA has moved towards analyzing continuous conformational variability, starting with a few methods that were sprouted almost a decade ago and kept evolving, such as HEMNMA [43, 123, 124] and Manifold Embedding [159, 160], and passing through a significant amount of new methods that have been developed in the past few years, mostly based on Deep Learning such as cryoSparc [161], Multi-CryoGan [162], cryoDrgn [163, 164] and others [165],[166], including a deep learning extension of HEMNMA referred to as DeepHEMNMA developed in the team at IMPMC (manuscript under review). These methods can potentially be extended to cryo-ET data; however, deep learning methods require large datasets in general to train. Unfortunately, so far, cryo-ET datasets are still of small size in terms of the number of particles, especially in situ. On the other hand, the methods developed in this thesis (HEMNMA-3D and TomoFlow), can deal with continuous conformational variability in cryo-ET data despite the small number of subtomograms since they have mathematical bases to extract interpretable results from datasets of any size and are not based on Deep learning.

It should be reminded that better statistics (including those produced by PCA) are obtained for larger datasets. The cryo-ET datasets that can be collected (thus, analyzed by HEMNMA-3D and TomoFlow) nowadays are still much smaller than those produced by single-particle cryo-EM. However, the goal of the PCA in HEMNMA-3D and TomoFlow is to reveal the major motions of the complex. In this context, a dataset of 2000 subtomograms that can be obtained nowadays may be considered large enough for such PCA and should allow revealing the main motions of the complex. In this thesis, we have shown the results of an experiment with 666 *in-situ* cryo-ET nucleosome subtomograms, which appear to be sufficient for revealing the breathing and gapping motions of the nucleosome, the two main motions of the nucleosome that have also been detected using two different methods *in situ* [28] as well as in a theoretical study [90].

In this thesis, two methods for continuous conformational variability in cryo-ET were introduced. The first method is called HEMNMA-3D, and is based on matching simulated movements using normal mode analysis with experimental subtomograms. It requires a prior selection of a subset of simulated movements (normal modes) that will be a basis for searching the data. The selection of normal modes is a critical step, and is recommended to be performed

using a criterion of low-frequency and high-collectively modes. When this prior is properly selected, HEMNMA-3D allows discovering the full range of macromolecular motion hidden in input data. Nevertheless, when this prior is not well selected (e.g., some important normal modes are not selected), HEMNMA-3D might generate results that might be misinterpreted. The second method is called TomoFlow, and is based on extracting movements from the input data by matching it with a reference based on a three-dimensional dense optical flow (OF) approach. TomoFlow does not use any prior information and is thus less prone to misinterpretation and misuse. However, to encounter large motion magnitudes, TomoFlow results in smoothing the OF and downscaling the true macromolecular motion. HEMNMA-3D and TomoFlow can be applied to the same dataset to cross-validate the obtained results. Both methods were applied to the case study of nucleosomes in cells and showed similar results, coherent with previous findings and theoretical anticipations of nucleosome conformations, mainly showing gapping and breathing motions of the nucleosome.

It should be noted that grouping similar structures and computing their averages for improving SNR, as it is done in the traditional StA workflows, is not the main objective of HEMNMA-3D and TomoFlow. Their main objective is to obtain the conformational landscape that can be easily visualized (in two or three dimensions determined by the first two or three principal axes) and explored in terms of molecular flexibility animations along different directions (animated displacements of a reference conformation). Yet, HEMMNA-3D and TomoFlow allow making such groups of similar structures and computing their averages, but in contrast to the traditional, discrete classification methods, the number of groups in HEMNMA-3D and TomoFlow is not defined prior to the analysis and it is selected according to the conformational distribution in the low-dimensional conformational space. Furthermore, using traditional StA workflows, less dominant conformations are likely to be undiscovered as being wiped out through the global or class averages blindly (no possibility of visualizing all conformations in a common frame and selecting the conformations to average accordingly). On the contrary, HEMNMA-3D and TomoFlow provide a visualization of the full conformational space and, thus, allows discovering less dominant conformations (as less dense regions in this space) and prevents from wiping such conformations out thanks to an interactive selection of the regions from which the averages will be calculated. However, HEMNMA-3D and TomoFlow are not designed for analyzing all types of structural variabilities such as macromolecular disassembly or binding and unbinding of ligands, but it can be combined with discrete classification methods to disentangle such structural variabilities and then analyze

continuous intraclass variability using class averages as references instead of the global subtomogram average.

Each of HEMNMA-3D and TomoFlow has its advantages and drawbacks. Interesting future work is to combine the two methods in single data processing pipeline. The main disadvantage of TomoFlow is that its optical flow backend enforces data smoothness, which helps avoid overfitting and keeps the method robust to noise. However, this smoothness does not allow the discovery of large movements. Relaxing the smoothing effect of optical flow is not possible in the presence of very high noise. The main disadvantage of HEMNMA-3D is that it is dependent on a prior selection of a set of normal modes, which can be subjective, and results may vary when data is analyzed with different prior selections of the modes. Recently, we have developed a way to project optical flow vectors on normal modes, which allows using normal modes in the TomoFlow pipeline without prior selection of a subset of normal modes. The logic is to project the results of TomoFlow on normal modes, use normal modes to update the starting reference, and use TomoFlow again. The disadvantage of this combination is the required processing time, which multiplies by the number of iterations performed (e.g., 5 times the processing time of TomoFlow, if 5 iterations are used).

