Molecular dynamics simulation based image analysis methods for structural studies of biological macromolecular complexesolecular complexes

Rémi Vuillemot

Doctoral Thesis

Molecular Dynamics Simulation-based Image Analysis Methods for Structural Studies of Biological Macromolecular Complexes

Rémi Vuillemot
ORCID 0000-0003-3714-9127

Submitted in total fulfilment for the jointly awarded degree of
Diplôme national de doctorat, Sorbonne Université
Doctor of Philosophy (Ph.D.), The University of Melbourne

Submitted on July, 28 2023

Directed by
Slavica Jonic (Sorbonne Université)
Isabelle Rouiller (The University of Melbourne)

Doctoral school of Computer Science, Communication and Electronics (ED130), IMPMC-UMR7590 CNRS, Sorbonne Université

Department of Biochemistry & Pharmacology, School of Biomedical Sciences, Faculty of Medicine, Dentistry and Sciences, Bio21 Molecular Science & Biotechnology Institute, The University of Melbourne

---

Presented and defended publicly on October 4th, 2023

In front of a jury composed of:

Yves Mechulam (DR1 CNRS, Ecole polytechnique)	Reporter
Gunnar Schroeder (A./Prof., Universität Düsseldorf)	Reporter
Irina Gutsche (DR1 CNRS, Université Grenoble-Alpes)	Examiner
Michael Parker (Prof., The University of Melbourne)	Examiner
David Perahia (DR2 Emeritus CNRS, ENS Paris-Saclay)	Examiner
Martin Weigt (Prof., Sorbonne Université)	Examiner
Slavica Jonic (DR2 CNRS, Sorbonne Université)	Thesis director
Isabelle Rouiller (A./Prof., The University of Melbourne)	Thesis director

Dedication

To my mother

Abstract

Cryo-Electron Microscopy (cryo-EM) is an imaging technique that allows observing nanoscale molecular machines such as proteins and DNA. These molecular machines are flexible and constantly undergo internal motions that change their conformations in order to achieve their biological function. Identifying conformational changes of macromolecules is of great biological importance and is used in the development of new drugs. Deciphering continuous conformational transitions of macromolecules, through the main cryo-EM processing techniques, Single Particle Analysis (SPA) and cryo-Electron Tomography (cryo-ET), is challenging partly due to the low signal-to-noise ratio and is currently an active field of research. During my PhD thesis, I investigated new image processing methods based on Molecular Dynamics (MD) simulations that allow extracting continuous conformational variability from SPA and cryo-ET data. My work resulted in the development of MDSPACE and MDTOMO, the two first methods using MD simulations, empowered by Normal Mode Analysis (NMA), to extract the continuous conformational landscape from SPA and cryo-ET datasets, respectively. The developed methods were employed to investigate the conformational behavior of diverse systems such as 80S ribosome and AAA ATPase p97 (MDSPACE) and SARS-CoV2 spike protein in situ (MDTOMO). The methods are competitive as they allow an atomic-scale determination of the conformational spaces of various macromolecules from heterogeneous cryo-EM data, which can be advantageous in structure-based drug development.

Keywords: Cryo-electron microscopy, cryo-electron tomography, conformational heterogeneity, molecular dynamics, normal mode analysis, molecular modeling, AAA ATPase p97, 80S ribosome, SARS-CoV-2 spike protein

Résumé

La cryo-microscopie électronique (cryo-ME) est une technique d'imagerie permettant d'observer à l'échelle nanoscopique et dans leur état naturel des machines moléculaires tel les protéines et l'ADN. Ces machines moléculaires sont flexibles et subissent constamment des déformations pour réaliser leur fonction biologique. Par conséquent, l'identification de leur flexibilité est importante au point de vue biologique, et est utilisé par exemple dans le développement de nouveaux médicaments. Cependant, prédire les transitions conformationnelles continues des macromolécules en utilisant les principales techniques de traitement d'images en cryo-ME, à savoir l'analyse en particule isolé (API) et la cryo-tomographie électronique (cryo-TE), est difficile en partie en raison d'un rapport signal sur bruit très bas, ce qui en fait un sujet de recherche activement étudié ces dernières années. Pendant ma thèse, j'ai travaillé sur de nouvelles méthodes de traitement d'images, utilisant la simulation de la dynamique moléculaire, permettant d'extraire la variabilité conformationnelle continue des données d'API et de cryo-TE. Mes travaux ont abouti au développement de MDSAPCE et MDTOMO, les deux premières méthodes basées sur la simulation de la dynamique moléculaire, accélérées par une analyse de mode normaux, permettant d'extraire des paysages conformationnels continues des données d'API et de cryo-TE, respectivement. Les méthodes développées ont permis d'analyser le comportement conformationnel de systèmes divers comme le ribosome 80S, l'AAA ATPase p97 (MDSPACE) ou encore la protéine spiculaire de SARS-CoV2 (MDTOMO). L'avantage de ces nouvelles méthodes est de prédire les états conformationnels d'un jeu de donné de cryo-ME à l'échelle atomique, qui peut être un atout pour la conception de médicament à partir de structures.

Mots clés : Cryo-microscopie électronique, cryo-tomographie électronique, hétérogénéité conformationnelle, dynamique moléculaire, analyse de mode normaux, modélisation moléculaire, AAA ATPase p97, ribosome 80S, protéine spiculaire SARS-CoV-2

Declaration

This is to certify that:

1. This thesis comprises only my original work towards the jointly awarded degree National Doctoral degree at Sorbonne Université and Doctor of Philosophy (PhD) at the University of Melbourne except where indicated in the preface;
2. Due acknowledgment has been made in the text to all other material used; and
3. The thesis is shorter than 100,000 words in length, exclusive of tables, maps, and bibliographies.

Rémi Vuillemot

Preface

My PhD thesis is an interdisciplinary work, at the frontier of structural biology and computer science. It focuses on the development of new computational methods for the analysis of conformational heterogeneity in cryo-electron microscopy (cryo-EM). Funded by the CNRS, this thesis is a cotutelle resulting from the collaboration between Dr. Slavica Jonic at Sorbonne University (France) and Assoc. Prof. Isabelle Rouiller at the University of Melbourne (Australia), within the Melbourne-CNRS Network (MCN) and the ANR project EMBioMolMov. Therefore, it also involves a collaboration with the partners on this ANR project, namely the Institute of Genetics, Molecular and Cellular Biology – IGBMC Strasbourg (France) and the University of Nagoya and RIKEN Kobe (Japan).

Before starting this thesis, I obtained, in 2019, a master's degree in computer science at the University of Rennes 1. Then, I worked for one year as a research engineer in the industry in the development of computational methods for inverse problems applied to EEG brain signals.

I started my PhD thesis in 2020 at the IMPMC (Sorbonne University) and was introduced to the field of cryo-EM by Dr. Slavica Jonic. In my first year at IMPMC, under the supervision of Dr. Slavica Jonic, I developed a novel algorithm, NMMD, for cryo-EM structural modeling. The second year of the thesis was performed at the University of Melbourne (my mobility was postponed to spring 2022 due to the sanitary situation related to COVID19 pandemic). During my 1-year stay at the University of Melbourne, I improved my knowledge of cryo-EM through interaction with Assoc. Prof. Isabelle Rouiller and her team at the University of Melbourne and through hands-on experience with cryo-EM sample preparation and data collection. Also, during my stay at the University of Melbourne, I continued my work on image processing methods development for analyzing continuous conformational heterogeneity in cryo-EM datasets, which led to two main contributions of this thesis, namely the development of MDSPACE (for SPA data analysis) and MDTOMO (for cryo-ET data analysis) methods. Back in the IMPMC in the spring of 2023, I focused on further testing the methods, preparing two additional journal article manuscripts, and preparing this PhD thesis manuscript.

This thesis work resulted in three published articles, adapted for this thesis manuscript. Chapter III is adapted from an article published by the *Journal of Molecular Biology* on April, 15 2022, Chapter IV is adapted from an article published by the *Journal of Molecular Biology* on May, 1 2023, and Chapter V is adapted from an article published by the *Scientific Reports* on June, 30 2023. In these published articles, I contributed in more than 50% of the work, prepared the first draft of the manuscript, and prepared all the figures.

During the thesis work, I collaborated with colleagues from my French and Australian labs and with the collaborators involved in the ANR EMBioMolMov project. In Chapter III, the method development was done in collaboration with Florence Tama (Nagoya University, Japan) and Osamu Miyashita (RIKEN, Japan), who implemented the first version of the flexible fitting method based on MD simulation in the software package GENESIS, which I modified to speed up the fitting by including normal modes in the simulation.

In Chapter IV, the method development was done in collaboration with Alex Mirzaei (Sorbonne University). Alex developed the first version of the code to analyze cryo-EM particle images with MD simulations, in collaboration with Florence Tama (Nagoya University, Japan) and Osamu Miyashita (RIKEN, Japan), which I improved in several ways (e.g., by allowing the analysis of large datasets and adding normal modes to speed up this analysis). The cryo-EM dataset of p97 used in Chapter IV was collected by Sepideh Valimehr (University of Melbourne) and the preliminary data analysis was done by Sepideh Valimehr and Mohsen Kazemi (University of Melbourne). The interpretation and validation of the results obtained with the data of yeast 80S ribosome was done in collaboration with Bruno Klaholz and Léo Frechin (IGBMC, France).

In Chapter V, the subtomograms and the initial rigid-body alignment parameters used in the experiment with SARS-CoV-2 spikes were provided by Beata Turoňová (Max Planck Institute of Biophysics, Frankfurt am Main, Germany), who also participated in the interpretation and validation of the results obtained in this experiment.

This thesis work will result in at least two more journal article manuscripts, which are currently at a very advanced preparation stage. One focuses on the analysis of the conformational flexibility of p97 and involves the results of MDSPACE presented in IV.3.3. The other focuses on the analysis of the conformational flexibility of HER2 complexes and shows the results of MDSPACE in this context (collaboration with Stéphane Bressanelli and Rémi Ruédas, I2BC, Université Paris-Saclay).

Acknowledgements

I would like to thank all:

My supervisors, Slavica Jonic and Isabelle Rouiller for the countless help, guidance, insights, and patience they provided throughout my thesis. I always felt supported in my work and encouraged to express my ideas, for that I am deeply grateful.

My collaborators for their contributions to my thesis, Florence Tama (Nagoya University, Japan), Osamu Miyashita (RIKEN, Japan), Bruno Klaholz and Léo Fréchin (IGBMC, Strasbourg), Stéphane Bressanelli and Rémi Ruédas (I2BC, Paris-Saclay), and Beata Turoňová (Max Planck Institute of Biophysics, Frankfurt am Main, Germany).

The CNRS for funding this thesis and the Melbourne-CNRS Network (MCN) for the cotutelle between the Sorbonne University and the University of Melbourne. I also acknowledge the support of the ANR project EMBioMolMov.

My team members at the IMPMC, Mohamad Harastani, Alex Mirzaei, Ilyes Hamitouche, and Nima Barati for their great scientific collaboration, their support, and friendship.

The Ian Holmes Imaging Center for giving me the opportunity to get hands with the microscopes, especially Sepideh Valimehr for her help and patience.

The IDRIS and CINES computing centers for the HPC resources that were used during this thesis.

My lab members, past and present, both at Sorbonne University and at the University of Melbourne, for making this thesis a fun journey.

My family and friends for the great support provided throughout this thesis.

Finally, this work would not be possible without the endless support of my partner Barbara Maunaye. I am so grateful for her unconditional encouragement, love, and care.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Cα – Carbon alpha

CC – Correlation coefficient

Cryo-EM – Cryo-electron microscopy

Cryo-ET – Cryo-electron tomography

CTF – Contrast transfer function

CV – Collective variable

EMDB – Electron microscopy data bank

ENM – Elastic network model

HMC – Hamiltonian monte-Carlo

MD – Molecular dynamics

MC – Monte-Carlo

MCMC – Markov chain monte-Carlo

NMA – Normal mode analysis

PCA – Principal component analysis

PDB – Protein data bank

REUS – Replica exchange umbrella sampling

RTB – Rotation-translation block

SPA – Single particle analysis

STA – Subtomogram averaging

UMAP – Uniform manifold approximation and projection

Introduction

Biological molecules and macromolecular complexes composed of proteins, DNA, and RNA, are the building blocks of life, and responsible for most biological processes in living organisms. The three-dimensional (3D) shape of their folded chain of amino acids (or nucleic acids), is closely associated with their function and their working mechanism. Most importantly, molecular structures are dynamic objects that are constantly changing their shape (i.e. conformation) in reaction to their interaction with the continuously evolving cellular environment. Conformational changes in macromolecules engender mechanical motions that are responsible for a large number of biological functions in cells including DNA replication and reparation, cell signaling, and molecule transportation. Visualizing the 3D structure and dynamics of these molecular machineries in their native cellular environment is essential for understanding their molecular mechanism and for structure-based drug discovery [1].

Cryogenic electron microscopy (cryo-EM) has recently emerged as a powerful technique for determining the structure of biological macromolecules [2] and is particularly well-suited for studying their conformational variability. In a cryo-EM, the sample is observed in a frozen-hydrated state at cryogenic temperature (-180 °C) obtained by ultra-rapid freezing (vitrification) which preserves the sample in a close-to-native state. Particularly, the fast freezing allows capturing the different conformations of the molecules existing in solution. On the contrary, X-ray crystallography, the first historically established structural biology technique, requires sample crystallization. This technique, by nature, limits protein conformations as a protein crystal is an ordered array of identical proteins. with a limited number of conformations.

In the main cryo-EM technique, single particle analysis (SPA), the vitrified sample is composed of the studied protein isolated by protein purification. Thus, the sample contains multiple copies of the protein embedded in a thin layer of ice (so-called particles). These particles can be seen as "snapshots" of the proteins at the time of the vitrification. At this time, according to the *in vitro* biochemical conditions, the proteins can be in a conformational equilibrium, that is, adopt different conformations. In other words, vitrification captures the distribution of conformation of the protein at the time of the freezing.

Then, the vitrified sample is exposed to an electron beam that interacts with the proteins in the sample to produce an image. However, an intrinsic consequence of using high-energy particles such as electrons to observe fragile biological samples is radiation damage. To limit damages and avoid the complete destruction of the sample during imaging, the sample is exposed to a very low electron dose (below 20 e^- per Å^2) resulting in extremely noisy images. In such noisy images,

direct observation of the structural information is difficult, thus, requires sophisticated image processing algorithms to retrieve useful information and reconstruct a 3D structure.

Since the recent cryo-EM resolution revolution [2], cryo-EM reconstructions often achieve near-atomic resolutions (higher than 3.5 Å) [3-7]. At this resolution, it is possible to derive the atomic positions from the data and build an atomic structure *de novo*. However, to achieve such resolutions, the reconstruction methods rely on averaging a massive number of particles (between 10^4 and 10^6 particles). Because of the random nature of noise, averaging multiple noisy images cancels the noise and increases the signal-to-noise ratio (SNR). However, averaging particles that are conformationally heterogeneous has several drawbacks. First, averaging out heterogeneous conformations induces a loss of signal in the flexible regions of the molecule, decreasing the resolution of the reconstructions in these regions [8-11]. Second, and most importantly, averaging loses the information on the conformational heterogeneity, as such, the distribution of conformation contained in the sample is discarded to recover a single conformational state.

The problem of conformational heterogeneity in the data is traditionally solved by classification algorithms [12-17]. These algorithms group conformationally homogeneous particle images into a finite number of discrete classes. However, the classification assumes that the heterogeneity results from a small number of well-separated conformational populations, which may not be a true representation of the conformational heterogeneity present in the sample. Most often, biomolecules adopt a continuous distribution of conformations with multiple metastable states. In that case, classification discards most of the particles with continuous heterogeneity and keeps only the particles associated with a few discrete states with the largest conformational population. Consequently, the recovered conformational distribution is largely incomplete. In the recent years, characterizing the conformational heterogeneity in cryo-EM datasets as a continuous distribution has aroused great interest [18-30]. This thesis work aims at developing new methods that recover a continuous distribution of states from heterogeneous cryo-EM datasets. However, the task of disentangling the continuity of conformations in the data is a great computational challenge due to the low SNR of the collected images, which is typically between 0.1 and 0.01, signifying that the signal in the particle could be a hundred times smaller than the noise.

In some cases, preliminary knowledge of the structure of the studied protein is available. This can be an atomic model derived from experimental data (e.g., cryo-EM or X-ray crystallography data) or predicted from the amino-acid sequence of the complex (e.g., by AlphaFold2 [31]). This structural information can be exploited for analyzing the conformational heterogeneity, which can facilitate the conformational identification that is hindered by the high level of noise of the particles. Predictions of the structural motions of molecules have been studied by biophysicists for

many years with computational tools such as molecular dynamics (MD) simulations [32]. MD simulation predicts the physical movements of atoms of a given molecular structure by calculating inter-atomic forces using physically-based potentials. The conformational motion simulated by MD provides important information for studying the dynamics of objects that would be difficult to observe experimentally [33]. Moreover, *in silico* methods like MD simulation are useful when combined with experimental data (referred to as hybrid methods). For instance, hybrid methods were used to interpret cryo-EM maps in terms of atomic coordinates by flexibly fitting an existing atomic model into a cryo-EM map of another conformation [34-37].

This PhD thesis aims at investigating and developing new hybrid methods using MD simulations to characterize the continuity of conformational states in cryo-EM datasets. This work led to the development of two methods, MDSPACE and MDTOMO, that analyze large and highly heterogeneous datasets of SPA and cryo-electron tomography (cryo-ET), respectively. While SPA allows studying purified molecules, cryo-electron tomography (cryo-ET) allows studying the molecules in their cellular environment (*in situ*), for instance, by imaging vitrified cell sections or lamella.

Chapter I provides general background on cryo-EM, presents the standard cryo-EM image processing workflow, and reviews the state-of-the-art approaches to characterize the conformational heterogeneity.

Chapter II reviews the methods to simulate the mechanical motion of macromolecules, including MD simulations, and their application in hybrid methods for cryo-EM data analysis.

Chapter III presents an algorithm for efficient flexible fitting of an atomic model into an EM map by combining MD simulations and normal mode analysis (NMA). NMA is an approach for studying the intrinsic dynamics of a molecular structure around its equilibrium state. NMA is particularly useful to characterize the flexibility of macromolecules at a low computational cost, in contrast to MD simulations which are computationally expensive. However, the motions in NMA are typically constrained to a small subset of highly collective motions and often induce distortions in the structure. The proposed method that combines NMA and MD simulation, named Normal Mode Molecular Dynamics (NMMD) was tested on synthetic and experimental EM maps and revealed to be advantageous in terms of computational efficiency compared to standard MD simulations, while not suffering from the NMA limitations. The NMMD method allowed opening the analysis of conformational heterogeneity of large cryo-EM datasets to MD-based approaches as MD simulation alone was computationally prohibitive. NMMD has been the object of an article publication in the *Journal of Molecular Biology* in 2022, where I was the first author [38]. The

material presented in Chapter III was extracted from the published manuscript [38] and adapted for this PhD thesis manuscript.

Chapter IV describes MDSPACE (Molecular Dynamics simulation for Single Particle Analysis of Continuous Conformational hEterogeneity), the first method to analyze continuous conformational heterogeneity of cryo-EM single particle datasets based on an MD-based fitting approach (i.e. NMMD). In this method, NMMD is used to estimate the conformation in each individual cryo-EM particle image by flexibly fitting the image with a given atomic model. The analysis results in one fitted atomic model per particle image that can be mapped onto a low-dimensional space for further analyses (3D reconstructions and motion animations). MDSPACE has been successfully applied to diverse systems, including experimental datasets of yeast 80S ribosome and AAA ATPase p97, and is competitive with other methods addressing the conformational heterogeneity problem in cryo-EM. MDSPACE has been the object of an article publication in the *Journal of Molecular Biology* in 2023, where I was the first author [25]. The material presented in Chapter IV was extracted from the published manuscript [25] and adapted for this PhD thesis manuscript.

Chapter V describes MDTOMO, a method that extends the idea behind MDSPACE to the analysis of cryo-ET data. Compared to SPA that studies purified molecules, cryo-ET allows studying samples in their cellular environment (*in situ*), for instance, by imaging vitrified cell sections or lamella. To alleviate the smaller amount of molecule of interest of the *in situ* samples, a tomographic acquisition is performed to collect images of the samples from multiple directions allowing to directly reconstruct a 3D volume. However, the SNR of the collected tilt images of cell sections is much lower than the SNR of the cryo-EM single particle images. This is due to a lower electron dose used in cryo-ET to limit the radiation damage during the multiple expositions of the same zone of the sample to the electron beam. Despite the many challenges in cryo-ET, conformational studies of *in situ* samples are highly biologically relevant as they inform on the conformational behavior of macromolecules directly in their native environment and MDTOMO allows such conformational studies. MDTOMO was tested on synthetic and experimental datasets, including a dataset of subtomograms of SARS-CoV2 spike proteins *in situ,* revealing several continuous conformational transitions that were not observed with standard approaches. MDTOMO has been the object of an article publication in the *Scientific Reports* journal in 2023, where I was the first author [26]. The material presented in Chapter V was extracted from the published manuscript [26] and adapted for this PhD thesis manuscript.