One of the challenging tasks is cryo-ET data analysis is tilt series alignment. An aligned tilt series, which is used for tomographic reconstruction, can still have some errors that limit the resolution of the tomogram. Hence, research has been going on to refine the tilt series alignment after STA, based on an idea commonly known as “per-particle per-tilt” [59, 134, 147, 167]. This refinement usually involves extracting a subtilt-series for every subtomogram (per particle tilt-series) and using subtomograms as markers to refine the global tilt series alignment based on the information of the STA. The current research direction is moving toward altering the STA procedure to align the per-particle tilt -series (sub-stacks) as images in such a way to have computationally cheaper STA (in terms of disk space and processing time). However, the latter idea is still under development, with a notably advanced work presented in a package called SUSAN (SUbStack ANalysis package, https://github.com/Kudryas-hevLab/SUSAN). HEMNMA-3D and TomoFlow have the potential to be used in refining tilt-series alignments in the presence of continuous conformational variability. Therefore, interesting future work is to adapt HEMNMA-3D and TomoFlow to analyze sub-stacks instead of subtomograms, as in SUSAN.

Finally, interesting work is the development of new methods to describe the displacement field between different molecular conformations, other than normal mode analysis and optical flow used in this thesis.

Bibliography

[1] Hooke R. Micrographia : or, Some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon / By R. Hooke. London: Printed by Jo. Martyn and Ja. Allestry, printers to the Royal Society, and are to be sold at their shop at the Bell in St. Paul's Churchyard; 1665.
[2] contributors W. Antonie van Leeuwenhoek. Wikipedia, The Free Encyclopedia.
[3] Knoll M, Ruska E. Das Elektronenmikroskop. Zeitschrift für Physik. 1932;78:318-39.
[4] Brenner S, Horne RW. A negative staining method for high resolution electron microscopy of viruses. Biochimica et Biophysica Acta. 1959;34:103-10.
[5] Hoppe W, Gassmann J, Hunsmann N, Schramm H, Sturm M. Three-dimensional reconstruction of individual negatively stained yeast fatty-acid synthetase molecules from tilt series in the electron microscope. Hoppe-Seyler's Zeitschrift fur physiologische Chemie. 1974;355:1483-7.
[6] Crowther RA, DeRosier DJ, Klug A. The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. Proceedings of the Royal Society of London A Mathematical and Physical Sciences. 1970;317:319-40.
[7] Cheng Y. Single-particle cryo-EM-How did it get here and where will it go. Science. 2018;361:876-80.
[8] Adrian M, Dubochet J, Lepault J, McDowall AW. Cryo-electron microscopy of viruses. Nature. 1984;308:32-6.
[9] Dubochet J, McDowall A. Vitrification of pure water for electron microscopy. Journal of Microscopy. 1981;124:3-4.
[10] Dubochet J, Lepault J, Freeman R, Berriman JA, Homo J-C. Electron microscopy of frozen water and aqueous solutions. Journal of Microscopy. 1982;128:219-37.
[11] Orlova EV, Saibil HR. Structural Analysis of Macromolecular Assemblies by Electron Microscopy. Chem Rev. 2011;111:7710-48.
[12] Reimer L. Transmission electron microscopy: physics of image formation and microanalysis: Springer; 2013.
[13] Dubochet J, Adrian M, Chang J-J, Homo J-C, Lepault J, McDowall AW, et al. Cryo-electron microscopy of vitrified specimens. Quarterly Reviews of Biophysics. 1988;21:129-228.
[14] Glaeser RM. Invited Review Article: Methods for imaging weak-phase objects in electron microscopy. Review of Scientific Instruments. 2013;84:312.
[15] Voortman LM, Franken EM, van Vliet LJ, Rieger B. Fast, spatially varying CTF correction in TEM. Ultramicroscopy. 2012;118:26-34.
[16] Koning RI, Koster AJ, Sharp TH. Advances in cryo-electron tomography for biology and medicine. Annals of Anatomy - Anatomischer Anzeiger. 2018;217:82-96.
[17] Doerr A. Single-particle cryo-electron microscopy. Nature Methods. 2016;13:23-.
[18] Dobro MJ, Melanson LA, Jensen GJ, McDowall AW. Chapter Three - Plunge Freezing for Electron Cryomicroscopy. In: Jensen GJ, editor. Methods in Enzymology: Academic Press; 2010. p. 63-82.
[19] Ramani Lata K, Penczek P, Frank J. Automatic particle picking from electron micrographs. Ultramicroscopy. 1995;58:381-91.
[20] Sorzano CO, de la Rosa Trevín JM, Otón J, Vega JJ, Cuenca J, Zaldívar-Peraza A, et al. Semiautomatic, High-Throughput, High-Resolution Protocol for Three-Dimensional Reconstruction of Single Particles in Electron Microscopy. In: Sousa AA, Kruhlak MJ, editors. Nanoimaging: Methods and Protocols. Totowa, NJ: Humana Press; 2013. p. 171-93.
[21] Medalia O, Weber I, Frangakis AS, Nicastro D, Gerisch G, Baumeister W. Macromolecular Architecture in Eukaryotic Cells Visualized by Cryoelectron Tomography. Science. 2002;298:1209-13.
[22] Moor H. Theory and Practice of High Pressure Freezing. In: Steinbrecht RA, Zierold K, editors. Cryotechniques in Biological Electron Microscopy. Berlin, Heidelberg: Springer Berlin Heidelberg; 1987. p. 175-91.
[23] Villa E, Schaffer M, Plitzko JM, Baumeister W. Opening windows into the cell: focused-ion-beam milling for cryo-electron tomography. Current Opinion in Structural Biology. 2013;23:771-7.