Chapter VI provides a general discussion and conclusion with perspectives for future work.

Chapter I.Background on cryo-electron microscopy

In this chapter, I start by providing a background and historical perspective on cryo-EM. Then, I present the basic principles of SPA and cryo-ET imaging techniques and show how the modern SPA and cryo-ET image processing methods achieve the reconstruction of a 3D structure. Finally, I review the methods that analyze the conformational heterogeneity in cryo-EM data.

I.1. The development of cryo-EM

The observation of biological samples at the nanoscopic scale is particularly challenging and there are only a few experimental techniques that allow such observation, including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy.

Electrons, like other sources of radiation such as light and X-rays, have a wave-like behavior that can be exploited for imaging purposes. Compared to visible light, electrons and X-rays have a much smaller wavelength, which in theory allows imaging objects below atomic resolution. An intense X-ray beam can be used to illuminate a crystal and the resulting diffraction pattern allows identifying the position of atoms in the crystallized sample. X-ray crystallography was the first technique that allowed the determination of the structure of biomolecules [39]. X-ray crystallography is still the technique that results in the determination of the largest number of biomolecular structures per year. However, the crystalline state of the sample required for the diffraction of X-rays is the major limitation of X-ray crystallography. Crystallization involves chemical harsh buffers that may stabilize the macromolecules in low-populated states, not always biologically relevant. Furthermore, many biological macromolecules and usually highly flexible macromolecules cannot form crystals of good quality.

On the contrary, sample preparation for cryo-EM allows for better conservation of its native state as the sample is maintained in a biologically compatible solute. In the 1980s, J. Dubochet and coworkers proposed a sample preparation method [40] where a biological sample is deposited on a mesh grid and plunge-frozen in liquid ethane, around -180 °C, which freezes the sample rapidly enough to prevent the formation of ice crystals (see Figure 1a). In this particular form of ice, called amorphous or vitreous ice, the water molecules form non-ordered matrices that create minimal interference when imaging the sample with a beam of electrons, while still preventing evaporation of water in the vacuum of the microscope. Electrons, compared to X-rays, can be easily focused with electromagnetic lenses allowing applications in microscopy. However, electrons have a much smaller penetration capability into objects compared to X-rays. Therefore, to detect a signal, the studied sample needs to be extremely thin (typically 10 to 100 nm [41]). Moreover, the column of

the electron microscope needs to be in a high vacuum condition to avoid the interaction of electrons with air molecules. The cryo-EM sample preparation tackles all these problems at once: the solution containing the molecule of interest is placed on holes of a mesh grid which, when frozen at cryogenic temperature, forms a thin suspended layer of ice (see Figure 1a); and the low temperature prevents the sample from drying in the microscope. This technique was revolutionary and allowed the authors to observe vitrified viruses in native conditions [40].

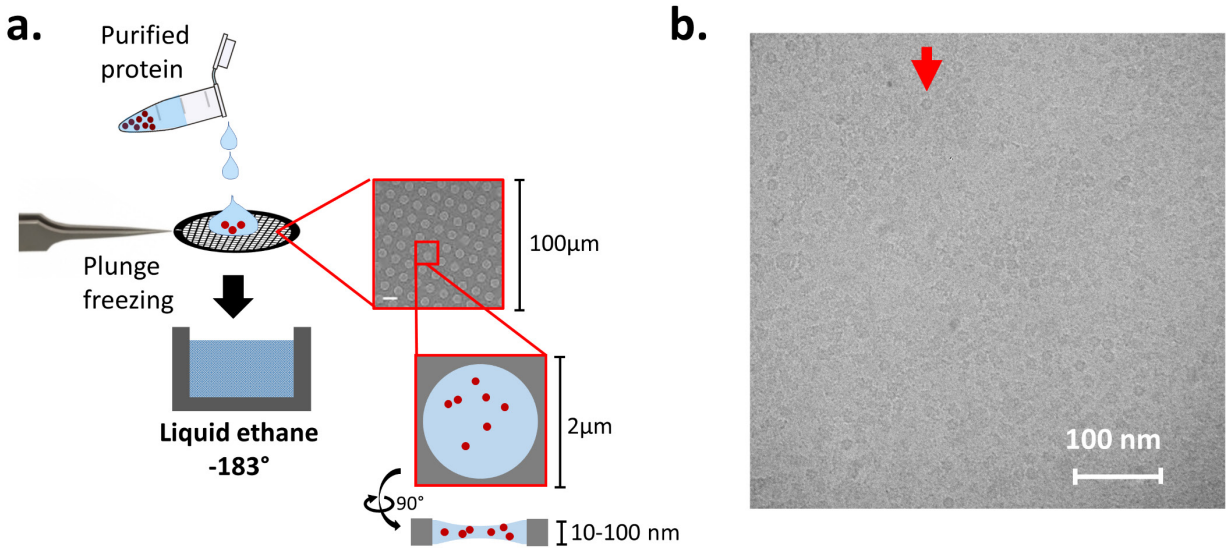

Figure 1: Cryo-EM single particle sample preparation. (a) cryo-EM sample preparation schematic (b) Micrograph (raw image) of AAA ATPase p97 that I collected with a 120kV microscope at the University of Melbourne. The red arrows show a particle of a top view of p97.

However, the biological samples are fragile and the interactions of electrons with the sample are extremely damaging. Thus, there is a tradeoff between the electron dose applied for collecting the data and the quality of structural elements that are imaged. Because the number of electrons acceptable to visualize high-resolution structural information of a biological sample is extremely limited (below 20 e^- per Å^2), the raw images have a low signal-to-noise ratio (SNR), between 0.1 and 0.01 as shown in Figure 1b. Such low SNR does not allow observing small structural features in raw images by eye making direct interpretation difficult. The pioneering work of J. Frank and coworkers in the development of mathematical tools for image analysis [42, 43] led to the basis of SPA (see section I.3). This approach relies on averaging a large number of images of the identical molecule into a reconstructed 3D density volume (known as "EM map") to drastically increase the SNR. At that time, the work of R. Henderson and coworkers on electron crystallography led, in 1990, to the first high-resolution structure (3.5 Å resolution) of bacteriorhodopsin from 2D crystals, demonstrating that electrons can be used to observe biomolecules at near-atomic resolution [44].

Nevertheless, for many years, determining high-resolution structures in the absence of crystals remained challenging due to the low SNR of the data and the movement of the particles under the electron beam preventing accurate alignment of the images. For these reasons, the resolution of most 3D reconstructions was generally too low to allow *de novo* structure modeling (worse than 5 Å resolution). About ten years ago, there was a notable hardware improvement with the apparition of a new generation of cameras known as direct electron detectors (DED). DEDs provided not only significant improvements in electron detection efficiency but also a much faster image acquisition than previous digital cameras [45]. The higher frame rate of DEDs allowed to record dose-fractioned series of images instead a single image, allowing to correct for beam-induced motions that greatly improved the resolution of the images [46]. The combination of hardware improvements with DEDs and notable software improvements [12, 47], lead to what is known as the "resolution revolution" [2] of single particle cryo-EM, where the structures of particularly challenging proteins were solved at near-atomic resolution (between 3 and 4 Å resolution) [48].

Since then, the number of structures solved by cryo-EM has followed an exponential growth, as reported by the number of atomic models derived from EM maps and deposited in the Protein Data Bank (PDB) over the years (Figure 2a), and a continuing increase in the resolution of the EM maps deposited in Electron Microscopy Data Bank (EMDB) (Figure 2b). Nowadays, cryo-EM is recognized as one of the most powerful imaging techniques for structure determination and has largely been adopted in the structural biology field. Nobel Prize in Chemistry was awarded to J. Dubochet, R. Henderson, and J. Frank in 2017.

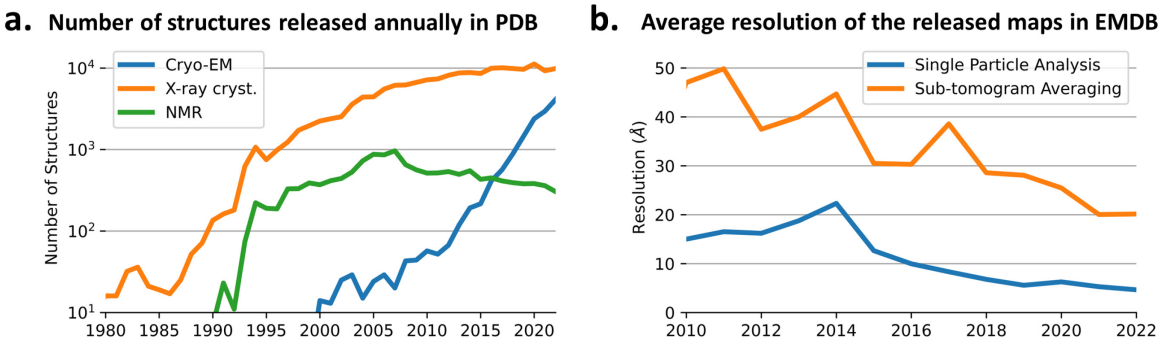

Figure 2 : a) The evolution of the number of structures deposited in the Protein Data Bank (PDB) each year using cryo-EM, X-ray crystallography, and nuclear magnetic resonance (NMR) spectroscopy. b) The evolution of the average resolution of the EM maps obtained by SPA and cryo-ET and deposited in the Electron Microscopy Data Bank (EMDB). Statistics from www.rcsb.org, accessed on May 1, 2023.

Despite the number of improvements in the SPA image processing methods, exploring the conformational heterogeneity in the single particle data remains one of the main challenges in cryo-EM. The reasons for that are the low SNR of the images and the difficulty to disentangle the

conformational heterogeneity from the rotational and translational heterogeneity. Currently, most used methods based on classification [12, 13] are driven by the aim of achieving the highest resolution of at least one cryo-EM map reconstruction from the given dataset, but they are ill-suited for the analysis of continuous conformational variability. Indeed, they most often discard all the particles that are not homogeneous with the most dominant conformational states (i.e. the intermediate states of continuous conformational transitions) resulting in a largely incomplete definition of the conformational distribution in the data. The analysis of continuous conformational variability poses a great computational problem, which is still open. The methods to address it are under active development (see section I.5.2) and is the major aim of this PhD thesis.

I.2. Image formation & CTF

In the microscope, the electron beam interacts with the matter present in the sample (the solvent and the biological matter). When an electron passes close to an atom of the sample, its phase is slightly delayed in the order of the milliradian. The scattered electron wave that is transmitted is recombined on the image plane and the phase-shifts of the scattered electrons produce interferences that can be measured. As the observed biological specimens are relatively ordered structures compared to the surrounding amorphous ice forming non-ordered matrices, the resulting phase-shift generates constructive interferences on the image plane. These phase-shifts are almost entirely responsible for the observed contrast in cryo-EM.

However, when the electron waves are recombined on the image plane in focus, most of the phase contrast disappears, resulting in almost no contrast at all. To accentuate the phase contrast, the images are acquired with a small defocus (typically between -0.5 and -2.0 μm). The defocus together with the optical properties of the microscope can be mathematically translated in a contrast transfer function (CTF) that modulates the acquired image. The CTF produces amplitude variations in the reciprocal space that take the form of a series of rings (known as Thon rings) as shown in the synthetic example in Figure 3. Such amplitude modulation in the reciprocal space means that some spatial frequencies are damped and some are totally missing in the image, therefore, producing signal distortions in the real space (Figure 3). To avoid such distortions in the reconstructions, an important step of the image processing workflow is the precise estimation and correction of the CTF [49-53]. In addition, to compensate for the loss of some spatial frequencies, the images are typically acquired with different defocus values. Each defocus being associated with the damping of specific spatial frequencies (as it changes the frequency of the Thon rings, see Figure 3), acquiring images with different defocus allows for the recover of all the spatial frequencies.

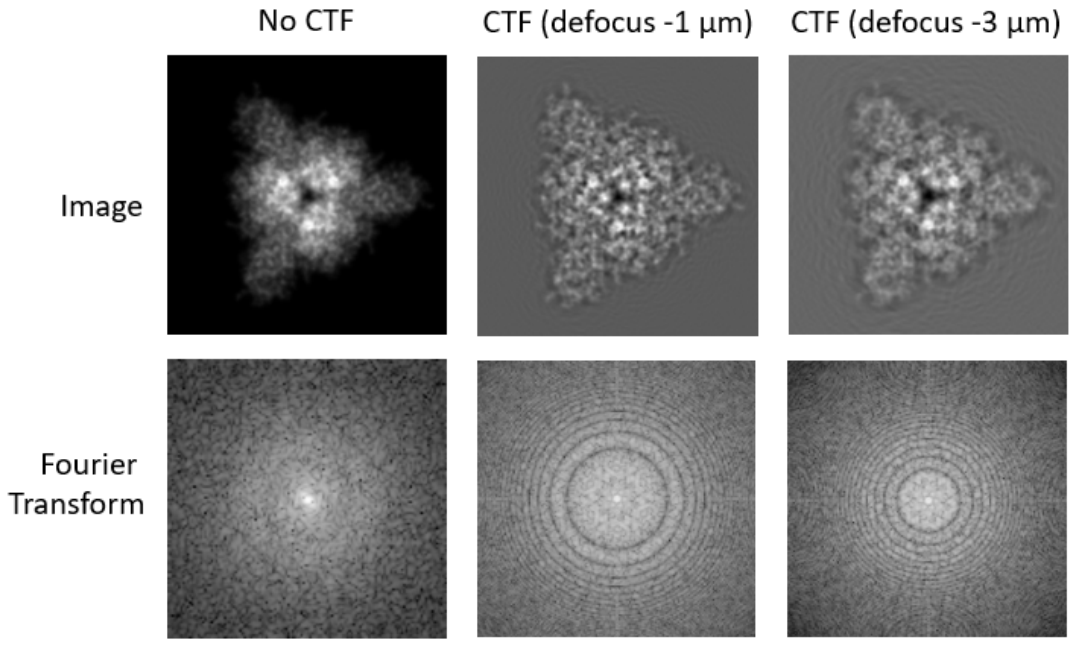

Figure 3: Effect of the contrast transfer function on a synthetic projection image of SARS-CoV2 spike without noise. The images were generated using Xmipp [54] by converting a model of SARS-CoV-2 spike (PDB 6VXX) into a density volume using a method based on scattering factors and projected into 2D images by simulating an electron microscope CTF.

I.3. Single Particle Analysis

The principle of SPA is to collect images of many identical purified macromolecules at different orientations in order to average them in 3D space and, in this way, increase the SNR and the resolution of the 3D reconstruction. Many image processing methods have been developed for SPA over the years.

The acquired image (called micrograph) contains projections of a large number of copies of the macromolecule at different orientations in 3D space. The SPA image processing methods rely on aligning and averaging all these copies to calculate a 3D reconstruction (the so-called 3D density map or EM map). The different image processing steps have been implemented in the principal software suites including Relion, Scipion, EMAN, and CryoSPARC [12, 13, 55, 56]. The basic principles of the SPA workflows are presented in Figure 4. In the following, I give an overview of the different steps that are involved in the computational SPA pipelines.

First, the collected micrographs are corrected for beam-induced motions during the acquisition [46, 57, 58], which increases significantly the achievable resolution of the 3D reconstruction. At the same stage, the CTF parameters are estimated from the micrographs [50-52].

Then, the sub-images corresponding to the particle of interest are identified and extracted from the micrographs (this procedure is known as particle picking). This can be done either manually (e.g., in the case of very challenging small complexes) or semi-automatically, by template-based approaches [59, 60] or deep-learning approaches [61, 62].

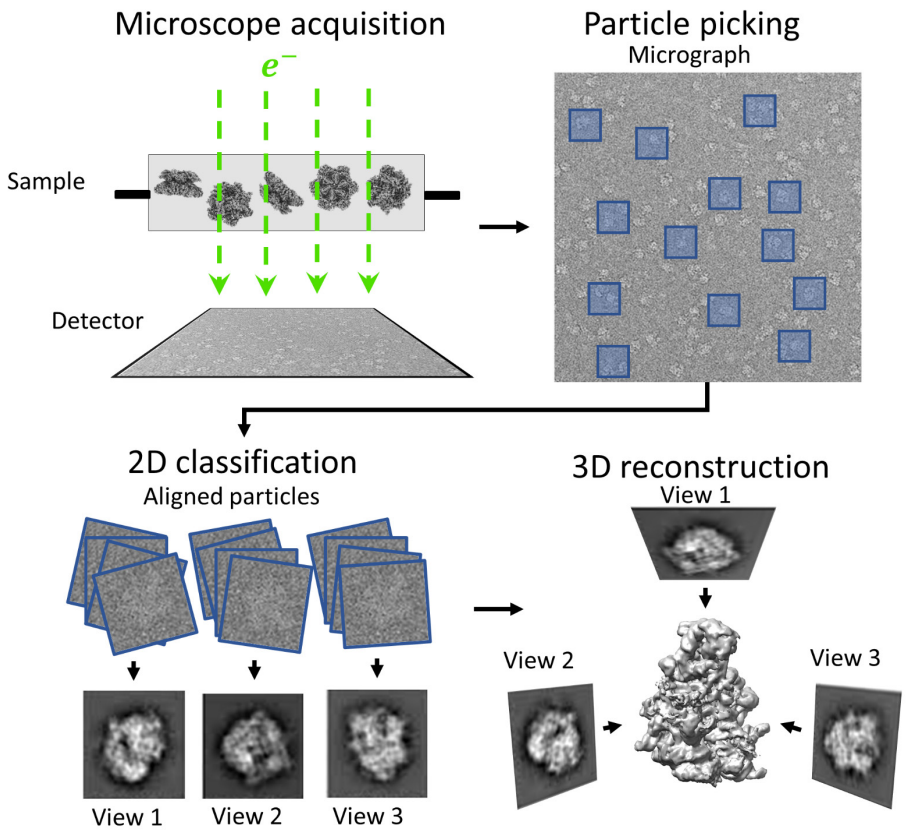

Figure 4: Basic principles of single particle analysis (SPA) The sample composed of purified proteins is imaged in the electron microscope. Then, particle picking is performed to extract images of individual proteins from the recorded micrograph. The 2D particles are classified into groups of similar images mainly of particles of the same orientation of projection with respect to in-plane rotations, averaged together allowing to increase the SNR and assess the quality of the sample. Finally, a 3D structure is reconstructed from the particles by estimating the alignment parameter of each particle toward a 3D reference and performing back-projection to obtain a 3D density map.

Once the particles are extracted, they are classified into groups of similar images of particles of the same orientation of projection with respect to in-plane rotations [12, 63]. This step, referred to as 2D classification, allows to check the sample homogeneity in terms of the distribution of sizes

and shapes of the particles, discard some of the particles that were wrongly picked, and show if some projection views are overrepresented or missing. Based on the results of 2D classification, the practitioners may decide to continue to analyze the data or to return to the lab to optimize the sample for cryo-EM, especially if their aim is a high-resolution 3D reconstruction of a single structure.

The next step is reconstructing a 3D EM map by estimating the rotational and translational projection parameters of each particle (i.e. particle pose). Reconstruction algorithms make use of the *central slice theorem* which associates the Fourier transform of the 2D projection of a given volume to a 2D slice that crosses the origin of the volume in the Fourier domain. The particle pose is generally estimated using maximum likelihood algorithms [12-14]. An important part of the reconstruction in the disambiguation of the conformational heterogeneity in the particles from the rotational and translational heterogeneity, traditionally done using classification algorithms, and will be detailed in the final section of this chapter (section I.5).

Once an initial volume or a set of initial volumes is obtained, their resolution is improved by refining the particle pose parameters iteratively and by correcting the CTF as well as by adding more particle images into the analysis. The refined volumes can be enhanced at the post-processing stage using volume sharpening algorithms based on structural knowledge [64, 65].

Finally, the final maps can be interpreted in terms of atomic positions by structural modeling. If the resolution is sufficiently high (3-4 Å), the structure can be modeled *de novo* – meaning that the atomic positions are derived based on the EM map without a reference model – using semi-automated tools [66, 67]. Otherwise, if a related atomic model is available, structural modeling can be performed using hybrid methods [34-36, 38, 68], which will be extensively studied in section II.4.

I.4. Cryo-Electron Tomography

Observing biological samples *in situ* is particularly important to study the native dynamics of biomolecules and their interactions within their cellular environment. Compared to purified molecules, the number of molecules of interest that can be extracted from *in situ* data is much smaller (and, therefore, the number of particle views as well). This is due to the dense cellular environment which is crowded with different proteins making the recognition task in the particle picking particularly difficult. Cryo-electron tomography (cryo-ET) [69] is a technique by which a series of 2D images of the sample are recorded by titling the sample holder in the microscope, usually around a single tilt axis. The tomographic acquisition increases the number of views of the complex of interest and the image tilt series is relatively easily assembled into a 3D volume (the

so-called tomogram) using back projection. During the image collection over multiple tilts, the sample deteriorates as it accumulates radiation damage. The dose accumulation is typically optimized by collecting images following a dose-symmetric tilt scheme by starting at 0° and increasing the tilt angle (with an increment of 1° or 2°) alternatively between the positive and negative tilt directions [70, 71]. However, the electron dose per image is still much lower than in SPA to spread the accumulated dose over the whole tilt range. This results in a very low SNR of the tomograms. On top of that, the maximum tilt angle is usually limited to +/-60°, due to the huge dose accumulation at larger tilt angles and the physical limitations of the sample holder, which results in anisotropic resolution of the 3D reconstructed volume (the so-called missing-wedge problem).

To obtain a high-resolution structure of a molecule of interest from cryo-ET data, a similar approach to SPA must be employed, where the molecule of interest is extracted from the tomogram into sub-volumes (called subtomograms), which are then aligned and averaged using an iterative scheme referred to as subtomogram averaging (StA). Figure 5 shows a typical cryo-ET image processing workflow. In the following, I describe the difference in the sample preparation and image processing methods between cryo-ET and SPA.

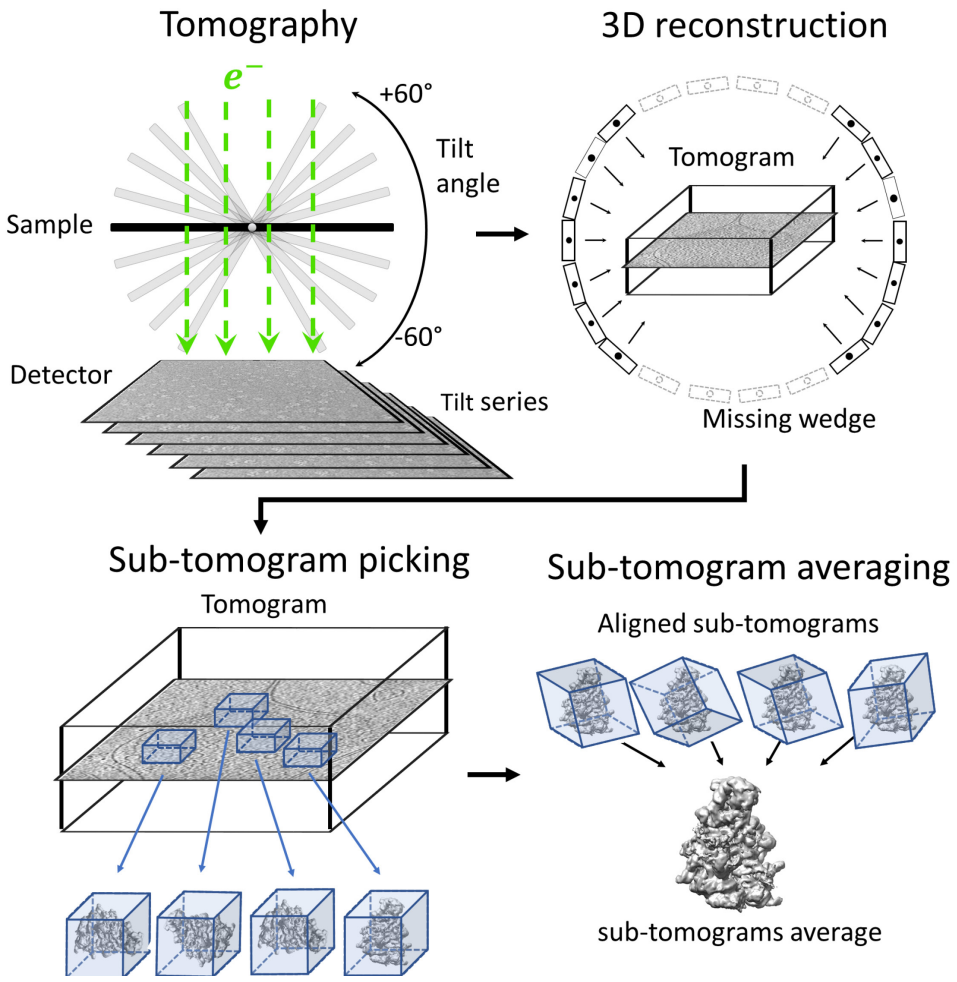

Figure 5: Cryo-ET workflow. A tilt-series is acquired by tilting the sample in the microscope and recording a series of images at different tilt angles (typically between -60° and 60°). The tilt series is back-projected to reconstruct a tomogram. The missing tilt angles in the 3D reconstruction correspond to the missing wedge, which causes anisotropy in the tomogram (the anisotropy is along the z-axis for the tilt performed around the y-axis). The sub-tomogram picking is used to extract the particle of interest from the tomogram into smaller volumes called sub-tomograms. The sub-tomogram averaging (StA) aligns and averages the sub-tomograms to increase the SNR.

The sample preparation in cryo-ET is usually much more complex than in SPA. While small specimens such as viruses, organelles, and small bacteria can be observed in the microscope *in toto* [72], larger cellular samples and tissues must be thinned using cryo-sectioning [73] or focused ion beam (FIB) milling [74] to obtain a section or a lamella that is electron transparent. Moreover, specimens exceeding 5 µm are too large to be vitrified using standard plunge-freezing (as the ice crystals have time to form due to the temperature that decreases too slowly) and require techniques such as high-pressure freezing [75].

Some cryo-ET image processing steps are similar to those used in SPA, such as motion correction of the images, CTF estimation, particle picking, iterative determination of particle

poses, etc. However, there are several differences and specificities. Firstly, as the sample holder tends to drift when changing the tilt angle, the images constituting the tilt series are misaligned which, if untreated, greatly limits the achievable resolution of the reconstructions. Therefore, tilt series are aligned in a common frame using computational methods [76, 77] or using high-contrast fiducial markers in the sample such as gold nanoparticles, which can be tracked by alignment software [78]. The estimation of CTF also differs in cryo-ET as the CTF depends on the particle height in the tomogram. Recent advances in the estimation of the 3D CTF allowed improvements in tomograms resolution [17, 53].

Another challenge in cryo-ET is identifying the protein of interest in the tomograms, which is particularly difficult due to the crowded cellular environment of *in situ* samples and the low SNR of the tomogram. Manual picking can be performed using different software packages, such as IMOD, Dynamo, Eman2, and ScipionTomo [16, 79-81]. However, high-resolution reconstruction relies on detecting a large set of particles, which requires automated methods. Template matching-based [16, 17, 82, 83] and semi-automated deep learning-based [62, 84, 85] methods can be employed on tomograms with large enough complexes, such as ribosomes, but they are prone to false-positive detection for smaller complexes. When studying complexes embedded in a larger biological entity, for instance, a membrane, geometry-based methods [16, 86] constrain the picking to a prior geometrical shape (e.g. the surface of the membrane) reducing the risk of picking false-positive.

The next step in cryo-ET workflow is STA, which aligns and averages subtomograms to increase the SNR and obtain more isotropic resolution (averaging of a large number of particles with different orientations tends to cancel the missing wedge). STA employs similar algorithms as SPA to optimize the rotational and translational alignment parameters of the subtomograms [16, 17, 81, 83, 87], which is performed by maximizing a metric such as cross-correlation between the subtomograms and a reference (the cross-correlation is constrained in the Fourier space to account for the missing wedge). Recent methods are proposing a “per particle per tilt” refinement that refines the 3D average and the tilt series alignment directly at the level of the 2D images of the particles in the tilt series [16, 17, 81, 88].

I.5. The heterogeneity problem

Macromolecules are flexible entities and therefore most cryo-EM samples contain conformational heterogeneity. If left untreated, the EM maps obtained by particle averaging via SPA and STA would suffer from a loss of signal in the flexible regions, resulting in a lower resolution of the maps locally in these regions. Moreover, as the flexibility of macromolecules is

directly linked to their biological functions, an accurate estimation of the conformational variability is of great interest for understanding how the macromolecules function. The currently most used methods sort particles that belong to globally similar conformational states into a set of finite number of classes and perform 3D reconstruction from the particles in each class. This allows for fast high-resolution 3D reconstructions of a few conformational averages, provided that the optimized biochemical procedures were used to maximize the homogeneity of the sample [3, 4, 89-94]. Such methods are often referred to as discrete classification methods [95-100] and assume that the heterogeneity in the data can be explained by a few discrete states that are well separated. For instance, when the aim is to determine a few metastable states, meaning the states that are the most present in the sample (e.g., open and closed states of a molecule), or compositional heterogeneity (e.g., different assembly states of a molecular complex), discrete classification methods give satisfactory results. However, samples often include less stable conformations (short-lived, that is less present at a given time), which are usually discarded by discrete classification approaches (due to a low resolution of the reconstruction from a heterogeneous class, such class is usually not kept for further refinement).

A more realistic representation of the conformational space would be a continuous probability distribution of the states. Methods for deciphering continuous conformational variability from cryo-EM single particle images do not make any assumption on the number of the different conformational states present in the sample, but rather consider that each particle image may come from a different particle conformation [18-23, 30, 101-105]. These methods are sometimes referred to as continuous-state methods, by analogy with discrete classification methods that are sometimes referred to as discrete-state methods [106]. In this section, I detail the available approaches to tackle the heterogeneity problem including discrete-state and continuous-state approaches.

I.5.1. Discrete-state approaches

I.5.1.1. *Classification of single particle images*

Multireference classification

Earlier classification approaches assume prior knowledge of the conformational states present in the sample [107, 108]. In such cases, a set of template volumes (i.e., references) representing the expected conformational states are used to initiate an iterative process to assign each particle to the most similar template. The particles are assigned to different references by comparing the cross-correlation between the particle and a library of 2D projections calculated from the references (i.e., by projection matching). Then, the references for the next iteration are updated by the reconstructed volumes corresponding to the particle class assignment obtained at the current iteration. The iterative process is performed until the reconstructed volumes stabilize (no

significant change in the volumes over two successive iterations). Multireference classification is simple and fairly efficient when the prior template volumes are accurately representing the heterogeneity.

The main problem is the difficulty to find such templates and that the classification is strongly biased towards the initial references. Another limitation is the "attraction problem" [109]. During the classification process, the particles tend to get "attracted" into the classes with the highest SNR, even if they do not belong to these classes. Consequently, some particles can get misclassified and some classes can get very little populated.

Maximum likelihood classification

The currently most-used classification methods avoid using an initial reference but rather infer the class assignment by regularized maximum likelihood (ML) algorithms [12-14, 110]. ML-based methods define a Bayesian model that places on the class assignment parameter a probability distribution that the particle belongs to any of the available classes. The classification algorithms simultaneously estimate the most probable (in the sense of the maximum a posteriori) class assignment together with the most probable alignment parameters using optimization algorithms such as expectation-maximization [12, 14] or stochastic gradient descent [13, 110].

The number of classes is the major parameter that needs to be predetermined. The guess of this parameter is challenging and can be obtained by preliminary knowledge of the system, or by testing different values until the class number is too large (e.g. if several 2 or 3 classes are the same). If the number of classes is too small, it leads to merged states with averaged flexibility. On the contrary, a large number of classes tend to get the optimization unstable while increasing the computational cost. Therefore, when the conformational changes are continuous, the number of classes cannot satisfyingly represent the sample heterogeneity, resulting in a number of heterogeneous classes. Most often, to go around such limitations, the classification is performed in an iterative and hierarchical scheme with human intervention at each stage to discard the particles that the classifier fails to assign to a relevant class [111]. In many cryo-EM studies, it is common to discard the majority of the particles if they are not sorted into a relevant class (Figure 6). Some of the incorrectly classified states correspond to wrongly picked particles, but some other may correspond to interesting continuous-transition states. Therefore, such discrete classification methods result in a largely incomplete representation of the conformational heterogeneity of the sample. Note that ML-based classification is also subject to the "attraction problem".

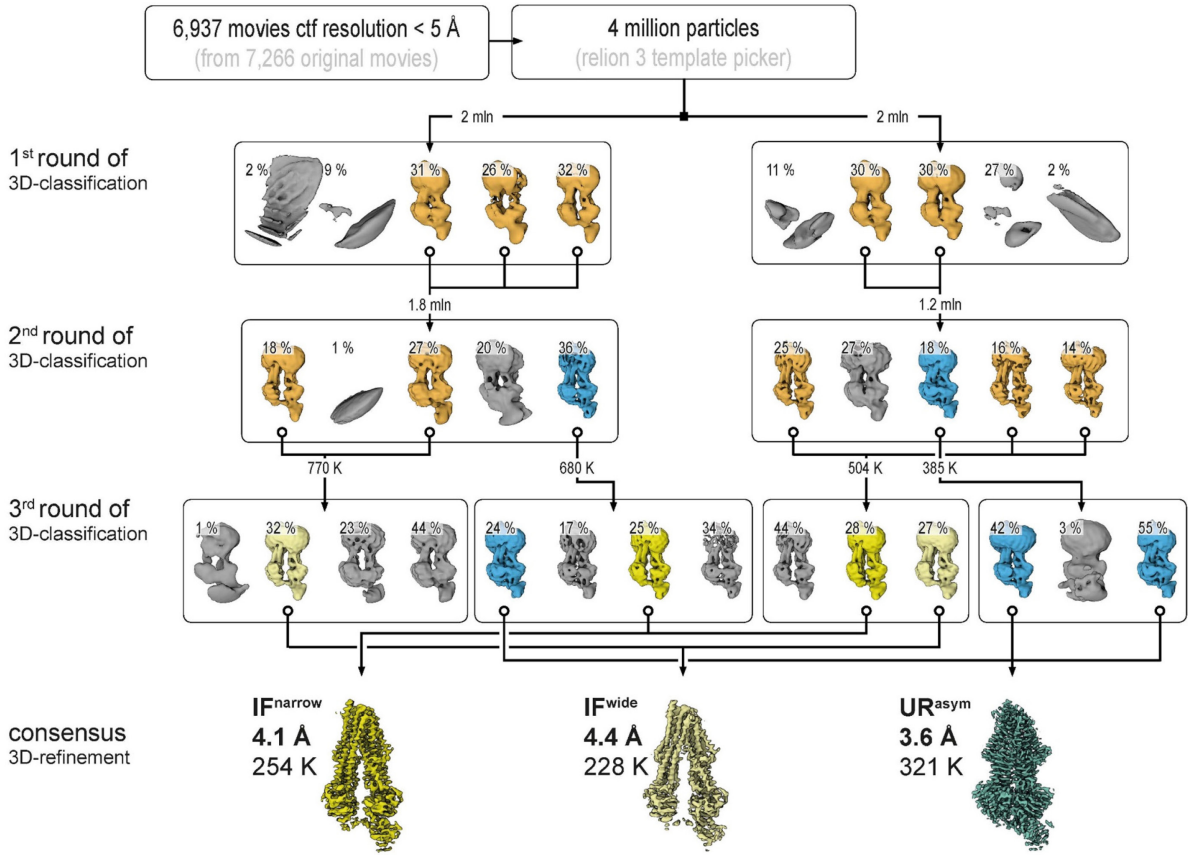

Figure 6: Example of a 3D classification performed in a hierarchical scheme to identify a few discrete states of an ABC exporter [4]. At first, 4 million particles are analyzed, and after 3 rounds of classification, the analysis results in 3 conformational states corresponding to 3 discrete classes cumulating 800K particles (20% of the original set of particles). Adapted from [4].

Focused and multi-body classification

It is observed that conformational changes of macromolecules often involve continuous motion of one or several domains that remain rigid during the motion. Therefore, the macromolecule can be seen as a set of rigid bodies continuously moving relative to each other and having a stable internal structure. In such cases, standard 3D classification would most often fail to precisely reconstruct the moving domain. Nevertheless, the classification can be focused on a specific region of the map, for instance by segmenting the targeted domain and masking the other domains by subtracting the corresponding density in each image. This method referred to as focused classification is focused on analyzing local variability in order to improve the resolution of the reconstruction in the focused region. This process can be extended to multiple rigid domains. For instance, the reference map can be manually segmented into its assumed rigid domains and the particles aligned and classified separately in each identified rigid domain using focused classification. This process is referred to as multi-body refinement or classification and was popularized by its implementation in the software package Relion [112].

Although this method achieves high-resolution of flexible domains in some cases of complexes, it may not be suitable for studying all complexes. Also, it does not solve the problem of identifying the continuous distribution of conformations. For instance, it cannot estimate the distribution of conformations of several domains moving relative to each other. Moreover, the prior made on the rigidity of the domains is not suited to all the types of continuous motions. Finally, the segmentation of the rigid regions must be done manually and must consider that the classification can get unstable when the size of the focused region is too small (smaller than 100 kDa) due to the small amount of signal (pixels or voxels) remaining in the focused region.

I.5.1.2. *Classification of subtomograms*

In the context of structural heterogeneity in cryo-ET subtomograms, the joint classification and alignment of subtomograms have been performed using multireference classification [113] and ML-based classification [15, 87]. These approaches are similar to multireference classification and ML-based classification approaches used in SPA. Therefore, their advantages and limitations are globally the same as those of similar methods used in SPA. However, it should be noted that the classification of volumes is much more computationally expensive than the classification of images.

An alternative to these methods is post-alignment classification [113-115]. Indeed, once the subtomograms are aligned with respect to a single reference (using STA) one can directly calculate the covariance matrix of the subtomograms. This covariance matrix can be used for classification, for instance using dimensionality reduction techniques such as principal component analysis (PCA) followed by K-means clustering [115], or by passing the covariance matrix directly to a hierarchical clustering algorithm [113, 114].

Post-alignment classification has the advantage of being computationally efficient and not suffering from the "attraction problem". Additionally, the number of classes is not required to be selected in advance of using post-alignment classification. For instance, when using hierarchical clustering, the user can visualize the dendrogram tree of the clustering before deciding the number of classes. However, the post-alignment classification is strongly dependent on the quality of the STA alignment, which can be a problem for low SNR datasets or highly heterogeneous datasets.

I.5.2. Continuous-state approaches

Recently, several approaches have been developed to represent conformational heterogeneity in a continuous space. The common point of all the methods is that they represent the conformational variability by estimating an underlying manifold of conformations in a low-dimensional space. Such conformational space can be plotted for instance in two or three

dimensions, to obtain a comprehensive view of the distribution of conformations. The different methods mostly vary in the way they estimate the conformational space.

Another important consideration of continuous-state approaches is the particle pose determination. In some continuous-state approaches, an approximate estimation of the pose parameters (3 Euler angles and 2 in-plane shifts for each particle in SPA, 3 Euler angles and 3 shifts for each subtomogram in cryo-ET) is performed in advance (e.g., using discrete-state approaches) and these pre-determined particle poses are either kept fixed during the search for the particle conformation or refined during the conformational search. The conformational search methods that keep the pre-determined particle poses fixed assume that the conformational heterogeneity of the particles is much smaller than their orientational heterogeneity so that the prior orientational search is not strongly affected by the conformational heterogeneity and thus results in an enough accurate angular assignment. In practice, this is valid only for conformational changes induced by local motions (e.g., motions of a small portion of the complex). In the case of large-amplitude global conformational rearrangements, the prior particle pose assignment is unreliable for an accurate determination of conformations. In such cases, a refinement of the particle poses during the search for conformations is required. However, it is worth noting that a few continuous-state approaches perform a quasi-simultaneous determination of the particle poses and conformations. In this section, I review the existing continuous-state approaches.

I.5.2.1. *Continuous-state approaches for SPA*

Linear combination of principal volumes

One approach to represent a set of given volumes is by decomposing their variability into a linear combination of "principal components" or "principal volumes". This representation assumes that each conformation can be represented by a sum of a reference volume and a linear combination of principal volumes. Several methods define the principal volumes by calculating the 3D covariance matrix from 2D images [105, 116, 117]. Note that these approaches were combined with classification approaches to recover discrete conformational states. Other methods directly estimate the principal volumes using a probabilistic PCA [20, 22, 118]. For instance, cryoSPARC [13], largely used in the SPA field, includes a probabilistic PCA method named 3DVA [22].

Although these methods can be useful to identify the regions of high variability in the reference volume, the linear approximation usually deforms the density in a way that does not preserve the object's size and thus creates non-physical deformations of the object. For instance, a large-amplitude motion of an object in a linear space tends to increase the size of the object [119]. Therefore, the motion can be retrieved at high resolution only in the case of small amplitudes of motion. Moreover, conformational changes often involve non-linear motions that could not be

represented with these methods. It should be noted that these approaches consider that the particle pose is known in advance and, therefore, they are strongly dependent on the accuracy of the prior pose determination. Finally, these algorithms can only produce an estimation of the continuous conformational variability through deformations around the reference volume. In the case of SPA, they cannot perform 3D reconstruction directly using the 2D images [10].