[24] Al-Amoudi A, Chang J-J, Leforestier A, McDowall A, Salamin LM, Norlén LP, et al. Cryo-electron microscopy of vitreous sections. The EMBO Journal. 2004;23:3583-8.
[25] Al-Amoudi A, Studer D, Dubochet J. Cutting artefacts and cutting process in vitreous sections for cryo-electron microscopy. Journal of Structural Biology. 2005;150:109-21.
[26] HAN H-M, ZUBER B, DUBOCHET J. Compression and crevasses in vitreous sections under different cutting conditions. Journal of Microscopy. 2008;230:167-71.
[27] Pierson J, Ziese U, Sani M, Peters PJ. Exploring vitreous cryo-section-induced compression at the macromolecular level using electron cryo-tomography; 80S yeast ribosomes appear unaffected. Journal of Structural Biology. 2011;173:345-9.
[28] Eltsov M, Grewe D, Lemercier N, Frangakis A, Livolant F, Leforestier A. Nucleosome conformational variability in solution and in interphase nuclei evidenced by cryo-electron microscopy of vitreous sections. Nucleic Acids Res. 2018;46:9189-200.
[29] Grünewald K, Medalia O, Gross A, Steven AC, Baumeister W. Prospects of electron cryotomography to visualize macromolecular complexes inside cellular compartments: implications of crowding. Biophysical Chemistry. 2002;100:577-91.
[30] Lucic V, Förster F, Baumeister W. Structural studies by electron tomography: from cells to molecules. Annual review of biochemistry. 2005;74:833.
[31] Wan W, Briggs JAG. Chapter Thirteen - Cryo-Electron Tomography and Subtomogram Averaging. In: Crowther RA, editor. Methods in Enzymology: Academic Press; 2016. p. 329-67.
[32] Li Z, Chen S, Zhao L, Huang G, Pi X, Sun S, et al. Near-atomic structure of the inner ring of the Saccharomyces cerevisiae nuclear pore complex. Cell Research. 2022.
[33] Levoy M. Volume rendering using the fourier projection-slice theorem: Computer Systems Laboratory, Stanford University; 1992.
[34] De Rosier DJ, Klug A. Reconstruction of Three Dimensional Structures from Electron Micrographs. Nature. 1968;217:130-4.
[35] Mastronarde DN, Held SR. Automated tilt series alignment and tomographic reconstruction in IMOD. Journal of Structural Biology. 2017;197:102-13.
[36] Nogales E, Scheres Sjors HW. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. Molecular Cell. 2015;58:677-89.
[37] Radermacher M. Weighted Back-projection Methods. In: Frank J, editor. Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell. New York, NY: Springer New York; 2006. p. 245-73.
[38] Wolf D, Lubk A, Lichte H. Weighted simultaneous iterative reconstruction technique for single-axis tomography. Ultramicroscopy. 2014;136:15-25.
[39] Andersen AH, Kak AC. Simultaneous Algebraic Reconstruction Technique (SART): A superior implementation of the ART algorithm. Ultrasonic Imaging. 1984;6:81-94.
[40] Zhai X, Lei D, Zhang M, Liu J, Wu H, Yu Y, et al. LoTToR: An Algorithm for Missing-Wedge Correction of the Low-Tilt Tomographic 3D Reconstruction of a Single-Molecule Structure. Scientific Reports. 2020;10:10489.
[41] Moebel E, Kervrann C. A Monte Carlo framework for missing wedge restoration and noise removal in cryo-electron tomography. Journal of Structural Biology: X. 2020;4:100013.
[42] Boisset N, Penczek PA, Taveau J-C, You V, de Haas F, Lamy J. Overabundant single-particle electron microscope views induce a three-dimensional reconstruction artifact. Ultramicroscopy. 1998;74:201-7.
[43] Jin Q, Sorzano Carlos Oscar S, de la Rosa-Trevín José M, Bilbao-Castro José R, Núñez-Ramírez R, Llorca O, et al. Iterative Elastic 3D-to-2D Alignment Method Using Normal Modes for Studying Structural Dynamics of Large Macromolecular Complexes. Structure. 2014;22:496-506.
[44] Behrmann E, Loerke J, Budkevich Tatyana V, Yamamoto K, Schmidt A, Penczek Pawel A, et al. Structural Snapshots of Actively Translating Human Ribosomes. Cell. 2015;161:845-57.
[45] Elad N, Clare DK, Saibil HR, Orlova EV. Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. Journal of Structural Biology. 2008;162:108-20.
[46] Förster F, Medalia O, Zauberman N, Baumeister W, Fass D. Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. Proceedings of the National Academy of Sciences. 2005;102:4729-34.