Per-view manifolds of particle images and manifold embedding

Another way to estimate the manifold of conformations is by first calculating the manifold of conformations at the level of the 2D images and in a second step reconstructing the manifold of the 3D volumes [18, 101]. Assuming that the particle pose is given and accurate enough, considering the conformational heterogeneity, the projection directions of the particles can be sorted into a number of viewing directions on a sphere discretized using a given angular step. In the first stage, the particles from each subset of views are analyzed using manifold-learning algorithms (see section II.3.1) based on diffusion maps [120] that calculate a manifold of conformational change in each viewing direction. In the second stage, the manifolds from different views are embedded into a common space of conformation using manifold embedding based on nonlinear Laplacian spectral analysis (NLSA). Finally, the reconstructed manifold can then be converted into an energy landscape [18] (see section II.3.2).

The approach successfully extracted energy landscape from experimental data [18, 101, 121, 122]. However, the method is strongly affected by the accuracy of the predetermined poses of the particles, which directly affects the accuracy of the per-view manifold determination. Moreover, it is common to have missing projection directions in SPA samples. Therefore, the disparity in the distribution of the projection directions is a limitation, as the per-view manifolds cannot be calculated for certain views.

Approaches based on deep learning

In recent years, several deep learning approaches emerged to tackle the heterogeneity problem [23, 27, 30, 123, 124]. These methods aim at inferring a low-dimensional latent variable that describes the conformational heterogeneity by using an encoder-decoder deep neural network. An encoder is trained to project the particle images on the latent variable and a decoder maps the latent variable onto a 3D structure. Different types of deep network architectures have been proposed in the literature (for a detailed review, see [125]). The existing deep learning methods also differ in the volume representation. Some methods use a discretized representation of the object, with a voxel grid representation of the volume in the Fourier space [23, 27, 126]. Some other methods use a mixture of Gaussian functions in a continuous space to represent coarse-grained atomic or pseudo-atomic models [30, 123, 124] (see section II.1).

As discussed previously, it is difficult to decouple the determination of the particle pose from the determination of the conformation. If predetermined, the particle pose parameters should be refined during the estimation of the particle conformation. Currently, the majority of deep learning methods keep the particle pose constant during the conformational estimation, such as 3DFlex of cryoSPARC software package [27]. Only a few deep learning methods determine the particle pose and the conformation jointly. For instance, a more recent version of cryoDRGN (cryoDRGN2) uses a classical, non-deep-learning global search of rigid-body parameters at some iterations of deep learning of the deformable volume in order to refine the particle poses for a more accurate learning of the volume [127]. Another method jointly learns the poses and the conformations with a complex encoder-decoder network, but difficulties to train both simultaneously have been reported and the method has not yet been tested using experimental data [124]. Other challenges of these methods are their sensitivity to initial conditions and to tuning of hyperparameters. For instance, the dimension of the latent variable in cryoDRGN [23] is usually molecule-dependent and can greatly affect the obtained conformational space [125].

I.5.2.2. *Continuous-state approaches for cryo-ET*

Flexible alignment of sub-tomograms using optical flow

Recently, a method named Tomoflow [24] has been developed by colleagues from my lab for analyzing continuous conformational variability in cryo-ET subtomograms. In Tomoflow, a computer vision technique of dense 3D optical flow (OF) is used in an iterative process to quasi-simultaneously refine the rigid-body alignment of each subtomogram with respect to the current-iteration subtomogram average (3 Euler angles and 3 shifts) and estimate the flexible motion field by displacing the voxels of the subtomogram average towards the corresponding voxels in each given subtomogram.

In the first step, the subtomograms are corrected for missing-wedge artifacts and rigid-body aligned with respect to a single reference (e.g., a preliminary subtomogram average). In the second step, the displacement of voxels of the current subtomogram average towards the corresponding voxels of each given subtomogram is calculated using 3D dense OF. This results in a set of deformation (OF) vectors, which is obtained for each subtomogram and called 3D OF. The obtained 3D OFs are then used to warp to the subtomogram average to get denoised versions of the subtomograms. The warped versions of the subtomogram average are referred to as “matched subtomograms”. In the third step, these “matched subtomograms” are used to refine the rigid-body alignment parameters using Fast-Rotational Matching (FRM), which leads to calculating a refined subtomogram average for the next iteration. These steps are repeated until convergence. Once the

iterative procedure has converged, the estimated 3D OFs are projected onto a low-dimensional space through PCA.

The method was used to analyze an *in situ* dataset of nucleosome subtomograms, producing the results that are in accordance with those of another method tested using the same data set (HEMNMA-3D, an hybrid method, see next section I.5.2.3 and section II.5.1) and with the results of theoretical studies [128]. One limitation of this method is that the OF are computed on the subtomogram in the real space and are not compensating for the missing wedge. Instead, the missing wedge is corrected beforehand. The missing-wedge correction is performed by filling the region of the subtomograms corresponding to missing-wedge in the Fourier space by the corresponding region in the Fourier space of the subtomogram average. Although this method corrects the anisotropic resolution of the subtomograms, it tends to attenuate the amplitude of conformational changes towards the conformation used to fill the missing-wedge (the subtomogram average).

Approach based on deep learning

Very recently, a deep learning-based approach was proposed for cryo-ET, TomoDRGN [126]. The method uses an encoder-decoder deep neural network as other deep learning-based methods for SPA and shares similar a design with cryoDRGN [23], with the principal difference that the encoder learns cryo-ET data instead of SPA data. Instead of analyzing subtomogram volumes (as in TomoFlow [24]), TomoDRGN analyzes the sub-images of the tilt-series corresponding to the subtomogram. This has the advantage of being more computationally efficient (as it processes images instead of volumes) and removes the need to correct for missing wedge artifacts (as missing-wedge artifacts arise only when reconstructing the volume from the tilt-series). The method was tested on synthetic data and experimental data of ribosome *in situ* [126]. Note that in TomoDRGN, the pose alignment of the subtomograms is kept constant during the conformational estimation.

I.5.2.3. *Hybrid methods*

When an atomic model of the macromolecule under study is available, this existing structural information can be exploited to infer the conformational heterogeneity in the cryo-EM data [19, 28, 29, 102]. In this context, the experimental data analysis can be combined with simulations of the molecular mechanics (e.g., molecular dynamics simulations or normal mode analysis), where the simulation produces conformational states that are compared with the experimental data to extract those conformations that are present in the data. These methods are referred to as hybrid methods. My PhD thesis was focused on the development of hybrid methods for SPA and subtomogram data analysis based on molecular dynamics simulations, for the conformation and

pose determination in the context of large conformational variability and large datasets. My PhD thesis resulted in the publication of three journal articles, which describe three new hybrid methods [25, 26, 38]. The hybrid methods that existed before the start of my PhD thesis and those published later are extensively described in the next chapter (see section II.5). The hybrid methods developed in my PhD thesis are described in full detail in separate chapters (Chapters III, IV, V).

Chapter II. Simulating the mechanics of macromolecules and application to hybrid methods in cryo-EM

As presented in the previous chapter, cryo-EM provides an experimental technique particularly suited to studying the conformational variability of macromolecules. The methods developed in this thesis aim at characterizing this conformational variability by taking advantage of *a priori* structural information (e.g. an atomic model). By using methods that simulate the molecular mechanics of macromolecules, it is possible to generate a large number of conformational states, which can be compared to cryo-EM data to infer conformational variability. Such approaches, referred to as hybrid, combine analysis of experimental data and computational simulation of molecular mechanics. However, simulating accurately the conformational variability is challenging due to the complex structure of macromolecules and requires computational models integrating physical and chemical knowledge.

In this chapter, I detail the principal approaches to simulate the mechanics of macromolecules, from the most realistic models to the most computationally efficient, and show applications to hybrid methods for structural modeling of cryo-EM maps and conformational heterogeneity analysis from cryo-EM data. The techniques detailed in this section are the basis on which I developed the methods presented in this thesis.

II.1. Representation of molecular systems

Simulating the mechanics of a molecular system relies on *a priori* structural information of the studied macromolecule, which is most often defined by combining information of its amino acid (or nucleic acid) sequence and the position of its N constituting atoms in a 3D Cartesian space, referred to as atomic model:

$$\boldsymbol{r} = (\boldsymbol{r}_1, \dots, \boldsymbol{r}_N)^T, \boldsymbol{r}_n = \left(r_n^x, r_n^y, r_n^z\right). \quad \text{(II-1)}$$

Such 3D atomic model can be derived from experimental data (e.g. X-ray crystallography, cryo-EM) or estimated based on the protein sequence using homology modeling [129, 130] and more recently deep learning-based structure prediction [31, 131] such as AlphaFold2 [31].

Methods that simulate molecular mechanics predict a model of the molecular structure, usually through a potential energy function, that translates the interactions between atoms that maintain

the structure in an energetically favorable state. This model of the potential energy is then used by simulation methods, either by directly drawing samples or trajectories from the potential energy (e.g. Molecular Dynamics (MD) [32]) or by computing directions of motions that minimally increase the energy (e.g. Normal Mode Analysis [132]). The potential energy function defines the rules of interaction between the atoms. This function defines the sum of all terms of the potential energy applied to the atoms of the system, including the macromolecule and its surrounding environment (e.g. solvent molecules, ligands, lipid bilayer for membrane proteins). Most detailed models even predict sub-atomic interactions by considering each nucleic and electronic particle and estimating their quantum effects. However, quantum mechanics simulations are limited to very small systems (systems smaller than a nanometer width) due to the high computational cost of integrating the Schrödinger equation and are out of the scope of this thesis. In this section, I will detail the principal approaches for modeling the potential energy of molecular systems.

II.1.1. All-atom models

Realistic models of potential energy define a set of semi-empirical potentials describing the physical inter-atomic interactions. The potentials are a set of equations and parameters referred to as force fields that can slightly differ, depending on the model, with popular examples being CHARMM [133] and AMBER [134]. In most force fields, the potential energy function $U(\boldsymbol{r})$ is divided into two categories of inter-atomic interactions, the bonded potentials, which describe the interactions between atoms that form covalent bonds, and the non-bonded potentials, which define long-range forces such as electrostatic forces and Van der Waals forces.

$$U(\boldsymbol{r}) = U_{bonded}(\boldsymbol{r}) + U_{non-bonded}(\boldsymbol{r})$$

Bonded potentials preserve the distances and the angles between bonded atoms. These interactions generally include the terms that account for the length of the covalent bond, the angle between two bonds, and the "dihedral angle" or "torsion angle" formed by three successive bonds. The typical bonded potential term can be modeled by a sum of two simple harmonic potentials, for the bond length and angle between two bonds, and one periodic potential for the torsion angle:

$$\begin{aligned} U_{bonded}(\boldsymbol{r}) = & \frac{1}{2}\sum_{bond} k_d\left(d_{ij} - d_0\right)^2 \\ & + \frac{1}{2}\sum_{angle} k_\theta\left(\theta_{ijk} - \theta_0\right)^2 \\ & + \frac{1}{2}\sum_{torsion}\sum_{m} k_\phi\left(1 + \cos\left(m\phi_{ijkl} - \gamma_m\right)\right), \end{aligned}$$

where d_{ij} is the distance between connected atoms i and j, θ_{ijk} is the angle between two successive bonds formed by atoms i, j, k, and ϕ_{ijkl} is the torsion angle formed by three successive bonds related to atoms i, j, k, l. All the other terms are semi-empirical constants defined by the force-field parametrization.

Non-bonded potentials account for the interactions between atoms that are not connected by a covalent bond. The simplest models include two most essential potentials, namely the Lennard-Jones potential, which allows modeling the short-range repulsion and the Van der Waals forces, and the electrostatic potential, which accounts for the electrostatic forces between charged atoms. The Lennard-Jones potential is defined as follows:

$$U_{LJ}(\boldsymbol{r}) = 4\epsilon \sum_{pairs} \left[\left(\frac{\sigma}{d_{ij}} \right)^{12} - \left(\frac{\sigma}{d_{ij}} \right)^{6} \right],$$

where d_{ij} is the distance between atoms i and j. The power-of-12 term is repulsive and dominant in a very short range (less than one angstrom), to avoid clashes between atoms, whereas the power-of-6 term is attractive and dominant in a longer range, to account for Van der Waals forces. The other essential potential is the electrostatic potential based on the Coulomb law:

$$U_{Coulomb}(\boldsymbol{r}) = \sum_{pairs} \frac{q_i q_j}{4\pi\varepsilon_0 {d_{ij}}^2},$$

where q_i is the charge of the atom i and ε_0 is the permittivity of free space.

Note that, in principle, the non-bonded potentials are computed between each pair of atoms, which leads to a complexity of $\mathcal{O}(N^2)$. In practice, the long-range potentials can be truncated after a certain cutoff distance or can be approximated with advanced methods, such as the particle mesh Ewald [32], which decreases the complexity, resulting in a complexity between $\mathcal{O}(N)$ and $\mathcal{O}(N \log N)$.

II.1.2. Coarse-grained models

Due to their large number of degrees of freedom, all-atom models are computationally expensive and are either limited to small systems or short processes. Coarse-graining is the process of reducing the resolution of the model by representing a group of several atoms by a single node which is connected to others nodes with bonds (see Figure 7b). Coarse-grained models were developed to be more computationally efficient while still permitting to simulate accurately macromolecular motions (e.g. protein folding, conformational transitions). Various levels of resolution were proposed [135-139] for coarse-graining, for instance by representing only the heavy atoms of the protein (i.e. non-hydrogen atoms: N, O, C, etc.), or by restricting to the atoms

of the backbone (without the side-chains), or even with only a single node per residue (e.g. at the position of the Cα atoms for proteins and P atoms for nucleic acid chains). For an exhaustive review, the reader can refer to [140].

Coarse-grained models also differ in the strategy used to define their forcefield. Physically-based forcefields are directly translating classical all-atom forcefield to a coarse-grained representation, which is difficult as it necessitates modifying the potentials to account for the degrees of freedom lost by the coarse-graining and adjusting the forcefield parametrization accordingly. An example of successful physically-based coarse-grained forcefield is the MARTINI model [141], which uses a coarse-graining scheme with groups of four heavy atoms represented by one node.

Another class of force fields, referred to as knowledge-based force fields, uses statistical information from a large amount of known structures, such as those available in the Protein Data Bank (PDB), to predict the atomic interactions in the coarse-grain representation [142].

Structure-based force fields, also known as Gō-like models, as they share similar concepts with the model proposed by Gō [143], are defined based on the "native state" of the structure (i.e. the configuration of the initial model). In these models, the bonded interactions typically follow classical potentials while the long-range potentials are restricted to pairs of atoms that are natively in contact in the initial model, simplifying the calculations of the non-bonded potentials. Gō models can successfully predict folding and stretching of proteins [144, 145]. However, they tend to bias the dynamics towards the initial structure [146]. An example of coarse-grained Gō-like model is the Cα Gō model described by Clementi and collaborators [147], which represents the proteins with Cα atoms of the amino acid chain.

II.1.3. Elastic Network Model

A further level of simplification can be obtained by the Elastic Network Model (ENM) proposed by Tirion [148]. In this model, the specificity of the different atoms is ignored and all the atoms (or pseudo-atoms) are similarly connected by simple harmonic springs within a certain distance cutoff (see Figure 7c):

$$U = \frac{1}{2} \sum_{i,j;\, r_{ij}<R} k\left(d_{ij} - {d_{ij}}^0\right)^2 \qquad \text{(II-2)}$$

where d_{ij} is the distance between atom i and j, R is the cutoff distance beyond which the inter-atomic interactions are ignored, ${d_{ij}}^0$ is the initial inter-atomic distance in the initial model and k is the spring constant that is typically chosen to be the same for all pairs of atoms. This simple

spring model allows for elastic deformations around the initial model and showed to be very effective at predicting conformational transitions of macromolecule [149], with the main advantage of being simple and efficient to sample, for instance using normal mode analysis (see section II.2.3). Another advantage of ENM is that it can be applied to molecular systems without a prior atomic model, for instance, using a pseudo-atomic representation of a low-resolution cryo-EM map [19, 150]. In that case, a pseudo-atomic model can be obtained by placing a 3D Gaussian function on each pseudo-atom in the way to well represent the EM density. With the ENM, such pseudo-atomic models without prior information of the sequence or atomic bonds showed similar flexibility as actual atomic models [150]. However, the ENM has strong limitations, especially that it is biased towards the initial conformation and therefore, remains valid only in the vicinity of this conformation [10]. Moreover, the existing springs between atoms in the initial conformation cannot be detached, and new springs cannot be formed, which does not account for motions such as domain association and dissociation [151].

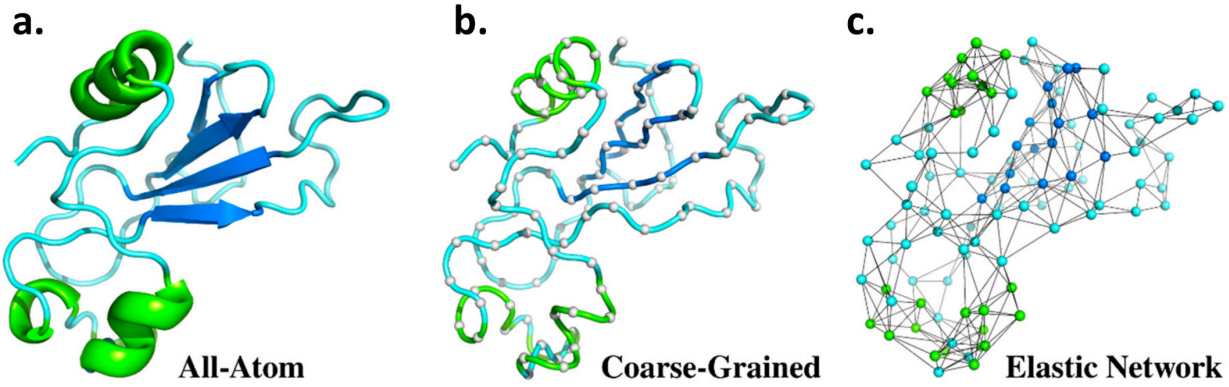

Figure 7: Example of different models of representation of protein structure (accession code in the Protein Data Bank: 1A2P). (a) All-atom model including all the atoms of the protein structure shown as a ribbon diagram. (b) Coarse-grained model with a single node per residue (Cα atoms) connected by bonds (shown as tubes). (c) Elastic network model with a single node per residue (Cα atoms) connected by harmonic springs within a cut-off distance of 7 Å. Adapted from [152].

II.2. Sampling the conformational space

Once the potential energy function is defined for the studied system, different algorithms can be used to generate new samples or trajectories from this potential energy function and explore the conformational space.

II.2.1. Molecular Dynamics

A widely-used approach to explore the conformational space is Molecular Dynamics (MD) simulations [32]. MD generates deterministic trajectories directly sampled from the potential

energy function. The trajectories are obtained based on classical mechanics by predicting the motion of atoms by integrating Newton equation of motion:

$$\boldsymbol{F} = \boldsymbol{M} \cdot \ddot{\boldsymbol{r}}(t),$$

where $\boldsymbol{M}$ is the diagonal matrix, with atomic masses m_i on the diagonal, and $\ddot{\boldsymbol{r}}(t)$ is the second order derivative of the atomic positions with respect to time and $\boldsymbol{F}$ is the sum of forces applied on atoms i, which can be calculated from the potential energy function U as:

$$\boldsymbol{F} = -\frac{dU(\boldsymbol{r})}{d\boldsymbol{r}}. \tag{II-3}$$

Put together, these equations form a second order differential equation of the atomic positions $\boldsymbol{r}$ that is typically solved using numerical integration such as the velocity Verlet algorithm:

$$\begin{aligned} \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \dot{\boldsymbol{r}}(t)\Delta t + \frac{1}{2}\ddot{\boldsymbol{r}}(t)\Delta t^2 \\ \dot{\boldsymbol{r}}(t+\Delta t) &= \dot{\boldsymbol{r}}(t) + \frac{\ddot{\boldsymbol{r}}(t) + \ddot{\boldsymbol{r}}(t+\Delta t)}{2}\Delta t, \end{aligned}$$

where Δt the time step of the integrator. Using this integration scheme, one can perform a simulation as long as desired, with the limitation that the time step Δt is typically very small (in the order of a femtosecond) and therefore long simulations are computationally expensive. The simulation length with all-atom models ranges between a few picoseconds and a few microseconds., whereas for coarse-grained models the simulation can be extended to a few milliseconds and even a few seconds depending on the coarse-graining level [140].

II.2.2. Monte-Carlo methods

On the contrary of MD simulation that generates deterministic trajectories, Monte-Carlo (MC) methods are a class of stochastic algorithms that can be used to explore the conformational space from a potential energy function. MC are general algorithms designed to generate samples from a given probability distribution. In our case, the probability $\mathrm{P}(\boldsymbol{r})$ for a molecular system to be in a certain state $\boldsymbol{r}$ can be defined by a Boltzmann distribution:

$$\mathrm{P}(\boldsymbol{r}) = \exp\left(-\frac{U(\boldsymbol{r})}{k_B T}\right), \tag{II-4}$$

where k_B is the Boltzmann constant, T the temperature of the system, and U the potential energy of the system. As an analytical solution of such high-dimensional probability is not possible and a common approach to estimate the distribution of $\mathrm{P}(\boldsymbol{r})$ is to use Markov Chain Monte Carlo

(MCMC) algorithms [153-155]. In this class of algorithms, the target distribution is estimated by a series of samples that define a Markov chain, which means than each sample $\boldsymbol{r}^n$ depends only on the preceding element of the chain $\boldsymbol{r}^{n-1}$, and the stationary distribution of the Markov chain is the target distribution $\mathrm{P}(\boldsymbol{r})$. The initial state $\boldsymbol{r}^0$ of the Markov chain is typically the initial structure, then the longer the chain is sampled, the closer it gets to the target distribution, so that after convergence, each new sample is a valid state from the probability $\mathrm{P}(\boldsymbol{r})$. The principal challenge of MCMC methods is to efficiently define the transition between $\boldsymbol{r}^{n-1}$ and $\boldsymbol{r}^n$ in the Markov chain.

II.2.2.1. *Metropolis algorithm*

A widely-used sampling algorithm is the Metropolis algorithm [153], which builds the Markov Chain by "random walk" transitions that consist of two steps. First, random perturbations are applied to the structure $\boldsymbol{r}^n$ at the current nth iteration, from a transition kernel $\mathcal{T}$ (usually a normal distribution), and a candidate structure $\tilde{\boldsymbol{r}}$ is generated. The second step adjusts the transition kernel to the target distribution $\mathrm{P}(\boldsymbol{r})$ by accepting or rejecting this candidate structure $\tilde{\boldsymbol{r}}$ with the acceptation probability α :

$$\alpha = min(1, \frac{\mathrm{P}(\tilde{\boldsymbol{r}})/\mathcal{T}(\tilde{\boldsymbol{r}}|\boldsymbol{r}^n)}{\mathrm{P}(\boldsymbol{r}^n)/\mathcal{T}(\boldsymbol{r}^n|\tilde{\boldsymbol{r}})}). \tag{II-5}$$

This algorithm is considered as a random walk because the new candidate is obtained by performing a random deformation of the initial structure and it is accepted only if it decreases the potential energy (or minimally increases the potential energy with a certain probability). However, for most structures, a random displacement of atoms is very likely to be rejected, which in practice means that a lot of candidates must be generated before accepting a new one. For this reason, the Metropolis algorithm is highly inefficient for conformational sampling.

II.2.2.2. *Hamiltonian Monte Carlo*

A method such as Hamiltonian Monte-Carlo (HMC) [154, 155] was developed to enhance the Metropolis approach, especially by performing a more efficient proposal step by removing the random walk behavior. In HMC, the new candidate structure $\tilde{\boldsymbol{r}}$ is obtained by a gradient-based displacement along the potential energy landscape, similar to running a short MD simulation. The candidate structure is then accepted with the acceptation scheme of the Metropolis algorithm (equation (II-5)). HMC was originally called Hybrid Monte-Carlo as it combines deterministic MD trajectory and stochastic Metropolis acceptation scheme [154]. The advantage of this gradient-based proposal step compared to a random walk is that the candidate structure already follows the topology of the potential energy landscape (the atoms are conjointly displaced rather than randomly), which allows for a very high acceptation rate and consequently an efficient sampling.

In theory, HMC could be a more efficient sampler than MD simulations because the time step in the HMC proposal steps can be larger than in MD and because long MD simulations tend to come back to the steps of the MD trajectory that were already visited. Therefore, in practice, the efficiency of HMC requires an accurate tuning of the length and time step of the trajectory in the proposal step, which are challenging. A method that gained popularity recently, as it alleviates this limitation by performing an automated tuning of these parameters, is the No-U-Turn sampler (NUTS) [156]. However, the No-U-Turn sampler is mostly used as MCMC sampler in machine-learning and statistics and is not currently applied to sampling conformational space of molecular systems.

II.2.3.Normal Mode Analysis

Instead of sampling conformational states directly from the potential energy function, another way of simulating dynamics is to describe the flexibility of a macromolecule around a given state using normal mode analysis (NMA) [132, 149, 157]. NMA provides a solution to the equation of motion under a harmonic approximation of the potential energy [132]. In NMA, the initial conformation $\boldsymbol{r}^0$ is assumed to be at the energy minimum. Then, we can write the difference in the energy of state $\boldsymbol{r}$ with respect to the state $\boldsymbol{r}^0$ using the following Taylor expansion:

$$U(\boldsymbol{r}) \cong U(\boldsymbol{r}^0) + \frac{1}{2}\Delta\boldsymbol{r}^T\boldsymbol{H}\Delta\boldsymbol{r},$$

where $\Delta\boldsymbol{r} = \boldsymbol{r} - \boldsymbol{r}^0$ is the conformational displacement and $\boldsymbol{H}$ the $3N \times 3N$ Hessian matrix corresponds to the matrix of the second order derivatives of the potential energy with respect to the atomic positions. Considering a harmonic approximation of the potential energy around state $\boldsymbol{r}^0$ (a small oscillation of the system), the Hessian matrix is diagonalized as follows [148]:

$$\boldsymbol{H} = \boldsymbol{A}\boldsymbol{\lambda}\boldsymbol{A}^T,$$

where $\boldsymbol{A}$ corresponds to the matrix of eigenvectors (A_k) of the Hessian matrix and $\boldsymbol{\lambda}$ is the diagonal matrix containing the corresponding λ_k eigenvalues. The eigenvectors A_k describe the different modes of atomic motions (the directions along which atoms move) and are called normal modes. Therefore, NMA is based on the diagonalization of this Hessian matrix to calculate the normal modes A_k. In this context, the eigenvalues λ_k are associated with the frequency to which each atom vibrates in a certain mode [10, 158]. The lowest-frequency normal modes typically show slow collective motions that correspond to large-scale conformational changes while the high-frequency modes correspond to fast and local dynamics. The eigenvectors A_k form an orthonormal basis so that it is possible to express an atomic displacement $\Delta\boldsymbol{r}$ with the following linear combination:

$$\Delta \boldsymbol{r} = \sum_k \alpha_k \boldsymbol{A}_k, \tag{II-6}$$

where α_k is the amplitude of motion along normal mode A_k. The atomic displacement $\Delta \boldsymbol{r}$ added to the initial conformation $\boldsymbol{r}^0$ produces a new conformational state which depends on the values of amplitudes α_k. The contribution of a specific normal mode l on the atomic displacement is typically observed by varying the value of α_l in equation (II-6) and keeping all the other normal mode amplitudes α_k to zero, with $k = \{1, \dots, K\}, k \neq l$.

In NMA, it is common to reduce the number of conformational degrees of freedom by selecting only the lowest-frequency normal modes that account for global conformational changes (around 10 normal modes), which was shown to be efficient in describing functional conformational changes of macromolecules [150]. Once the normal modes are calculated and the lowest-frequency modes selected, the computational expense of an atomic displacement is negligible compared to, for instance, MD simulations. However, the calculation of the Hessian matrix can be difficult when using standard all-atom force fields, for this reason, NMA is often performed using an ENM model which simplifies the Hessian calculation. The diagonalization of the Hessian matrix is still a challenging computational task due the size of the matrix ($3N \times 3N$) that scale with the number of atoms (or pseudo-atoms) in the system. An extension of NMA, the rotation-translation block (RTB) method [157] propose a further level of coarse-graining to reduce the dimension of the Hessian matrix, where the atomic structures are converted into blocks, each block containing multiple residues and having 6 degrees of freedom (3 rotations and 3 translations).

The principal limitations of NMA is that the harmonic approximation of the potential energy lead to a linear combination of normal modes that is valid only within short-amplitude displacements around the conformation of the initial model. Large-amplitude displacements along the normal modes tend to distort the secondary structure elements and break bonds between atoms. One approach proposed an extension of the RTB method to model non-linear displacements using normal modes [159] which attempts to preserve the structure from the distortions. Furthermore, reducing the number of degrees of freedom to a few low-frequency normal modes limits the exploration of the energy landscape to only certain directions in this landscape.

II.3. Analysis of molecular ensemble

Methods that simulate molecular mechanics, such as MD or MC simulations, produce a large number of conformational states that form an ensemble of molecular structures. Visualizing conformational variations of such molecular ensembles is challenging as an ensemble lies on a high dimensional space, typically a $(M \times 3N)$-dimensional space for a system with N atoms and

M samples. Therefore, it is important to find strategies to identify a few dimensions that are sufficient to describe the conformational variability. For that, different dimensionality reduction methods can be employed that aim at finding the best possible embedding of an ensemble of structures onto a low-dimensional space (e.g., two or three dimensions that can be plotted).

To reduce the dimension of a molecular ensemble, an approach is to define appropriate functions of the atomic positions that represent collective motions of the atoms. In MD, these functions are referred to as collective variables (CV) and can be built on prior knowledge of the expected geometric variations (e.g., atomic distances or bond angles, secondary structure elements, domains etc.). This allows to have standard metrics that can be used for comparison of simulations in different conditions. However, in the absence of knowledge of appropriate CVs, dimensionality reduction algorithms can be used to automatically calculate an appropriate low-dimensional space. In this section, I describe the principal dimensionality reduction algorithms.

II.3.1. Dimensionality reduction algorithms

II.3.1.1. *Principal component analysis*

The most widely-used dimensionality reduction algorithm is by far PCA [160, 161]. PCA approximate the dissimilarities in the dataset by a linear combination of principal components. PCA allows to project a dataset of high-dimension such as a molecular ensemble $\boldsymbol{X} = \{\boldsymbol{r}^1, \dots, \boldsymbol{r}^M\}$ of $3N \times M$ size composed of M molecular structures $\boldsymbol{r}^m$ including N atoms onto a low-dimensional linear subspace of size $L << 3N$:

$$\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X},$$

where $\boldsymbol{P}$ is matrix of size $L \times 3N$ with the rows being the principal components and $\boldsymbol{Y}$ is the matrix $L \times M$ of the low-dimensional projection of the data on the principal components. The principal components are typically obtained by eigen decomposition of the covariance matrix of the data, for instance using singular value decomposition (SVD), where the principal components $\boldsymbol{P}$ are the eigenvectors that form an orthogonal basis on which the data is projected. In the eigen decomposition, the eigenvalues are associated to the amount of variance explained by each principal component, therefore it is usual to keep only the largest eigenvalues that have the strongest contribution to the variance in the data.

PCA is frequently used to identify the representative conformational states of a molecular ensemble as obtained by methods such as MD simulations [162]. It has the advantage of being efficient and simple to interpret, for instance, any point $\boldsymbol{y} = (y_1, \dots, y_L)$ in the L-dimensional space

can be interpreted by a corresponding atomic structure $\boldsymbol{x}$ by performing the inverse projection $\boldsymbol{x} = \boldsymbol{yP}$.

It is important to note that PCA is a linear approximation of a high-dimensional manifold that could be non-linear. PCA calculates a linear subspace that passes through the data. For instance, if the reduced dimension is one, it finds the best projection of the data points onto a line; if the dimension is two, it finds the best projection onto a plane; and so on. However, if the manifold is non-linear and corresponds to large conformational transitions a linear approximation will create distortions [163, 164].

II.3.1.2. *Manifold learning algorithms*

Manifold learning algorithms also referred as non-linear dimensionality reduction methods aim at estimating the underlying non-linear manifold in a low-dimensional space. Instead of a linear approximation, manifold learning methods aim at reconstructing a manifold in a low-dimensional space while attempting to preserve local distances between the data points. Several methods are available using different criteria to measure the distance between data points, including diffusion maps [120], Isomap [165], t-distributed stochastic neighbor embedding (t-SNE) [166], and uniform manifold approximation and projection (UMAP) [167]. UMAP is currently the most widely-used manifold learning algorithm and has the advantage of being computationally efficient and scaling well on very high-dimensional dataset as large conformational ensembles. UMAP performs local manifold approximations by estimating the geodesic distance between the points in the high-dimensional manifold. The geodesic distance, in comparison to the Euclidean distance, follows the curvature of the manifold. Then, the local patches obtained are assembled to construct a topological representation in a low-dimensional space. UMAP has been employed as dimensionality reduction method for MD trajectories of macromolecules and revealed to be particularly effective at separating the different conformational populations while being efficient (computation time comparable to PCA) [164].

II.3.2.Representation as free energy landscapes

Once a conformational landscape is constructed, either by defining an appropriate CV or by applying a dimension reduction algorithm, the points distribution obtained in the low-dimensional space (different conformational states) can then be interpreted as the difference of the free energy ΔG with respect to the state with the minimum of energy, through the Boltzmann factor:

$$\frac{\Delta G}{k_B} T = -\ln\left(\frac{n}{n_0}\right),$$

by counting the number of particles n in each region of the space and the number of particles in the most populated region n_0 (which can be considered to be the state with the minimum of energy, as being the most probable state). Here, k_B is the Boltzmann constant and T is the temperature of the system.

II.4. Hybrid methods for macromolecular modeling in cryo-EM

Hybrid methods are commonly employed in cryo-EM in the case where the resolution of the reconstructed EM map is too low to perform *de novo* structure modeling. In that case, hybrid methods can estimate the conformational state in the EM map in a process referred to as flexible fitting, where mechanics of an atomic model are simulated to flexibly align the atoms with the density observed in the EM map. In this section, I describe the flexible fitting methods that have been developed using the different simulation approaches described in the previous section. The objective of the section is to describe the methods that perform a fully-automated fitting. Therefore, the different software packages proposing manual or semi-automated model builders will not be described.

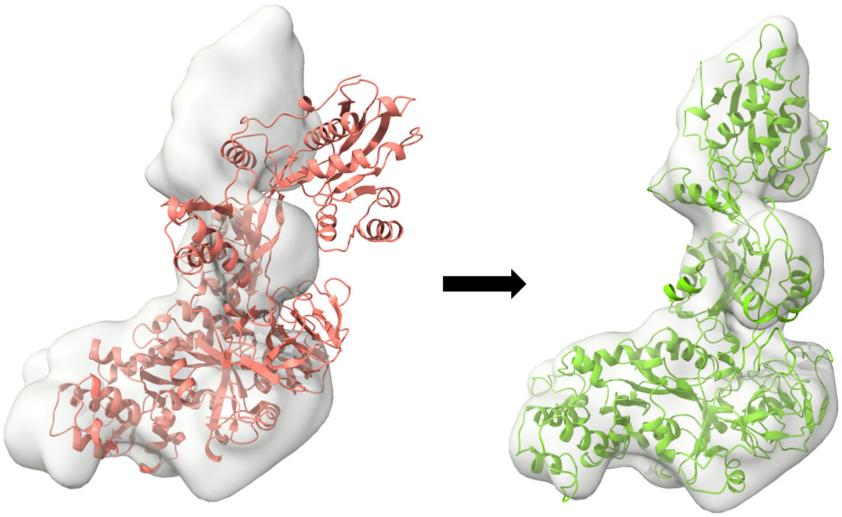

Figure 8: Example of flexible fitting of an atomic model of elongation factor 2 (accession code in the Protein Data Bank: 1N0V) into a synthetic EM map. The flexible fitting employed here (NMMD) combines MD simulations with NMA (see Chapter III).

II.4.1. Flexible fitting using MD simulations

The use of MD simulations to model the conformational state contained in a given EM map is not straightforward. A direct approach would be to run a MD simulation until eventually the trajectory reaches the conformational state that agrees with the experimental data. However, this

approach would be extremely inefficient as there is no guarantee that the trajectory ever reaches the observed state. Therefore, to enhance the sampling, the fitting needs an extra force that drives the simulation from the initial state (determined by the given atomic model) to the target state (determined by the given EM map). A common approach is to insert a virtual potential in the potential energy function, that biases the simulation towards the region in the energy landscape that agrees with the target state [34, 36, 168-171]. This new potential is referred to as a *biasing potential* and is directly added to the total energy:

$$U = U_p + k\, U_f, \tag{II-7}$$

where U_f is the biasing potential and U_p is the potential energy function of the standard MD simulation. The factor k is the force constant that defines the balance between the biasing potential and the classical potential and must be tuned with caution as it strongly affects the fitting results, which will be discussed below. The objective of the biasing potential is to bias the energy so that the minimum of energy is centered onto the target state. Therefore, by simply running an MD simulation with the hybrid energy function in (II-7), the atoms will be dragged to the low-energy regions and converge to the target state. To define the biasing potential, one must incorporate a metric that assess the goodness of the fitting between the atomic model $\boldsymbol{r}(t)$ at time t and the experimental EM density map $\boldsymbol{\rho}_{exp}$ composed of ${N_{vox}}^3$ voxels. To compare $\boldsymbol{r}(t)$ in the Cartesian space of atomic coordinates and $\boldsymbol{\rho}_{exp}$ in the voxel space, $\boldsymbol{r}(t)$ should be converted into a density map. A simple method for that is to use 3D isotropic Gaussian kernels at the atomic positions in the Cartesian space and project (sum up) the Gaussian densities in the voxel space [34]. The simulated density map can be written as follows:

$$\boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r}) = \sum_{n=0}^{N} \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} exp\left(-\frac{1}{2\sigma^2} \parallel (i,j,k)^T - \boldsymbol{r}_n \parallel^2\right), \tag{II-8}$$

where $\boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})$ is the simulated density at the voxel in the center of the coordinate (i,j,k), $\boldsymbol{r}_n$ are the Cartesian coordinates of atom n, N is the number of atoms, and σ is the variance of the Gaussian kernels. The variance σ is usually constant for all the atoms and determines the resolution of the simulated map.

Now that $\boldsymbol{r}$ is converted into a density, it can be compared to the experimental map. A widely-used metric for comparing density maps is the cross-correlation (CC), which has been largely

employed for flexible fitting purposes [34, 36, 170], in which case the biasing potential is as follows:

$$U_f = 1 - CC. \tag{II-9}$$

The CC has values between 0 and 1 and measures the similarity between the given cryo-EM map and the map simulated from the atomic structure during the fitting. A value of 1 means that the densities fit perfectly, whereas a low CC value corresponds to a low similarity. The potential is proportional to $1 - CC$ to guide the simulation to the highest CC values. The CC is defined as follows:

$$CC = \frac{\sqrt{\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})\, \boldsymbol{\rho}_{exp}^{i,j,k}}}{\sqrt{\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})^2} \sqrt{\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{exp}^{i,j,k\,2}}}, \tag{II-10}$$

where $\boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})$ and $\boldsymbol{\rho}_{exp}^{i,j,k}$ are the simulated and experimental density maps, respectively, at the voxel at the position (i, j, k).

The biasing forces are calculated by finding the gradient of U_f with respect to $\boldsymbol{r}$ and as in equation (II-3). Analytical gradient with respect to the atomic positions $\partial CC/\partial \boldsymbol{r}_n$ is therefore calculated [34]:

$$\begin{aligned}\frac{\partial CC}{\partial \boldsymbol{r}_n} = & \frac{\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{exp}^{i,j,k} \frac{\partial \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})}{\partial \boldsymbol{r}_n}}{\sqrt{\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{exp}^{i,j,k\,2} \sum_{l=1}^{N_{vox}} \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})^2}} \\ & - \frac{\left(\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r}) \frac{\partial \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})}{\partial \boldsymbol{r}_n}\right)\left(\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})\, \boldsymbol{\rho}_{exp}^{i,j,k}\right)}{\sqrt{\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{exp}^{i,j,k\,2}} \left(\sum_{i,j,k}^{N_{vox}} \boldsymbol{\rho}_{sim}^{i,j,k}(\boldsymbol{r})^2\right)^{\frac{3}{2}}}.\end{aligned} \tag{II-11}$$

The CC is employed by a large number of algorithms for cryo-EM image processing and has several advantages, for instance it allows to calculate analytical gradient of the biasing forces. Instead of the CC-based potential, another method uses a potential field that pushes the atoms to high-density regions of the EM map [35]. However, contrary to the CC-based potential, the strong bias imposed by the EM map in the potential-field-based method requires additional restraints to

conserve the secondary structure elements and avoid over-fitting while producing comparable fitting results.

II.4.1.1. *Optimization of the force constant*

The force constant k in equation (II-7) is an important hyperparameter that defines the strength of the biasing potential in comparison with the other potentials. This parameter is system-dependent and greatly influences the fitting results. If k is high, there is a risk of overfitting as the atoms are rapidly forced to adopt a particular position regardless to their native dynamics, which induces distortions of the structure such as breaking secondary structures or covalent bonds. Inversely, if k is too low, the simulation is not constrained enough and the atoms are not driven to the target conformation, which is similar as running an unbiased MD simulation. For most flexible fitting methods, this hyperparameter is present and is tuned with the trial-and-error method.