[47] Xu M, Beck M, Alber F. High-throughput subtomogram alignment and classification by Fourier space constrained fast volumetric matching. Journal of Structural Biology. 2012;178:152-64.
[48] Chen Y, Pfeffer S, Hrabe T, Schuller JM, Förster F. Fast and accurate reference-free alignment of subtomograms. Journal of Structural Biology. 2013;182:235-45.
[49] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems. 1987;2:37-52.
[50] Navarro PP, Stahlberg H, Castaño-Díez D. Protocols for Subtomogram Averaging of Membrane Proteins in the Dynamo Software Package. Frontiers in molecular biosciences. 2018;5:82-.
[51] Förster F, Pruggnaller S, Seybert A, Frangakis AS. Classification of cryo-electron sub-tomograms using constrained correlation. Journal of Structural Biology. 2008;161:276-86.
[52] Hrabe T, Chen Y, Pfeffer S, Kuhn Cuellar L, Mangold A-V, Förster F. PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. Journal of Structural Biology. 2012;178:177-88.
[53] Scheres SHW, Melero R, Valle M, Carazo J-M. Averaging of Electron Subtomograms and Random Conical Tilt Reconstructions through Likelihood Optimization. Structure. 2009;17:1563-72.
[54] Stölken M, Beck F, Haller T, Hegerl R, Gutsche I, Carazo J-M, et al. Maximum likelihood based classification of electron tomographic data. Journal of Structural Biology. 2011;173:77-85.
[55] Scheres SH. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol. 2012;180:519-30.
[56] Förster F, Hegerl R. Structure Determination In Situ by Averaging of Tomograms. Methods in Cell Biology: Academic Press; 2007. p. 741-67.
[57] Scheres SH, Gao H, Valle M, Herman GT, Eggermont PP, Frank J, et al. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat Methods. 2007;4:27-9.
[58] Bharat TAM, Scheres SHW. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. Nature Protocols. 2016;11:2054-65.
[59] Tegunov D, Xue L, Dienemann C, Cramer P, Mahamid J. Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 A in cells. Nat Methods. 2021;18:186-93.
[60] Dong H, Qin S, Zhou HX. Effects of macromolecular crowding on protein conformational changes. PLoS Comput Biol. 2010;6:e1000833.
[61] Zeng X, Xu M. Gum-Net: Unsupervised geometric matching for fast and accurate 3D subtomogram image alignment and averaging. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2020. p. 4073-84.
[62] Zhang L, Ren G. IPET and FETR: Experimental Approach for Studying Molecular Structure Dynamics by Cryo-Electron Tomography of a Single-Molecule Structure. PLOS ONE. 2012;7:e30249.
[63] Bendich AJ, Drlica K. Prokaryotic and eukaryotic chromosomes: what's the difference? BioEssays. 2000;22:481-6.
[64] Olins AL, Olins DE. Spheroid Chromatin Units (ν Bodies). Science. 1974;183:330-2.
[65] Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 1997;389:251-60.
[66] Karch K, DeNizio J, Black B, Garcia B. Identification and interrogation of combinatorial histone modifications. Frontiers in Genetics. 2013;4.
[67] Tan S, Davey CA. Nucleosome structural studies. Current Opinion in Structural Biology. 2011;21:128-36.
[68] Polach KJ, Widom J. Mechanism of Protein Access to Specific DNA Sequences in Chromatin: A Dynamic Equilibrium Model for Gene Regulation. Journal of Molecular Biology. 1995;254:130-49.
[69] Tolsma Thomas O, Hansen Jeffrey C. Post-translational modifications and chromatin dynamics. Essays in Biochemistry. 2019;63:89-96.
[70] Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution††We dedicate this paper to the memory of Max Perutz who was particularly inspirational and supportive to T.J.R. in the early stages of this study. Journal of Molecular Biology. 2002;319:1097-113.
[71] White CL, Suto RK, Luger K. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. The EMBO Journal. 2001;20:5207-18.