An approach for automated tuning of k was proposed using Replica Exchange Umbrella Sampling (REUS) [36], where multiple replicas are run in parallel and exchange force constants (highest force constants to replicas with the highest CC), to gradually increase the biasing force over the simulation length [36]. In this approach, multiple parallel MD simulations are performed (called replicas), each with a different value of the force constant. During the fitting process, the force constants are periodically exchanged to adjust the optimal k value for each replica. In the REUS method, the probability W for the replica exchange is given by:

$$W(X \to X') = \begin{cases} 1 & \text{for } \Delta \leq 0 \\ \exp(-\Delta) & \text{for } \Delta > 0 \end{cases},$$

$$\Delta = \frac{1}{k_B T}(k_n - k_m)\left(CC^j - CC^i\right),$$

where CC^i is the CC value for the i-th replica and k_n is the n-th force constant, T the temperature, and k_B the Boltzmann constant. This algorithm tends to exchange highest force constants to replicas with highest CC. The effect of these exchanges is to increase gradually the force constant and improve the results of the fitting.

Besides the automated tuning of k, REUS has the advantage of producing different MD trajectories that can be used to estimate the uncertainty of the fitting. If the CC variation in the different replicas is low, it would suggest that all the replicas converged to very similar conformations, which increases the trust in the fitting. On the contrary, high variations in CC suggest that the replicas results are diverging and indicate an uncertainty in the results.

II.4.2.Bayesian model for flexible fitting

The flexible fitting can also be defined in the Bayesian formalism [172], where MC sampling replaces MD simulations. Starting from Bayes' law, we can write the posterior distribution as a function of the model given by atomic coordinates $\boldsymbol{r}$ and the experimental EM density map $\boldsymbol{\rho}_{exp}$:

$$\mathrm{P}(\boldsymbol{r}|\boldsymbol{\rho}_{exp}) \propto \mathrm{P}(\boldsymbol{\rho}_{exp}|\boldsymbol{r})\mathrm{P}(\boldsymbol{r}) \tag{II-12}$$

In this model, $\mathrm{P}(\boldsymbol{r})$ is the prior distribution of the atomic coordinates $\boldsymbol{r}$ that translates the properties of the structure into the model and, similarly as for MC sampling methods, can be defined as a Boltzmann distribution (equation (II-4)). The likelihood distribution $\mathrm{P}(\boldsymbol{\rho}_{exp}|\boldsymbol{r})$ is the probability that a given model with atomic coordinates $\boldsymbol{r}$ agrees with $\boldsymbol{\rho}_{exp}$ and can be assumed to be normally distributed and centered on the density $\boldsymbol{\rho}_{sim}(\boldsymbol{r})$ that is simulated from the atomic coordinates $\boldsymbol{r}$ (equation (II-8)), as follows:

$$\mathrm{P}(\boldsymbol{\rho}_{exp}|r) = \frac{1}{\sqrt{2\pi\sigma_\rho^2}} exp\left(-\frac{1}{2\sigma_\rho^2} \parallel \boldsymbol{\rho}_{exp} - \boldsymbol{\rho}_{sim}(\boldsymbol{r}) \parallel^2\right). \tag{II-13}$$

By assembling (II-4) and (II-13) in equation (IV-2), and by taking the negative logarithm of this function, we obtain the hybrid energy function of the following form:

$$\log \mathrm{P}\left(\boldsymbol{r}\middle|\boldsymbol{\rho}_{exp}\right) = \frac{1}{k_B T} U(\boldsymbol{r}) + \frac{1}{2\sigma_\rho^2} \parallel \boldsymbol{\rho}_{exp} - \boldsymbol{\rho}_{sim}(\boldsymbol{r}) \parallel^2 - \log\left(\frac{1}{\sqrt{2\pi\sigma_\rho^2}}\right).$$

Note that this function is similar to the biased energy function used in MD-based flexible fitting in equation (II-7), with the difference that the CC-based biasing potential is replaced by the Gaussian likelihood in equation (II-13), which can be shown to be closely related [172]. Such Bayesian model can be inferred using MC sampling such as HMC and is implemented in ISD software [172-174]. As the energy function is equivalent for MD-based fitting and for the Bayesian model, the difference between the two approaches lies in the sampling algorithm used (MD, HMC, Metropolis, etc.) as described in the section II.2.

II.4.3.Flexible fitting using normal mode analysis

Flexible fitting based on MD or MC simulation typically incorporates a large number of degrees of freedom (three Cartesian coordinates per atom or pseudo-atom). In the context of flexible fitting

of low-resolution EM maps (e.g. worse than 10 Å), such approaches are computationally expensive whereas a few degrees of freedom might be sufficient to globally describe the target conformation. On the other hand, NMA provide an efficient conformational sampling using a small number of degrees of freedom (typically 10 lowest-frequency normal modes) that is well-suited for flexible fitting of low-resolution EM maps. When using NMA, the CC in equation (II-10) can be directly maximized using gradient-based optimization algorithms by calculating analytically the gradient on the normal mode amplitudes $\partial CC/\partial \alpha_k$ [175]. By finding the optimal normal mode amplitudes, the equation (II-6) can be used to recover the corresponding atomic displacements with respect to the initial conformation. Different optimization schemes were developed [68, 150, 175-177], for instance using Newton's optimization method [175], or using MC sampling on the normal mode amplitudes [68].

However, the limitations of the NMA sampling also apply to NMA-based flexible fitting, for instance, the explored conformational states are restricted to those permitted by the selected low-frequency normal modes. In the case where the normal modes do not well describe the conformational transitions between the initial conformation and the conformation in the EM map, the fitting will most likely fail to overlap the atoms with the EM density. Besides, large-scale conformational transitions will generate distortions on the structure as discussed in 0. These problems can be attenuated by iteratively applying the flexible fitting and recalculating normal modes at each iteration [68, 178] or using non-linear normal modes [159]. However, NMA still suffers from difficulties to fit local structural changes, and as the resolution of recent EM maps tends to be higher, it requires methods that account for local flexibility. Therefore, MD-based or MC-based methods would be preferred, especially as current computational resources frequently allow to use them.

II.4.4.Flexible fitting methods combining local and global atomic displacements

It is frequent that the hybrid energy function contains several local minima in which the simulation can get trapped. In such cases, the simulation may arrive to a state that does not well align with the EM density. This is especially true when a large number of degrees of freedom are permitted such as in MC-based or MD-based methods described above, where the energy function is particularly rugged. Using NMA-based fitting smoothens the energy function as it greatly reduces the number of degrees of freedom. On top of that, NMA sampling is much more efficient, which is particularly important when fitting multiple EM maps as available with cryo-EM studies. However, other limitations arise when using NMA, such as structural distortions for large amplitudes of conformational transitions and limited conformational sampling due to the common

use of a small subset of normal modes (low-frequency high-collectivity modes) instead of all normal modes. Furthermore, while the biasing potential (or likelihood distribution for MC-based methods) can be smooth when fitting low-resolution EM maps, it tends to get rugged for medium to high resolution maps (e.g., 6 to 2 Å). The effect of this is the increase of the chances of getting trapped into a local minimum for higher resolutions of EM maps, which are becoming more and more common. Therefore, flexible fitting methods are required to perform efficiently both global and local fitting. In this subsection, I describe various methods that in one way or another address this problem.

II.4.4.1. *First fitting globally then fitting locally*

An idea of the refinement is to start fitting the EM map globally, and once the fit is close enough to the target conformation, to start fitting the EM map locally. A similar idea was implemented in the multi-resolution MD-based method, CDMD [170]. Here, a biased-MD fitting is performed and the resolution parameter of the simulated EM map (σ in equation (II-8)) is gradually decreased during the simulation, as well as the intensity of the biasing potential (the force constant k) is increased. The idea is that the initial conformational difference between the atomic model and the EM map is large and, therefore, large (global) atomic displacements are needed. Therefore, initially, the resolution of the simulated map is low (the energy function is smooth) and global atomic displacements are performed. The closer the simulation to the EM map, the higher the resolution of the simulated map gets and local conformational differences can then be fitted. Although the method does not explicitly define global degrees of freedom, the low-resolution of the simulated map at the beginning of the simulation allows to avoid getting trapped into local minima.

II.4.4.2. *Stochastic sampling combined with an ENM*

In another approach implemented in the software DireX [179, 180], an ENM is combined with a stochastic conformational sampling to maximize a CC-based potential. Here, local conformational changes are allowed by the sampling. However, the combination with the ENM tends to reduce the degrees of freedom to global motions permitted by the ENM. In this sampling method, the structure is first perturbed by a Gaussian noise (similarly as in random walks MC methods) to generate a candidate structure. Then, a set of corrections are applied to the candidate structure using restraints that move the atoms toward the EM map density while preserving the structural geometry. Three corrections, based on a set of potentials are applied iteratively: i) an ENM potential (equation (II-2)) calculated from the initial conformation, which accounts for global displacements; ii) a CC-based potential calculated using the EM map and the initial conformation (equation (II-9)), which moves the atoms towards the target conformation in the EM

map; and iii) a potential that maintains the structural geometry by combining a bonded potential, which regulates the distances and angles between bonded atoms, and a non-bonded potential with a repulsion term to avoid clashes between the atoms [181]. The three corrections are applied iteratively until the candidate structure fulfills several structural restraints (or the candidate structure is rejected after a certain number of iterations, if the structural restraints are not fulfilled). When the candidate structure is accepted, the parameters of the current structure (bond angles and distances, atom-pair distances, etc.) are used to update the ENM, the CC-based potential, and the distance restraints, and the process is repeated until convergence of the CC. Although the conformational sampling incorporates $3N$ degrees of freedom as all the atoms are permitted to move individually, the ENM encourages the atoms to follow motions permitted by the ENM which reduces the effective conformational degrees of freedom and, in turn, helps preventing overfitting [179]. The limitation of such conformational sampling is the random walk proposal that could get inefficient for large macromolecules, although the corrections applied to the candidate structure tend to alleviate the random walk behavior.

II.4.4.3. *MD simulation combined with NMA*

In an approach called MDeNM-EMfit, proposed by Costa and collaborators [182], local and global displacements are combined by employing MD simulations excited in NMA directions. In this method, an MD simulation is initiated using a linear combination of normal modes whose amplitudes are sampled with an MC approach so as to guide the simulation toward the conformation in the given EM map. Given an initial structure and a selected set of its low-frequency normal modes, this method tries randomly generated combinations of normal-mode amplitudes to find the one that produces the NMA-based structural displacement in the direction of the CC increase. The obtained linear combination of normal modes is then used to initiate the velocities of a short, 2-ps MD simulation. This procedure is iterated until the CC convergence, considering the structure resulting from the MD run at the previous iteration as the initial structure for the next iteration, which involves recalculating normal modes of the initial structure at each iteration. The advantage of this approach is that NMA excitation may helps MD to escape from local minima: if the MD simulation gets trapped in a local minimum, the NMA excitation at the next step can help moving out from this energy region. A limitation of this method is the need to continuously calculate normal modes, every each 2 ps of simulation that represents an additional computational cost to the already expensive MD simulation.

II.5. Hybrid methods for heterogeneity analysis in cryo-EM

In the context of the conformational heterogeneity in cryo-EM, hybrid methods can be employed to analyze 2D single particles images (or 3D subtomograms in cryo-ET) and extract the

conformational variability from the data. In that case, the hybrid methods described in section II.4 are modified to incorporate the ensemble of particle images (or subtomograms) instead of a single EM map. These methods aim at estimating an ensemble of conformational states and use a dimensionality reduction method to construct a conformational landscape that provides the best possible description of the experimental data.

II.5.1. Flexible fitting of particles using NMA

The first hybrid method for heterogeneity analysis in cryo-EM was published in 2014 and is referred to as HEMNMA [19] (Hybrid Electron Microscopy Normal Mode Analysis). In HEMNMA, an atomic model or a pseudo-atomic model (e.g., an EM map converted into a collection of 3D Gaussian functions) is flexibly fitted using an NMA-based flexible fitting to each particle individually (one fitting per particle) and the normal mode amplitudes corresponding to the best possible fit are collected to construct the conformational landscape. HEMNMA uses a 3D-to-2D flexible fitting that maximizes the CC between a 2D particle and a 2D projection of the 3D structure being fitted. HEMNMA uses an iterative procedure where the update of normal mode amplitudes (search of conformational parameters) is alternated with the update of rotations and translations (a rigid-body parameter search that accounts for rotational and translational displacements when updating the conformation). Once the normal modes amplitudes are estimated, they are projected onto a low-dimensional space using a dimensionality reduction method, or directly plotted in the normal mode space. The obtained conformational space is further analyzed by animating motions of the molecular complex along different directions drawn through the data in the densest regions of this space (the so-called trajectories of motions) or by clustering to reveal conformational populations. The points in the conformational space are directly associated to a particle and an estimated pose, therefore, grouping points in this space allows performing 3D reconstruction to observe the different conformations directly on the reconstructed density maps. This point is particularly important as it provides a form of validation of the conformational variability estimated in terms of atomic or pseudo-atomic models (the motions observed when animating the reconstructed EM maps should be equivalent to the motions observed when animating the fitted models).

The limitation of performing one fitting per particle is the computational cost that scales with the number of particles, which makes the analysis on very large datasets difficult. Recently, a deep learning extension of HEMNMA, named DeepHEMNMA [102] was published, which involves a deep neural network that can be trained to emulate the HEMNMA operations, achieving a HEMNMA speed-up of at least 40 times in the experiments shown in [102].

HEMNMA has been also extended to the analysis of subtomograms in cryo-ET [183], and this method is named HEMNMA-3D. In HEMNMA-3D, the fitting using normal modes is performed at the level of subtomograms (3D density maps) and includes an extra step that compensates for the missing wedge during the flexible fitting (the CC is constrained in the Fourier space in the missing-wedge region).

The major limitation of HEMNMA and HEMNMA-3D is the strong prior imposed to the conformational exploration by selecting a few low-frequency high-collectivity normal modes (see section II.2.3. for more details). Indeed, the conformational exploration is limited to the choice of normal modes in order to speed up the analysis of a large number of particle images or subtomograms. Enhancement of this approach would use an MD-based or MC-based fitting to avoid NMA-induced limitations, however the computational cost is prohibitive as MD or MC methods include a high number of degrees of freedom and are several orders of magnitude more computationally expensive than NMA. Strategies to overcome these limitations are the subject of Chapter IV and Chapter V.

II.5.2.Reweighting of a predefined conformational space

In a more recent study, another strategy was employed to define the conformational heterogeneity in cryo-EM by reweighting a prior conformational space. Here the conformational space is not directly estimated, but a predetermined conformational space is adjusted to the experimental particle images to reflect the distribution of each conformation in the data.

In the method named cryo-BIFE [28], a free energy landscape is derived from a set of cryo-EM particles by using a predetermined 1D conformational path. A Bayesian model connects the conformational path to the particle images through a likelihood model that accounts for rotational and translational parameters, CTF parameters and noise [184]. The Bayesian inference estimates the probability distribution of all the parameters together with the probability of each conformational state along the predefined conformational path using a random walk Metropolis sampling. The inferred probability distribution of each state along the conformational path is used for “reweighting” the initial free energy landscape in accordance with the cryo-EM data.

The difficulty of the method is to define the prior conformational path that represents a relevant conformational transition in accordance with the experimental data. In the case where two atomic models are available, a path can be obtained between the two states using steered MD simulations as presented for an experimental dataset of TMEM16F ion channel [28]. However, this is a strong prior as a conformational space can be more complex than a single 1D path between two metastable states. Moreover, in practice, it can be difficult to find two available states for a given system.

This approach has been enhanced to analyze a conformational ensemble instead of a 1D conformational path [29] to extract a free energy landscape. A similar Bayesian model is employed and inferred with the NUTS sampler [156] which is much more efficient than the Metropolis algorithm (see section II.2.2.2). The method was tested on a synthetic data of chignolin folding generated with an unbiased MD simulation. The method was able to accurately perform the reweighting of the ensemble based on the synthetic cryo-EM particles. However, the method is still very dependent on the prior conformational ensemble, indeed, if this ensemble (e.g., obtained by an unbiased MD simulation) does not represent all the possible conformations present in the data (e.g. if the preliminary simulation was too short, if the system was in a different condition, etc.), then the reconstructed free energy landscape will be inaccurate.

Chapter III. NMMD: efficient combination of Normal Mode Analysis and Molecular Dynamic simulation and its use to flexibly fit an atomic model into a cryo-EM map

This chapter presents NMMD, a new hybrid method for flexible fitting of an atomic model into a cryo-EM map, which combines MD simulations with NMA. NMMD was published in 2022 [38]. This chapter describes the NMMD method and its performance with synthetic and experimental data. The Methods, Results, and Discussion-Conclusion sections of this chapter were extracted from the published manuscript [38] and adapted for inclusion in this PhD thesis manuscript.

III.1. Introduction

One challenge of hybrid methods for flexible fitting (see section II.4) is to fit global and local conformational motions simultaneously. On one hand, methods that incorporate a large number of degrees of freedom (MD-based or MC-based) can fit accurately local dynamics but can get trapped into local minima while fitting large-scale global conformational changes, due to the ruggedness of the potential energy function of the system. Also, these methods suffer from a high computational cost. The computational cost of the fitting gets increasingly important when considering fitting of multiple EM maps, as more and more EM maps are getting available in current cryo-EM studies, or even more when considering the use of the fitting in conformational heterogeneity analysis of individual cryo-EM single particle images (see Chapter IV and Chapter V). On the other hand, NMA-based flexible fitting smoothens the energy function by reducing the number of degrees of freedom and explores efficiently the global collective motions, while being much less computationally expensive. However, NMA has a few limitations, among which a difficulty to fit local dynamics unless all the available normal modes are used for fitting, which in turn would increase the computational cost (see section II.2.3).

To combine efficiently the two approaches, I devised a method that combines a NMA-based displacement with a classical displacement model based on atomic-level degrees of freedom (such as in MD-based or MC-based hybrid methods), which results in a speed-up of the fitting and an accurate description of both global and local motions. The proposed method can be seen as an extension of the standard MD-based flexible fitting [34]. The originality of the proposed method lies in empowering the MD-based displacement with a normal mode displacement, which

accelerates the fitting and compensates for the NMA limitations. Prior to the current implementation of NMMD, I have tested the idea of combining NMA and MD simulations by implementing a Bayesian model sampled using HMC [185]. This preliminary work has proved that such hybrid methods can, indeed, be efficient for fitting both local and global displacements and that they are less computationally expensive than the methods that do not use normal modes. The Bayesian, HMC-based method and the results of its use with synthetic data were published in 2021, in the EUSIPCO conference proceeding [185]. This implementation was a proof of concept of the approach, but it required further improvements. One of the conclusions of that preliminary study was that an implementation of the approach in an efficient MD software would significantly speed-up the force calculations. For that, we decided to use GENESIS, an open-source MD simulation software [186], considering a long-term collaboration of our group regarding theoretical molecular modeling with the team of Florence Tama (currently based in Nagoya University and RIKEN-Kobe, Japan) and some previous methods' implementations of the Tama group in GENESIS, in particular an EM-map flexible fitting method solely using MD simulation [34], which served as the basis for implementing the new approach, named NMMD (Normal Mode Molecular Dynamics) [38]. Thus, I implemented the NMMD approach in GENESIS by modifying the existing code of the method for MD-based flexible fitting of an atomic model into an EM map. The NMMD approach is based on the same idea as in the case of the method implemented in the Bayesian framework, namely combining atomic displacements with full (MD) and reduced (NMA) numbers of freedom degrees, but the potential energy function is sampled using MD simulations instead of using HMC.

More precisely, the model in NMMD modifies the atomic displacement permitted in classical MD simulations to incorporate a displacement due to a linear combination of normal modes. The amplitudes of the linear combination of normal modes are estimated simultaneously with the atomic-level displacements, by integrating the equation of motion in the same fashion as in MD simulations. The amplitudes of normal modes are updated at every step of the simulation with almost no extra integration cost, as the number of normal modes is much smaller than the number of atomic coordinates (Cartesian coordinates). NMMD method has been coupled with a multireplica procedure using REUS [36] (see section II.4.1.1) to improve fitting by adjusting the force constant of the biasing potential.

Compared to MDeNM-EMfit (see section II.4.4), a different approach aiming at an efficient flexible fitting of both global and local displacements, NMMD uses an analytical gradient to find the optimal normal mode directions, instead of the random walk approach used in MDeNM-EMfit, which is both faster and more accurate. Also, NMMD requires a single calculation of normal modes (at the beginning of the fitting process) and a single MD simulation run, without any

restriction regarding the duration of the MD simulation. Furthermore, NMMD updates normal-mode amplitudes in each MD simulation step, which improves accuracy of their estimation.

In this chapter, I describe the NMMD approach and present its results using synthetic cryo-EM maps of four molecular complexes (LAO binding protein (LAO) [187], Adenylate kinase (AK) [188, 189], Lactoferrin (LF) [190, 191], and Elongation factor 2 (EF2) [192]) and experimental cryo-EM maps of two complexes (p97 ATPase (p97) [3] and ABC exporter (ABC) [4]). Also, using the same data, I show that the fitting using a combination of MD and NMA samples the conformational space more efficiently than the fitting using MD only, by comparing the results of the NMMD fitting method and the MD-based fitting method that served as the basis for the NMMD development and was available in GENESIS [34]).