[72] Harp JM, Hanson BL, Timm DE, Bunick GJ. Asymmetries in the nucleosome core particle at 2.5 A resolution. Acta Crystallographica Section D. 2000;56:1513-34.
[73] Chua EYD, Vogirala VK, Inian O, Wong ASW, Nordenskiöld L, Plitzko JM, et al. 3.9 Å structure of the nucleosome core particle determined by phase-plate cryo-EM. Nucleic Acids Research. 2016;44:8013-9.
[74] Koopmans WJA, Brehm A, Logie C, Schmidt T, van Noort J. Single-Pair FRET Microscopy Reveals Mononucleosome Dynamics. Journal of Fluorescence. 2007;17:785-95.
[75] Armeev GA, Kniazeva AS, Komarova GA, Kirpichnikov MP, Shaytan AK. Histone dynamics mediate DNA unwrapping and sliding in nucleosomes. Nat Commun. 2021;12:2387.
[76] Andrews AJ, Luger K. Nucleosome structure (s) and stability: variations on a theme. Annual review of biophysics. 2011;40:99-117.
[77] Tachiwana H, Kagawa W, Shiga T, Osakabe A, Miya Y, Saito K, et al. Crystal structure of the human centromeric nucleosome containing CENP-A. Nature. 2011;476:232-5.
[78] Li G, Widom J. Nucleosomes facilitate their own invasion. Nature Structural & Molecular Biology. 2004;11:763-9.
[79] Ordu O, Lusser A, Dekker NH. Recent insights from in vitro single-molecule studies into nucleosome structure and dynamics. Biophysical Reviews. 2016;8:33-49.
[80] Böhm V, Hieb AR, Andrews AJ, Gansen A, Rocker A, Tóth K, et al. Nucleosome accessibility governed by the dimer/tetramer interface. Nucleic Acids Research. 2010;39:3093-102.
[81] Ngo TT, Ha T. Nucleosomes undergo slow spontaneous gaping. Nucleic Acids Res. 2015;43:3964-71.
[82] Mangenot S, Leforestier A, Vachette P, Durand D, Livolant F. Salt-Induced Conformation and Interaction Changes of Nucleosome Core Particles. Biophysical Journal. 2002;82:345-56.
[83] Bilokapic S, Strauss M, Halic M. Histone octamer rearranges to adapt to DNA unwrapping. Nature Structural & Molecular Biology. 2018;25:101-8.
[84] Cutter AR, Hayes JJ. A brief review of nucleosome structure. FEBS Letters. 2015;589:2914-22.
[85] Zhou K, Gaullier G, Luger K. Nucleosome structure and dynamics are coming of age. Nature Structural & Molecular Biology. 2019;26:3-13.
[86] Koyama M, Kurumizaka H. Structural diversity of the nucleosome. The Journal of Biochemistry. 2017;163:85-95.
[87] Zhou M, Dai L, Li C, Shi L, Huang Y, Guo Z, et al. Structural basis of nucleosome dynamics modulation by histone variants H2A.B and H2A.Z.2.2. The EMBO Journal. 2021;40:e105907.
[88] Chittori S, Hong J, Bai Y, Subramaniam S. Structure of the primed state of the ATPase domain of chromatin remodeling factor ISWI bound to the nucleosome. Nucleic Acids Research. 2019;47:9400-9.
[89] Zhou B-R, Jiang J, Ghirlando R, Norouzi D, Sathish Yadav KN, Feng H, et al. Revisit of Reconstituted 30-nm Nucleosome Arrays Reveals an Ensemble of Dynamic Structures. Journal of Molecular Biology. 2018;430:3093-110.
[90] Zlatanova J, Bishop TC, Victor JM, Jackson V, van Holde K. The nucleosome family: dynamic and growing. Structure. 2009;17:160-71.
[91] Bloom K. Centromere dynamics. Current Opinion in Genetics & Development. 2007;17:151-6.
[92] Sivolob A, Lavelle C, Prunell A. Flexibility Of Nucleosomes On Topologically Constrained DNA. New York, NY: Springer New York; 2009. p. 251-91.
[93] Chereji RV, Ramachandran S, Bryson TD, Henikoff S. Precise genome-wide mapping of single nucleosomes and linkers in vivo. Genome Biology. 2018;19:19.
[94] Rhee Ho S, Bataille Alain R, Zhang L, Pugh BF. Subnucleosomal Structures and Nucleosome Asymmetry across a Genome. Cell. 2014;159:1377-88.
[95] Cai S, Böck D, Pilhofer M, Gan L. The in situ structures of mono-, di-, and trinucleosomes in human heterochromatin. Molecular Biology of the Cell. 2018;29:2450-7.
[96] Bednar J, Garcia-Saez I, Boopathi R, Cutter AR, Papai G, Reymer A, et al. Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. Molecular Cell. 2017;66:384-97.e8.
[97] Campos-Ortega JA, Hartenstein V. Stages of Drosophila Embryogenesis. The Embryonic Development of Drosophila melanogaster. Berlin, Heidelberg: Springer Berlin Heidelberg; 1997. p. 9-102.
[98] Mastronarde DN. Automated electron microscope tomography using robust prediction of specimen movements. Journal of Structural Biology. 2005;152:36-51.

[99] Hagen WJH, Wan W, Briggs JAG. Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. Journal of Structural Biology. 2017;197:191-8.
[100] Kremer JR, Mastronarde DN, McIntosh JR. Computer Visualization of Three-Dimensional Image Data Using IMOD. Journal of Structural Biology. 1996;116:71-6.
[101] Kunz M, Frangakis AS. Three-dimensional CTF correction improves the resolution of electron tomograms. Journal of Structural Biology. 2017;197:114-22.
[102] Hansson T, Oostenbrink C, van Gunsteren W. Molecular dynamics simulations. Current Opinion in Structural Biology. 2002;12:190-6.
[103] Skjaerven L, Hollup SM, Reuter N. Normal mode analysis for proteins. Journal of Molecular Structure: THEOCHEM. 2009;898:42-8.
[104] Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Physical review letters. 1996;77:1905.
[105] Tama F, Wriggers W, Brooks CL. Exploring Global Distortions of Biological Macromolecules and Assemblies from Low-resolution Structural Information and Elastic Network Theory. Journal of Molecular Biology. 2002;321:297-305.
[106] Bahar I, Lezon TR, Bakan A, Shrivastava IH. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. Chem Rev. 2010;110:1463-97.
[107] Laub AJ. Matrix analysis for scientists and engineers: Siam; 2005.
[108] Nogales-Cadenas R, Jonic S, Tama F, Arteni AA, Tabas-Madrid D, Vázquez M, et al. 3DEM Loupe: analysis of macromolecular dynamics using structures from electron microscopy. Nucleic Acids Research. 2013;41:W363-W7.
[109] Jonić S, Sorzano CÓS. Coarse-graining of volumes for modeling of structure and dynamics in electron microscopy: Algorithm to automatically control accuracy of approximation. IEEE Journal of Selected Topics in Signal Processing. 2016;10:161-73.
[110] Durand P, Trinquier G, Sanejouand Y-H. A new approach for determining low-frequency normal modes in macromolecules. Biopolymers. 1994;34:759-71.
[111] Tama F, Gadea FX, Marques O, Sanejouand Y-H. Building-block approach for determining low-frequency normal modes of macromolecules. Proteins: Structure, Function, and Bioinformatics. 2000;41:1-7.
[112] Tama F, Miyashita O, Brooks CL. Flexible Multi-scale Fitting of Atomic Structures into Low-resolution Electron Density Maps with Elastic Network Normal Mode Analysis. Journal of Molecular Biology. 2004;337:985-99.
[113] Schröder GF, Brunger AT, Levitt M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure. 2007;15:1630-41.
[114] Lopéz-Blanco JR, Chacón P. iMODFIT: Efficient and robust flexible fitting based on vibrational analysis in internal coordinates. Journal of Structural Biology. 2013;184:261-70.
[115] Shah STH, Xuezhi X. Traditional and modern strategies for optical flow: an investigation. SN Applied Sciences. 2021;3:289.
[116] Lefébure M, Cohen LD. Image Registration, Optical Flow and Local Rigidity. Journal of Mathematical Imaging and Vision. 2001;14:131-47.
[117] Norvig PR, Intelligence SA. A modern approach: Prentice Hall Upper Saddle River, NJ, USA:; 2002.
[118] Hilsmann A, Eisert P. Deformable object tracking using optical flow constraints. 2007.
[119] Farnebäck G. Two-Frame Motion Estimation Based on Polynomial Expansion. In: Bigun J, Gustavsson T, editors. Image Analysis. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 363-70.
[120] Harastani M, Eltsov M, Leforestier A, Jonic S. HEMNMA-3D: Cryo Electron Tomography Method Based on Normal Mode Analysis to Study Continuous Conformational Variability of Macromolecular Complexes. Frontiers in molecular biosciences. 2021;8:663121.
[121] Harastani M, Jonic S. Comparison between HEMNMA-3D and Traditional Classification Techniques for Analyzing Biomolecular Continuous Shape Variability in Cryo Electron Subtomograms. 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)2021. p. 01-4.

[122] Harastani M, Jonic S. Methods for analyzing continuous conformational variability of biomolecules in cryo electron subtomograms: HEMNMA-3D vs. traditional classification. bioRxiv. 2021:2021.10.14.464366.
[123] Harastani M, Sorzano COS, Jonić S. Hybrid Electron Microscopy Normal Mode Analysis with Scipion. Protein Science. 2020;29:223-36.
[124] Sorzano COS, de la Rosa-Trevín JM, Tama F, Jonić S. Hybrid Electron Microscopy Normal Mode Analysis graphical interface and protocol. Journal of Structural Biology. 2014;188:134-41.
[125] Brüschweiler R. Collective protein dynamics and nuclear spin relaxation. The Journal of Chemical Physics. 1995;102:3396-403.
[126] Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Eng. 2001;14:1-6.
[127] Tama F, Charles L. Brooks I. SYMMETRY, FORM, AND SHAPE: Guiding Principles for Robustness in Macromolecular Machines. Annual Review of Biophysics and Biomolecular Structure. 2006;35:115-33.
[128] Suhre K, Navaza J, Sanejouand Y-H. NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. Acta Crystallographica Section D. 2006;62:1098-100.
[129] Wang Y, Rader AJ, Bahar I, Jernigan RL. Global ribosome motions revealed with elastic network model. Journal of Structural Biology. 2004;147:302-14.
[130] Delarue M, Dumas P. On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. Proceedings of the National Academy of Sciences. 2004;101:6957-62.
[131] Sanchez Sorzano Carlos O, Alvarez-Cabrera Ana L, Kazemi M, Carazo Jose M, Jonić S. StructMap: Elastic Distance Analysis of Electron Microscopy Maps for Studying Conformational Changes. Biophysical Journal. 2016;110:1753-65.
[132] Vanden Berghen F, Bersini H. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. Journal of Computational and Applied Mathematics. 2005;181:157-75.
[133] Mahalanobis PC. On the generalized distance in statistics. National Institute of Science of India; 1936.
[134] Galaz-Montoya JG, Flanagan J, Schmid MF, Ludtke SJ. Single particle tomography in EMAN2. Journal of Structural Biology. 2015;190:279-90.
[135] Müller CW, Schlauderer GJ, Reinstein J, Schulz GE. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure. 1996;4:147-56.
[136] Peng L-M, Ren G, Dudarev S, Whelan M. Robust parameterization of elastic and absorptive electron atomic scattering factors. Acta Crystallographica Section A: Foundations of Crystallography. 1996;52:257-76.
[137] Iwasaki W, Miya Y, Horikoshi N, Osakabe A, Taguchi H, Tachiwana H, et al. Contribution of histone N-terminal tails to the structure and stability of nucleosomes. FEBS Open Bio. 2013;3:363-9.
[138] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
[139] Kufareva I, Abagyan R. Methods of protein structure comparison. Homology Modeling: Springer; 2011. p. 231-57.
[140] Harastani M, Eltsov M, Leforestier A, Jonic S. TomoFlow: Analysis of Continuous Conformational Variability of Macromolecules in Cryogenic Subtomograms based on 3D Dense Optical Flow. J Mol Biol. 2022;434:167381.
[141] Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia; 1981.
[142] Bouguet J-Y. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. Intel corporation. 2001;5:4.
[143] Turoňová B, Schur FKM, Wan W, Briggs JAG. Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4Å. Journal of Structural Biology. 2017;199:187-95.