III.2. Methods

III.2.1. NMMD: combined MD and NMA based flexible fitting

Here, I describe the NMMD approach that combines NMA-based flexible fitting (small number of degrees of freedom well describing global motions) with MD-based flexible fitting (large number of degrees of freedom well describing local motions). In this approach, the computational cost of fitting large-scale conformational transitions is reduced thanks to the NMA-based fitting and the precision of fitting local dynamics is maintained thanks to the MD-based fitting. In standard MD simulations, the atomic positions $\boldsymbol{r}(t)$ at each time t is obtained by the estimated atomic displacement $\boldsymbol{\Delta r}(t)$ from the initial structure $\boldsymbol{r}_0$:

$$\boldsymbol{r}(t) = \boldsymbol{\Delta r}(t) + \boldsymbol{r}_0.$$

In NMMD, the standard atomic positions $\boldsymbol{r}$ is modified to add an NMA-based atomic displacement to the MD-based atomic displacement $\boldsymbol{x}(t)$ from the initial atomic position $\boldsymbol{r}_0$, as follows:

$$\boldsymbol{r}(t) = \boldsymbol{q}(t) \cdot \boldsymbol{A} + \boldsymbol{x}(t) + \boldsymbol{r}_0, \quad \text{(III-1)}$$

where $\boldsymbol{q}(t) \cdot \boldsymbol{A}$ is the displacement of N atoms induced by a linear combination of M normal modes (given by matrix $\boldsymbol{A}$) with amplitudes $\boldsymbol{q}(t)$ at time t (M unknown parameters, $M << N$), $\boldsymbol{x}(t)$ is the atomic displacement at time t from classical MD simulation ($N \times 3$ unknown parameters). It can be noted that the total number of unknown parameters in NMMD ($\boldsymbol{q}(t)$ and $\boldsymbol{x}(t)$) at a simulation step t is $M + N \times 3$, where $M << N$.

NMMD integrates over time both types of parameters, $\boldsymbol{q}(t)$ and $\boldsymbol{x}(t)$, whereas classical MD simulation integrates $\boldsymbol{x}(t)$ only. For the numerical integration, NMMD uses the Velocity Verlet

integrator, which has good numerical stability and is commonly used in MD-based approaches. The integration of parameters $\boldsymbol{x}$(t) is given by :

$$\begin{aligned}\boldsymbol{x}(t+\Delta t) &= \boldsymbol{x}(t) + \dot{\boldsymbol{x}}(t)\Delta t + \frac{1}{2}\ddot{\boldsymbol{x}}(t)\Delta t^2 \\ \dot{\boldsymbol{x}}(t+\Delta t) &= \dot{\boldsymbol{x}}(t) + \frac{\ddot{\boldsymbol{x}}(t) + \ddot{\boldsymbol{x}}(t+\Delta t)}{2}\Delta t,\end{aligned}$$

and the integration of parameters $\boldsymbol{q}(t)$ is given by:

$$\begin{aligned}\boldsymbol{q}(t+\Delta t) &= \boldsymbol{q}(t) + \dot{\boldsymbol{q}}(t)\Delta t + \frac{1}{2}\ddot{\boldsymbol{q}}(t)\Delta t^2 \\ \dot{\boldsymbol{q}}(t+\Delta t) &= \dot{\boldsymbol{q}}(t) + \frac{\ddot{\boldsymbol{q}}(t) + \ddot{\boldsymbol{q}}(t+\Delta t)}{2}\Delta t,\end{aligned}$$

where $\dot{\boldsymbol{x}}(t)$ and $\ddot{\boldsymbol{x}}(t)$ respectively are the first and second derivatives of $\boldsymbol{x}(t)$ with respect to time, $\dot{\boldsymbol{q}}(t)$ and $\ddot{\boldsymbol{q}}(t)$ respectively are the first and second derivatives of $\boldsymbol{q}(t)$ with respect to time, and Δt is the time step of the numerical integration.

Recalling the Newton's second law of motion for $\boldsymbol{x}(t)$, $\ddot{\boldsymbol{x}}(t)$ is given by:

$$\ddot{\boldsymbol{x}}(t) = \boldsymbol{M}_x^{-1}\boldsymbol{F}_x(t),$$

where the force $\boldsymbol{F}_x(t)$ is the negative gradient of the potential energy $U(t)$ with respect to the atomic positions $\boldsymbol{x}(t)$ and $\boldsymbol{M}_x^{-1}$ is the inverse of the mass matrix $\boldsymbol{M}_x$ that is a diagonal matrix with the atomic masses m_x^i $(i = 1, \ldots, N)$ as the entries of the diagonal.

In NMMD, we consider that the atomic displacement in equation (III-1) due to normal modes, more precisely $\boldsymbol{q}(t)$ because $\boldsymbol{A}$ does not change with time in this equation, follows the motion equation:

$$\ddot{\boldsymbol{q}}(t) = \boldsymbol{M}_q^{-1}\boldsymbol{F}_q(t), \qquad \text{(III-2)}$$

where $\boldsymbol{F}_q(t)$ is the negative gradient of the potential energy $U(t)$ with respect to $\boldsymbol{q}(t)$ and $\boldsymbol{M}_q^{-1}$ is the inverse of the matrix $\boldsymbol{M}_q$ that is a diagonal matrix with $m_q^i = m_q$ $(i = 1, \ldots, M)$ as the entries of the diagonal. In this equation, m_q^i could be interpreted as the "mass" assigned to the i-th normal-mode amplitude $(i = 1, \ldots, M)$. Different values of such "masses" for different normal modes could be used for giving more or less weights to some of the normal modes: a high value of this parameter would result in a low contribution of the normal mode to the motion while its low value would result in a strong contribution of the normal mode. Here, we assigned the same value of this parameter for each normal mode (i.e. $m_q^i = m_q, i = 1, \ldots, M$), to avoid imposing any prior

information about the individual contribution of the different normal modes. During the simulation, NMMD automatically determines these contributions by estimating the amplitude of each normal mode. Therefore, equation (III-2) can be written as follows:

$$\ddot{\boldsymbol{q}}(t) \quad = \frac{1}{m_q}\boldsymbol{F}_q(t).$$

Regarding the choice of the mass value m_q to assign to all the normal modes, it should be noted that a too large value of m_q slows down the integration of the normal-mode amplitudes $\boldsymbol{q}(t)$, whereas its too small value makes the system unstable. This parameter can be tuned manually to maximize the speed while ensuring the stability of the system. For all the structures presented in the study, we have obtained satisfactory speed and stability results using $m_q = 10$. The diversity of these structures suggests that this value shall give satisfactory results in the majority of cases. Therefore, we propose to use $m_q = 10$.

While $\boldsymbol{F}_x(t)$ is implemented in the integration scheme of any standard MD simulation software, the computation of $\boldsymbol{F}_q(t)$ from scratch is not trivial as the potential energy is the sum of multiple potentials including the biasing potential. Besides, the gradient computation is usually highly optimized in MD software as it corresponds to the main computational consumption. In this context, it is convenient to see that $\boldsymbol{F}_q(t)$ can be expressed as a function of the force vector on MD-induced atomic coordinate displacement $\boldsymbol{F}_x(t)$ by using the following chain rule :

$$\begin{aligned}
\boldsymbol{F}_q(t) \quad &= -\frac{\partial U}{\partial \boldsymbol{q}}(t) \\
&= -\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{q}}(t) \cdot \frac{\partial U}{\partial \boldsymbol{r}}(t) \\
&= -\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{q}}(t) \cdot \left(\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{x}}(t)\right)^{-1} \cdot \frac{\partial U}{\partial \boldsymbol{x}}(t) \\
&= \boldsymbol{A} \cdot \boldsymbol{F}_x(t).
\end{aligned}$$

The NMMD method is implemented in GENESIS software and uses $\boldsymbol{F}_x(t)$ of GENESIS to calculate $\boldsymbol{F}_q(t)$. More precisely, NMMD was implemented in GENESIS 1.4 by modifying the code of the MD-based flexible fitting method implemented without normal modes [34, 36]. The flexible fitting in GENESIS minimizes the total potential consisting of the classical MD-based potential and a biasing EM-map potential, where the contribution of the biasing EM-map potential is defined by the force constant according to equation (II-7). We modified the MD-based fitting to include NMA-based atomic displacement. For NMA, we use ENM to compute the normal modes.

To speed up the calculation of normal modes of the reference atomic structure, we used the Rotation Translation Block (RTB) method [193] as implemented in ElNemo [194].

It should be noted that the matrix of normal modes $\boldsymbol{A}$ is constant over time in the equations above. Indeed, in our experiments, we have not found useful to recalculate normal modes during the fitting with NMMD. If one wants to recalculate normal modes, the simulation should be reinitialized using $\boldsymbol{r}_0$ (equation (III-1)), which should be set to be the best-fitting conformation from the previous run as well as using the corresponding normal-mode matrix of that new $\boldsymbol{r}_0$.

III.3. Results

In this section, I show the fitting performance of NMMD and compare these results with the results of fitting using MD simulation without NMA. The two methods were compared regarding computational time and conformational sampling in GENESIS 1.4 using REUS procedure. The simulations were performed on two Intel Xeon Silver 4214 CPUs (24 cores at 2.60 GHz per CPU) with 64 GB RAM, using a single core per replica.

The comparison of the two methods is shown using synthetic data sets of four complexes (LAO, AK, LF, and EF2) and experimental data sets of two complexes (p97 and ABC). The size of these complexes goes from small (26 kDa for LAO) to large (542 kDa for p97) (Table 1). Each synthetic data set consists of two atomic structures corresponding to two different molecular conformations. One of the two atomic structures was used as the initial conformation to fit the EM map simulated from the other atomic structure (target conformation). Each experimental data set contains one atomic structure and one EM map, which correspond to two different conformations and were used as the initial and target conformations for fitting, respectively. Additionally, each experimental data set contains one atomic model derived from the given EM map, which was used for comparison with the atomic models obtained by fitting with NMMD and MD. All atomic structures and cryo-EM maps used in this study are available in the Protein Data Bank (PDB) and Electron Microscopy Data Base (EMDB) databases. Their PDB and EMDB codes are provided in Table 1.

Table 1: PDB and EMDB codes of the available structural data used for the tests of NMMD fitting method and its comparison with MD fitting in GENESIS

Biomolecular complex	Initial PDB	Target PDB	Target EMDB	Weight (kDa)
LAO binding protein	1LST	2LAO		26
Adenylate kinase	4AKE	1AKE		47
Lactoferrin	1LFG	1LFH		77
Elongation factor 2	1N0V	1N0U		186
ABC exporter	6RAF	6RAH	4775	150

p97 ATPase	5FTM	5FTN	3299	542

NMMD and MD are compared using the following two measures: 1) CC between the target EM map and the map simulated from the fitted atomic conformation (normalized CC, with values between 0 and 1); and 2) Root Mean Square Deviation between the atomic positions in the target and fitted atomic conformations (RMSD, in angstroms). In the case of synthetic EM maps, the atomic structure used to simulate the target-conformation EM map is the target (ground-truth) atomic conformation that should be retrieved by fitting of an initial atomic conformation into the simulated EM map. In the case of experimental EM maps, such unique ground-truth atomic conformation is unavailable but existing fitted atomic models are available in the PDB and used here as the target atomic conformations for the RMSD calculations.

III.3.1. Synthetic EM maps of LAO, AK, LF, and EF2

To synthesize an EM map of a complex, we reproduced the reconstruction procedure used in single particle analysis. First, we obtained a high-resolution density map (map size: $100 \times 100 \times 100$ voxels, voxel size: $1.5 \times 1.5 \times 1.5$ Å) by converting a given atomic structure (target conformation for fitting) using atomic scattering factors (Peng et al. 1996). Then, a library of 2D projections of the obtained map was calculated using a quasi-uniform distribution of the projection directions on a sphere, determined by a tilt-angle step of 5°, which produced 1647 projection images. To simulate the effect of the electron microscope, we applied on the images a CTF with a set of parameters that simulate a 200 kV microscope and a defocus of -0.5 μm, and added Gaussian noise resulting in the SNR of 0.5, following the method of [49]. The SNR of 0.5 was chosen as the noise level typically encountered in decent-quality single particle class averages (considering 1647 projection images as the class averages). Finally, a synthetic EM map was reconstructed from the images (using Fourier interpolation method) and low-pass filtered to 5 Å. Such procedure for EM map synthesis and the fact that NMMD and MD fitting methods use a different method for converting atoms into density (3D Gaussian functions to obtain density volumes at each iteration of the fitting) make the fitting more difficult. All steps of the procedure were performed using Xmipp 3 command-line programs [54].

III.3.2. Experimental EM maps of p97 and ABC

For ABC, the fitting was performed using the following two conformations: 1) extracellular-side closed (intracellular-side open) conformation, given by PDB:6RAF (an atomic model derived from a cryo-EM map of this conformation); and 2) extracellular-side open (intracellular-side closed) conformation, given by EMD-4775 (a cryo-EM map of this conformation at 2.8

Å resolution). For p97, the fitting was performed using the following two conformations: 1) conformation with N-terminal domains in "down" position, given by PDB:5FTM (an atomic model derived from a cryo-EM map of this conformation); and 2) conformation with N-terminal domains in "up" position, given by EMD-3299 (a cryo-EM map of this conformation at 3.3 Å resolution). Both cryo-EM maps used as the target conformations for fitting were down-sampled to the size of $128 \times 128 \times 128$ voxels to speed up the computation.

In our preliminary experiments, we identified mode 10 of the 5FTM p97 structure as the mode that moves up and down the N domains of all six monomers while preserving the overall p97 symmetry. Our preliminary experiments with ABC showed that mode 9 of the 6RAF ABC structure is one of the modes that contribute to opening and closing motions of ABC. A description of complex conformational transitions, such as those in the data used in this work, requires using more than one normal mode. We show below that NMMD achieves good fitting results for these two molecular complexes using a small set of normal modes (ten lowest-frequency normal modes, i.e., modes 7-17) in combination with MD.

III.3.2.1. *Parameters for running NMMD and MD fitting methods*

For each of the six molecular complexes (Table 1), we ran 16 replicas of both NMMD and MD fitting methods with a time step of 2 femtoseconds. The value of the force constant for each complex was adjusted with REUS, from a linear distribution of values in the range determined in preliminary experiments with each complex, as the optimal range is specific to each system (5000-10000 kcal/mol for AK and LAO, 10000-30000 kcal/mol for LF, EF2 and ABC, 50000-100000 kcal/mol for p97). For all complexes and both fitting methods, all-atom simulations including hydrogen atoms were performed using CHARMM 36 force fields and the temperature of 300 K regulated with the Langevin thermostat with friction coefficient of 1 ps^{-1}. MD and NMMD do not search for rotations and translations during the fitting, meaning that the atomic structure and the EM map must be aligned using rigid-body transformations prior to fitting. The rigid-body transformations were performed using ChimeraX function "Fit in map" [195]. To simulate the map from the fitted atomic structure using Gaussian functions, the standard deviation of the Gaussian functions was set to 2.5, resulting in a map of 5 Å resolution (the simulated map to be compared with the given map using the CC). NMMD fitting was run using normal modes 7-17, which are the ten lowest-frequency non-rigid-body modes that usually describe collective motions. Normal modes were calculated using the RTB block size of 10 residues and the ENM with the interaction cutoff radius of 8 Å (the cutoff defining the radius beyond which the nodes of the ENM are not connected with elastic springs).

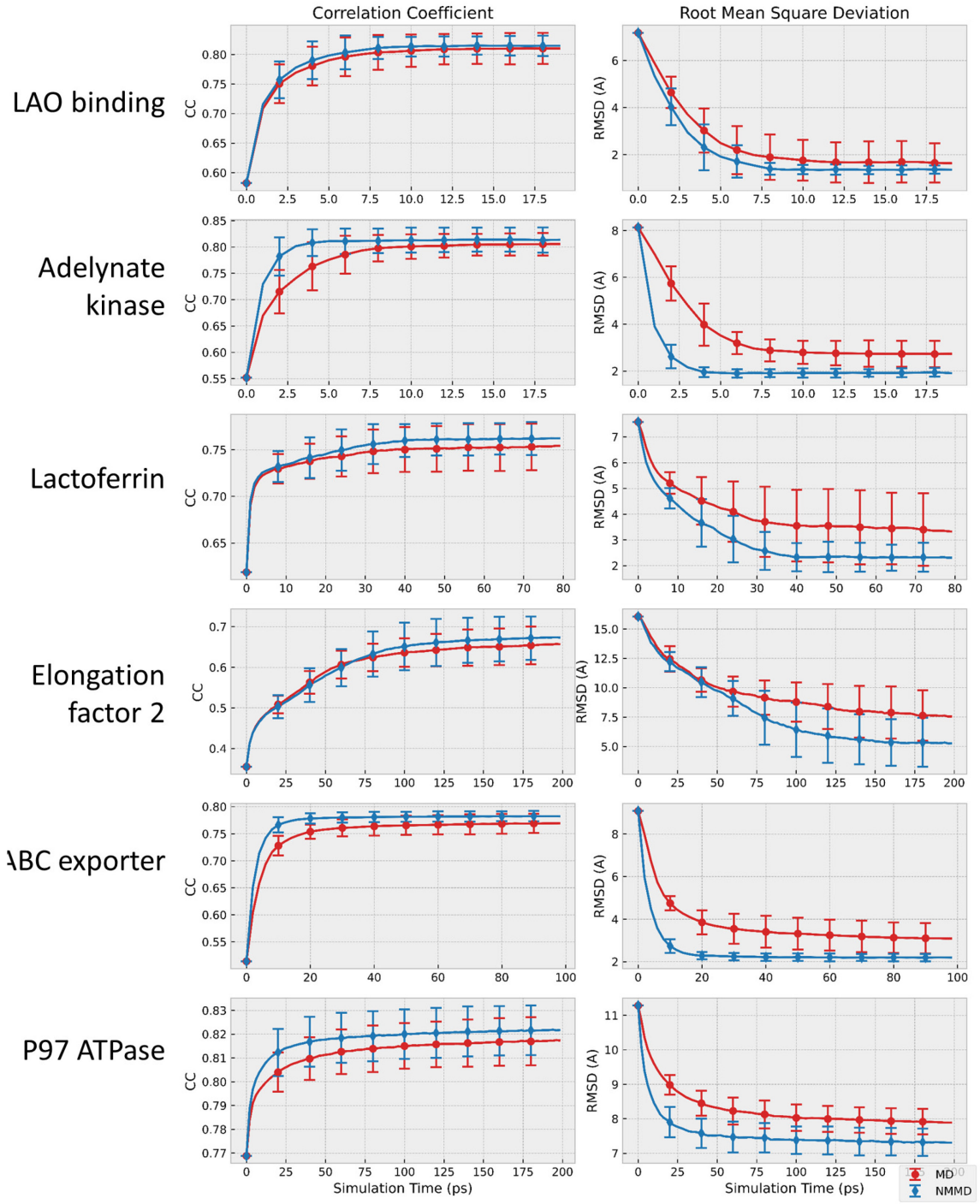

Figure 9 : Mean (curve) and standard deviation (error bar) of CC (left panels) and of RMSD (right panels) as a function of simulation time, for 16 replicas of NMMD fitting (blue) and the MD fitting (red). The results are shown for four synthetic data sets (LAO binding protein, Adenylate kinase, Lactoferrin, and Elongation factor 2) and two experimental data sets (ABC exporter and p97 ATPase). See Table 2 for additional information regarding the fitting results.

III.3.2.2. *Inclusion of normal modes speeds up fitting*

For each of the six molecular complexes (Table 1), the mean and the standard deviation of the CC and the RMSD over all replicas of NMMD and MD are plotted in Figure 9. For the replica that reached the lowest RMSD, Table 2 shows the achieved values of RMSD and CC, the total execution time (in the case of NMMD, the total time also includes the time required for computing normal modes), the convergence time (the time until the RMSD starts to change by less than 1% between the successive steps), and the measures of the obtained atomic structure quality (MolProbity score).