[144] Singla J, White KL, Stevens RC, Alber F. Assessment of scoring functions to rank the quality of 3D subtomogram clusters from cryo-electron tomography. Journal of Structural Biology. 2021;213:107727.
[145] Zhang L, Ren G. High-resolution single-molecule structure revealed by electron microscopy and individual particle electron tomography. J Phys Chem Biophys. 2012;2:10.4172.
[146] Chen M, Bell JM, Shi X, Sun SY, Wang Z, Ludtke SJ. A complete data processing workflow for cryo-ET and subtomogram averaging. Nature Methods. 2019;16:1161-8.
[147] Himes BA, Zhang P. emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging. Nature Methods. 2018;15:955-61.
[148] Slabaugh GG. Computing Euler angles from a rotation matrix. Retrieved on August. 1999;6:39-63.
[149] Ma J. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. Structure. 2005;13:373-80.
[150] Frenkel D, Smit B. Understanding molecular simulation: from algorithms to applications: Elsevier; 2001.
[151] Kobayashi C, Jung J, Matsunaga Y, Mori T, Ando T, Tamura K, et al. GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. Wiley Online Library; 2017. p. 2193-206.
[152] Jiménez de la Morena J, Conesa P, Fonseca YC, de Isidro-Gómez FP, Herreros D, Fernández-Giménez E, et al. ScipionTomo: Towards cryo-electron tomography software integration, reproducibility, and validation. Journal of Structural Biology. 2022;214:107872.
[153] Strelak D, Jiménez-Moreno A, Vilas JL, Ramírez-Aportela E, Sánchez-García R, Maluenda D, et al. Advances in Xmipp for Cryo–Electron Microscopy: From Xmipp to Scipion. Molecules. 2021;26:6224.
[154] de la Rosa-Trevín JM, Quintana A, del Cano L, Zaldívar A, Foche I, Gutiérrez J, et al. Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. Journal of Structural Biology. 2016;195:93-9.
[155] Vuillemot R, Miyashita O, Tama F, Rouiller I, Jonic S. NMMD: Efficient Cryo-EM Flexible Fitting Based on Simultaneous Normal Mode and Molecular Dynamics atomic displacements. Journal of Molecular Biology. 2022;434:167483.
[156] Harastani M, Jonic S. Methods for analyzing continuous conformational variability of biomolecules in cryo electron subtomograms: HEMNMA-3D vs. traditional classification. bioRxiv. 2021;DOI: 10.1101/2021.10.14.464366.
[157] Kobayashi C, Jung J, Matsunaga Y, Mori T, Ando T, Tamura K, et al. GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. Journal of Computational Chemistry. 2017;38:2193-206.
[158] Suhre K, Sanejouand Y-H. ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Research. 2004;32:W610-W4.
[159] Frank J, Ourmazd A. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. Methods. 2016;100:61-7.
[160] Moscovich A, Halevi A, Andén J, Singer A. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. Inverse Problems. 2020;36:024003.
[161] Punjani A, Fleet DJ. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. Journal of Structural Biology. 2021;213:107702.
[162] Gupta H, Phan TH, Yoo J, Unser M. Multi-CryoGAN: Reconstruction of Continuous Conformations in Cryo-EM Using Generative Adversarial Networks. In: Bartoli A, Fusiello A, editors. Computer Vision – ECCV 2020 Workshops. Cham: Springer International Publishing; 2020. p. 429-44.
[163] Zhong ED, Bepler T, Berger B, Davis JH. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. Nature Methods. 2021;18:176-85.
[164] Zhong ED, Lerer A, Davis JH, Berger B. CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)2021. p. 4046-55.
[165] Chen M, Ludtke SJ. Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. Nature Methods. 2021;18:930-6.

[166] Rosenbaum D, Garnelo M, Zielinski M, Beattie C, Clancy E, Huber A, et al. Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs. arXiv preprint arXiv:210614108. 2021.
[167] Tegunov D, Cramer P. Real-time cryo-electron microscopy data preprocessing with Warp. Nat Methods. 2019;16:1146-52.

Table of Figures

Table of Tables

HARASTANI Mohamad – Thèse de doctorat – 2019-2022

Développement de méthodes d'analyse d'images pour les études in vitro et in situ de la variabilité conformationnelle des complexes biomoléculaires par cryo-tomographie électronique : cas d'études de la structure et de la dynamique des nucléosomes

Résumé :

La tomographie électronique cryogénique (cryo-ET) est une technique de biologie structurale qui permet de déterminer la structure et la dynamique des complexes biomoléculaires dans leur environnement cellulaire natif. Au cours de la dernière décennie, la cryo-ET a connu divers développements instrumentaux et logiciels qui ont permis d'obtenir des informations structurelles sans précédent sur les processus fondamentaux de la vie cellulaire. Néanmoins, plusieurs défis restent à relever avant que la cryo-ET puisse atteindre son plein potentiel. Les données 3D des biomolécules produites à l'aide de la cryo-ET sont bruyantes, peu contrastées, souffrent d'anisotropies spatiales et sont donc très difficiles à analyser individuellement. Par conséquent, les méthodes courantes de traitement des données cryo-ET visent à faire la moyenne de plusieurs copies de données 3D de biomolécules individuelles pour obtenir une structure moyenne à une résolution plus élevée. Cependant, les biomolécules sont des entités flexibles, et le calcul de la moyenne cache des informations sur leur variabilité conformationnelle, alors qu'une compréhension complète des mécanismes fonctionnels des biomolécules ne peut être obtenue que si leur variabilité conformationnelle est prise en compte. Entravées par les défis mentionnés ci-dessus, les techniques précédentes abordent la variabilité biomoléculaire par la classification, en discrétisant les transitions continues des biomolécules en un nombre fini d'états. Par conséquent, un défi critique du traitement des données est de développer des méthodes qui seront capables d'interpréter les données en termes de transitions conformationnelles continues. Dans cette thèse, je présente les deux premières méthodes de traitement des données cryo-ET qui traitent de la variabilité conformationnelle continue des biomolécules, HEMNMA-3D et TomoFlow. HEMNMA-3D analyse les données expérimentales avec des directions de mouvement simulées par l'analyse en mode normal, et permet la découverte d'une large gamme de mouvements biomoléculaires cachés dans les données. Cependant, HEMNMA-3D dépend de ce préalable (directions de mouvement simulées), ce qui le rend susceptible d'être mal interprété et biaisé en cas de mauvaise utilisation. TomoFlow extrait les mouvements des données sans information préalable à l'aide d'une technique de vision par ordinateur appelée "Optical Flow". Il est donc moins susceptible d'être mal interprété et mal utilisé. Cependant, lorsqu'il rencontre de grandes amplitudes de mouvement, il produit une version lisse et réduite du mouvement biomoléculaire réel. HEMNMA-3D et TomoFlow ont des modèles mathématiques différents, mais tous deux sont capables d'explorer les paysages conformationnels biomoléculaires et sont supérieurs à la classification. Je valide systématiquement HEMNMA-3D et TomoFlow sur des ensembles de données synthétiques. Je montre également le potentiel de ces deux méthodes sur des données cryo-ET expérimentales de la variabilité conformationnelle des nucléosomes in situ, dans le cadre d'une étude en cours sur la structure et la dynamique des nucléosomes dans les cellules. Les deux méthodes présentent des résultats cohérents, permettant de mieux comprendre la variabilité conformationnelle des nucléosomes, en accord avec les analyses visuelles et théoriques précédentes des conformations des nucléosomes. Je démontre que ces méthodes produisent des résultats valables avec des données in situ difficiles des nucléosomes. Sur cette base, elles devraient également être utiles pour les études conformationnelles d'autres complexes biomoléculaires in vitro et in situ. Les logiciels HEMNMA-3D et TomoFlow sont accessibles au public, en tant que partie du pipeline de traitement des données cryo-ET du progiciel open-source ContinuousFlex, qui est actuellement un plugin du logiciel Scipion, largement utilisé dans le domaine, et qui utilise le logiciel backend Xmipp de Scipion. Je contribue à la maintenance et au développement de ces trois logiciels, en particulier de ContinuousFlex.

Mots-clés : Cryo-ET, Analyse continue de la variabilité conformationnelle, HEMNMA-3D, TomoFlow, Structure et dynamique des nucléosomes dans les cellules, ContinuousFlex, Scipion et Xmipp.

Image analysis methods development for in vitro and in situ cryo-electron tomography studies of conformational variability of biomolecular complexes: Case of nucleosome structural and dynamics studies

Abstract:
Cryogenic electron tomography (cryo-ET) is a structural biology technique that allows determining structure and dynamics of biomolecular complexes in their native cellular environment. During the last decade, cryo-ET has witnessed various instrumental and software developments that allowed unprecedented structural insights into the fundamental processes of cellular life. Nevertheless, several challenges remain to be addressed before cryo-ET can reach its full potential. 3D data of biomolecules produced using cryo-ET are noisy, low in contrast, suffer from spacial anisotropies, and thus are very difficult to analyze individually. Hence, mainstream cryo-ET data processing methods aim to average multiple copies of 3D data of individual biomolecules to obtain an average structure at a higher resolution. However, biomolecules are flexible entities, and the averaging hides information about their conformational variability, whereas a complete understanding of functional mechanisms of biomolecules can only be achieved when their conformational variability is taken into account. Hindered by the challenges mentioned above, previous techniques address biomolecular variability with classification, discretizing continuous transitions of biomolecules into a finite number of states. Hence, a critical data processing challenge is to develop methods that will be able to interpret data in terms of continuous conformational transitions. In this thesis, I present the first two cryo-ET data processing methods that address continuous conformational variability of biomolecules, HEMNMA-3D and TomoFlow. HEMNMA-3D analyses experimental data with motion directions simulated by Normal Mode Analysis, and allows the discovery of a large range of biomolecular motions hidden in the data. However, HEMNMA-3D depends on this prior (simulated motion directions), making it prone to misinterpretation and bias when misused. TomoFlow extracts movements from the data without prior information using a computer vision technique called the Optical Flow. Therefore, it is less prone to misinterpretation and misuse. However, when it encounters large motion magnitudes, it results in a smooth and downscaled version of the actual biomolecular motion. HEMNMA-3D and TomoFlow have different mathematical models, but both are able to explore biomolecular conformational landscapes and are superior to classification. I systematically validate HEMNMA-3D and TomoFlow on synthetic datasets. Also, I show the potential of these two methods on experimental cryo-ET data of nucleosome conformational variability in situ, taking part in an ongoing study of nucleosome structure and dynamics in cells. The two methods show coherent results, shedding insight into the conformational variability of nucleosomes, in line with previous visual and theoretical analyses of nucleosome conformations. I demonstrate that these methods produce valuable results with challenging in situ data of nucleosomes. Based on this, they are also expected to be useful for conformational studies of other biomolecular complexes in vitro and in situ. The software of HEMNMA-3D and TomoFlow is publicly available, as part of the cryo-ET data processing pipeline of the open-source software package ContinuousFlex, which is currently a plugin of Scipion software, extensively used in the field, and uses Scipion's backend software Xmipp. I contribute to the maintenance and development of these three software packages, especially ContinuousFlex.

Keywords: Cryo-ET, Continuous Conformational Variability Analysis, HEMNMA-3D, TomoFlow, Nucleosome Structure and Dynamics in cells, ContinuousFlex, Scipion and Xmipp.