Table 2: Comparison of NMMD and MD fitting results for each of four synthetic data sets (LAO binding protein, Adenylate kinase, Lactoferrin, and Elongation factor 2) and two experimental data sets (ABC exporter and p97 ATPase). The table shows the achieved CC and RMSD values for the replica with the lowest achieved value of RMSD (from 16 replicas), the total time of execution (for NMMD, this time includes the time required for computing normal modes), the convergence time (the time until the RMSD starts to change by less than 1% between the successive steps), the speed increase between NMMD and MD convergence time in percentage, and the measure of the obtained atomic structure quality (MolProbity score). See Figure 9 for the CC and RMSD plots over the simulation for all 16 replicas

Biomolecular complex	Fitting method	CC	RMSD (Å)	Total time (min)	Convergence time (min)	Speed increase	MolProbity score
LAO	MD	0.83	1.04	10.5	6.2		2.84
	NMMD	**0.83**	**1.06**	**10.6**	**4.2**	**32%**	**2.79**
Adenylate kinase	MD	0.83	1.88	9.8	5.3		2.61
	NMMD	**0.84**	**1.58**	**10**	**2**	**62%**	**2.59**
Lactoferrin	MD	0.79	1.57	132.2	122.2		2.33
	NMMD	**0.79**	**1.45**	**133.2**	**64.9**	**46%**	**2.37**
Elongation factor 2	MD	0.73	3.67	402.5	358.2		2.67
	NMMD	**0.75**	**2.13**	**405**	**322.6**	**10%**	**2.49**
ABC exporter	MD	0.79	2.18	229.9	179.3		2.21
	NMMD	**0.79**	**1.86**	**235.3**	**61.1**	**66%**	**2.21**
p97 ATPase	MD	0.83	7.17	1754.8	1526.7		2
	NMMD	**0.81**	**6.57**	**1796**	**1257.2**	**17%**	**1.99**

One can note that, generally, the CC increases (the RMSD decreases) faster with NMMD than with MD (Figure 9). The faster convergence of NMMD can be explained by the use of normal modes, as this is the main difference between the two fitting methods. Therefore, the addition of normal modes to MD-based fitting improves the sampling efficiency. The computation and integration of normal modes induces a small computational cost, included in the total execution time reported in Table 2. This computational cost is insignificant compared to the speed increase

induced by the inclusion of normal modes. Indeed, the convergence time is, in average, around 40% shorter for NMMD than for MD (Table 2). Additionally, it should be noted that NMMD generally reaches lower RMSD values than MD (in all tested cases of molecules except for LAO, where the two methods reached almost the same value of RMSD, Table 2). In the case of EF2, the speed increase is less important than for the other structures, but NMMD reaches a much lower RMSD value (2.13 Å) than MD (3.67 Å), probably thanks to this extra time compared to other structures.

III.3.2.3. *Inclusion of normal modes improves accuracy of fitting*

We observe (Table 2) that NMMD and MD retrieved the target structure (relatively low RMSD) for five out of six complexes (for all except for the experimental data of p97). As noted earlier, NMMD achieves a lower RMSD value than MD in all cases except in the case of LAO, where the achieved RMSD is almost the same for the two methods. In some cases, the achieved RMSD values differ more between the two methods (e.g, in the synthetic EF2 case and the experimental p97 case, Table 2).

The best achieved RMSD for the synthetic case of EF2 is 3.67 Å for MD and 2.13 Å for NMMD. The target and fitted conformations are different in the case of MD but very similar in the case of NMMD, meaning that NMMD produced better fitting than MD. For a visual assessment of the obtained conformations, we present the results of EF2 fitting in Figure 10. NMMD retrieved the right conformation (Figure 10c) while MD was able to retrieve the global target shape but not the details of the structure (some secondary structure elements are different or missing, Figure 10b). To assess the quality of the obtained atomic models, we evaluated their MolProbity scores [196] (Table 2). The MolProbity score is commonly used to assess quality of structures and corresponds to a combination of the number of atomic clashes, rotamer evaluations, and Ramachandran evaluations. In the case of EF2, the MolProbity score is significantly lower for NMMD than for MD, indicating that the conformation obtained with NMMD has better quality than the one obtained with MD. Generally, the quality of the atomic models obtained with NMMD is comparable to the quality of the models obtained with MD and, in some cases, NMMD produces models of better quality (see MolProbity score columns of Table 2).

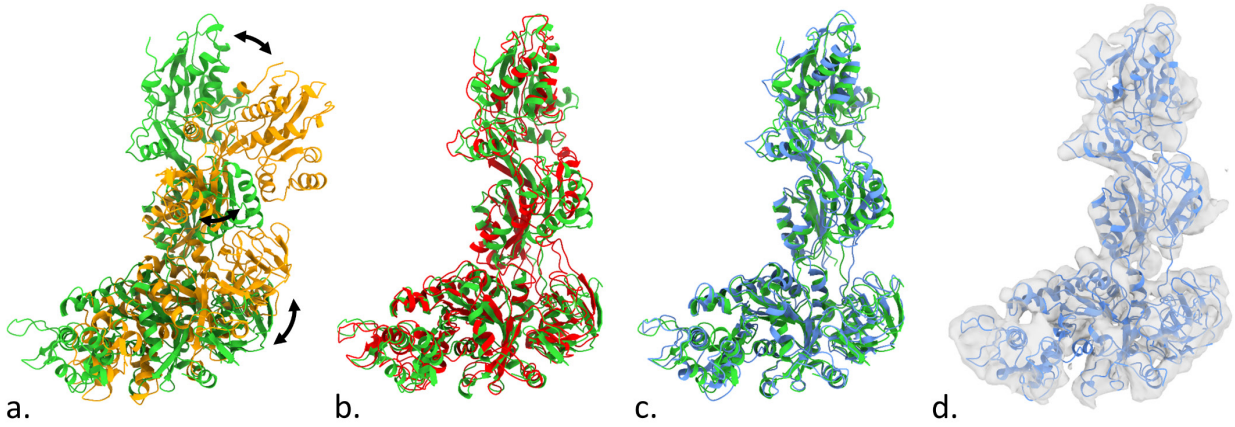

Figure 10 : Flexible fitting of a synthetic EM map of Elongation factor 2 for the NMMD and MD replicas reaching the lowest RMSD value. Target structure (green) is overlapped with the initial structure (a), MD fitted structure (b), NMMD fitted structure (c). NMMD-fitted structure is overlapped with the target EM map (d). The black arrows show the main conformational changes.

Figure 11 shows the results of fitting in the case of ABC experimental data. Despite the large conformational rearrangements of ABC, NMMD successfully retrieved the target conformation (RMSD = 1.86 Å), with less clashes than MD (Table 2). The MD achieved a slightly worse RMSD than NMMD (The difference between the two RMSDs is around 0.3) (Table 2).

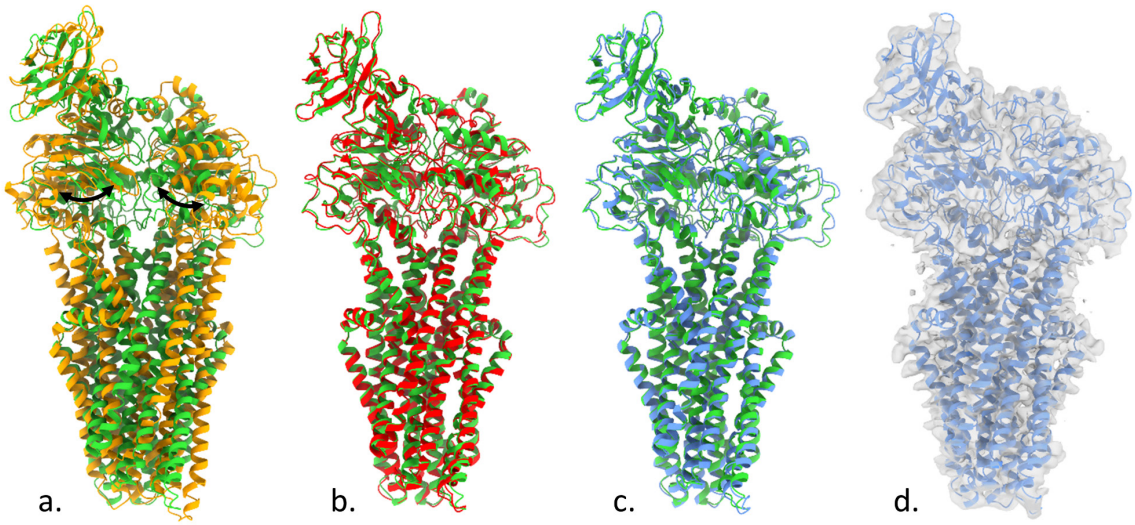

Figure 11 Flexible fitting of an experimental EM map of ABC exporter for the NMMD and MD replicas reaching the lowest RMSD value. Target structure (green) is overlapped with the initial structure (a), MD fitted structure (b), NMMD fitted structure (c). NMMD-fitted structure is overlapped with the target EM map (d). The black arrows show the main conformational changes.

Concerning the experimental p97 case, we observe that the achieved RMSD is high for both methods (RMSD > 6 Å) but still smaller for NMMD than for MD. These results indicate that the fitted structures are largely different from the target structure. Figure 12 presents the p97 fitting

results. We observe that both MD and NMMD were able to lift the N domain up (the N-domain motion is shown by a black arrow in Figure 12a) and to fit the protein globally. However, the final conformations obtained by both methods are different from the target conformation. Figure 12 shows significant conformational differences in the N-terminal domain of these three structures (Figure 12b,c).

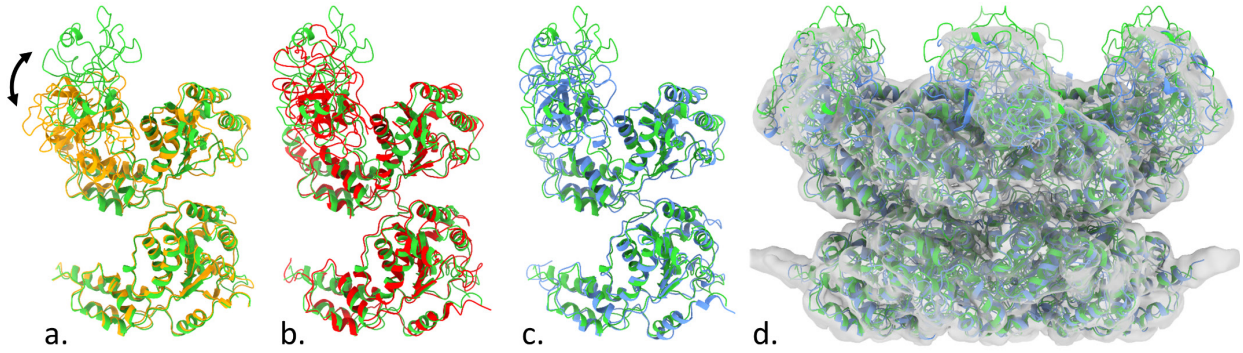

Figure 12 Flexible fitting of an experimental EM map of p97 ATPase for the NMMD and MD replicas reaching the lowest RMSD value (a single monomer is shown for better visibility in a-c). Target structure (green) of one monomer is overlapped with the initial structure (a), MD fitted structure (b), and NMMD fitted structure (c). Target (green) and NMMD-fitted (blue) structures of all six monomers are overlapped with the target EM map (d). The black arrows show the main conformational change.

To assess and compare the local resolutions of the two fitted experimental cryo-EM maps (p97 and ABC), we used MonoRes [197], an automatic method to determine the local resolution of an EM map (Figure 13). We can note that the ABC map has a similar local resolution for the whole structure, which is around 3 Å (Figure 13b). On the contrary, the p97 map has a much lower local resolution in the regions that correspond to the N-terminal domains (top parts of the map) than in the other regions (Figure 13a). More precisely, the local resolution in the major part of the p97 map is around 3 Å, but it is between 7 and 15 Å in the regions corresponding to the N domains. The differences in local resolutions of the cryo-EM maps of ABC and p97, especially for the N domains of p97, may explain the differences in the fitting results obtained for these two test cases.

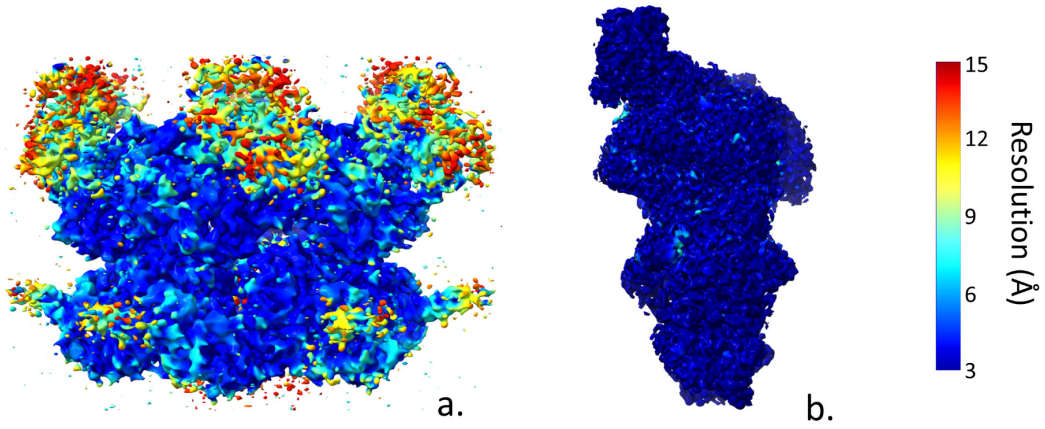

Figure 13 Local resolution obtained with MonoRes for the experimental cryo-EM maps fitted with MD and NMMD in this article. (a) Local resolution of p97 ATPase. (b) Local resolution of ABC exporter.

III.3.3. Inclusion of normal modes improves REUS adjustment of the force constant

The atomic models of the synthetic and experimental EM maps obtained by 16 replicas of NMMD and MD were analyzed by Principal Component Analysis (PCA). Figure 14 shows low-dimensional (2D) conformational spaces of LF, EF2, ABC and p97 (two synthetic and two experimental data cases), determined by the first two principal axes. In the case of LF, EF2 and ABC, the NMMD replicas (blue dots) are more concentrated around the target conformation (green dot), whereas the MD replica (red dots) are more scattered. This result indicates that REUS force constant adjustment produces more consistent fitting results with NMMD than with MD. In the case of p97, both MD and NMMD replicas are spread between the initial and target conformations, which is consistent with our previous observation that both approaches converged to a conformation that is different from the target conformation.

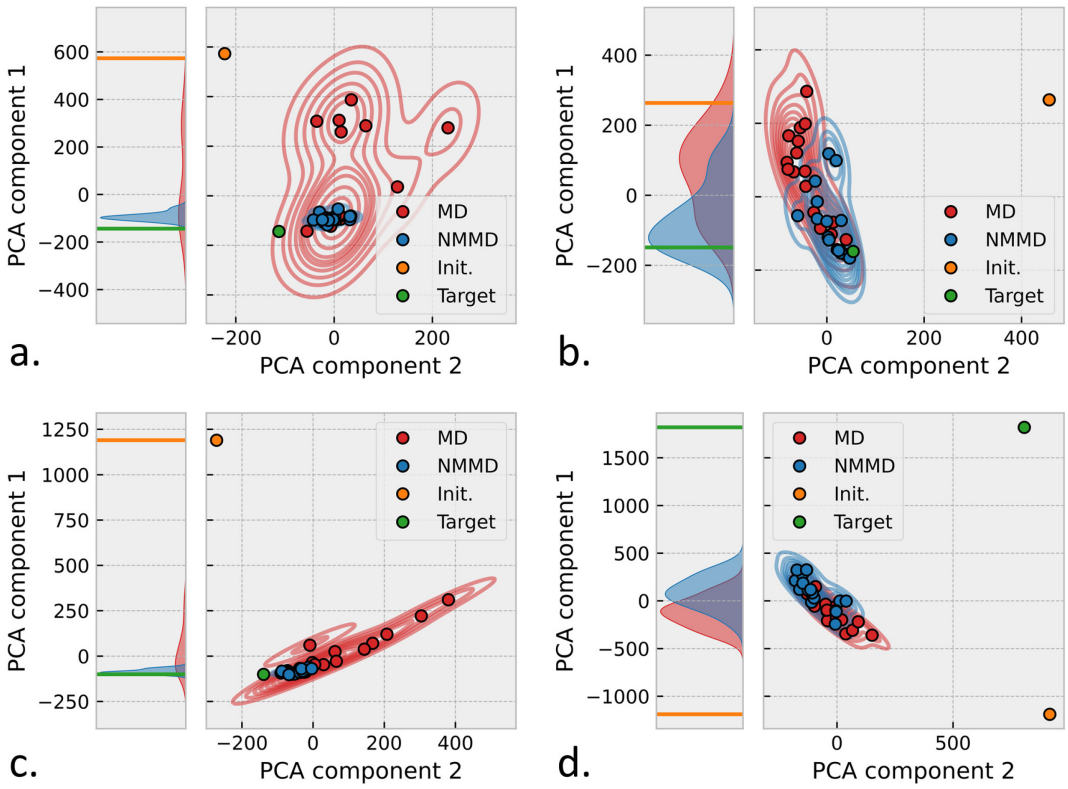

Figure 14 Two-dimensional conformational spaces determined by PCA of the atomic models from 16 replicas of MD (red) and NMMD (blue), together with the target structure (green) and the initial structure (orange), for Lactoferrin (a), Elongation factor 2 (b), ABC exporter (c), and p97 ATPase (d). The one-dimensional plots at the left side show the data distribution along the first PCA axis.

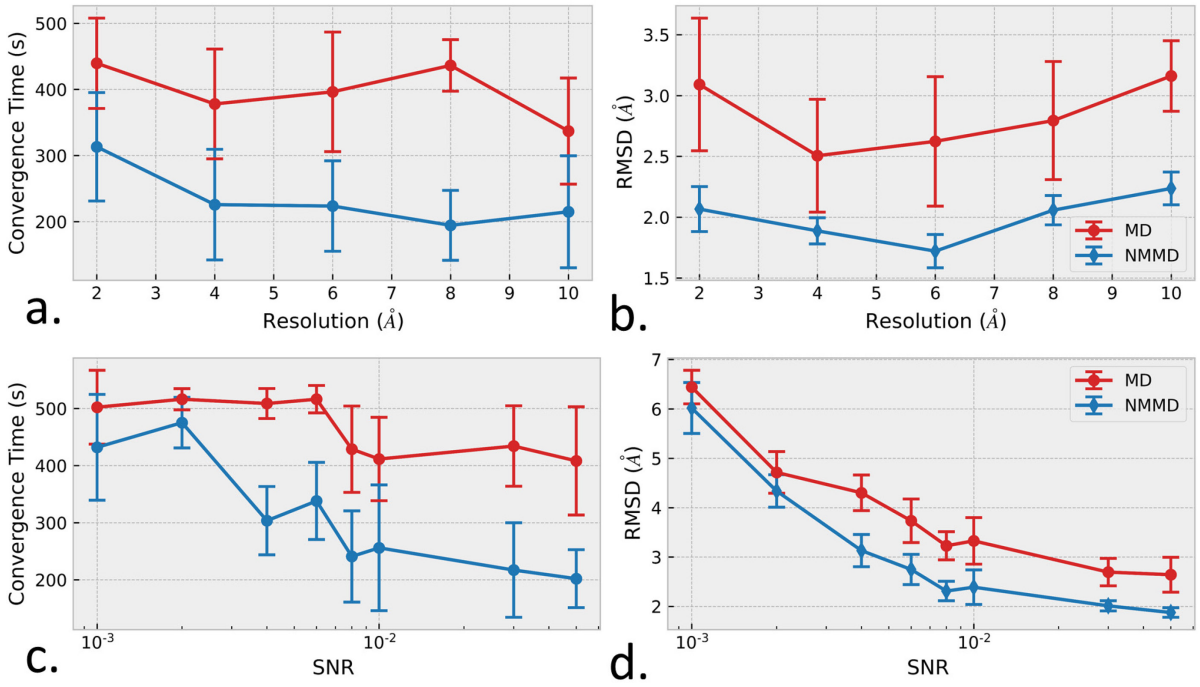

Figure 15: Impact of noise and map resolution on the results of fitting with MD (red curves) and with NMMD (blue curves). (a,c) Evolution of the convergence time with respect to the resolution (a) and the SNR (c). (b,d) Evolution of the minimum RMSD with respect to the resolution (b) and the SNR (d).

III.3.4. Impact of noise and map resolution

Additional tests of NMMD and MD sensitivity were performed with synthetic EM maps to study the impact of noise and map resolution. To this end, we performed several fitting processes on AK by varying the SNR of the images used for 3D reconstruction and the resolution of a synthetic map of AK. For the study of the noise impact, we employed the same process to synthesize the maps as the one described in section III.3.1 with the SNR of images in the range from 0.5 to 0.001. Note that it is almost impossible to see the particle in the synthetic map for the SNR of 0.001. For the study of the resolution impact, we applied a low-pass filter onto the high-resolution synthetic map obtained by converting the AK atomic structure into density as described in section III.3.1. The cut-off frequency of the low-pass filter was varied from 2 to 10 Å, with a step of 2 Å. For each of the resulting maps, we run ten NMMD and ten MD fitting processes with a force constant of 10000 kcal/mol (ten runs with random initial velocities). The average resulting convergence time and minimum RMSD are shown in Figure 15a,b for the resolution variation, respectively, and in Figure 15c,d for the SNR variation, respectively.

Figure 15a,b shows similar RMSD and convergence time values over the tested resolution range for each of the two methods. Figure 15c,d shows that the RMSD and the convergence time tend to increase with the decrease in the SNR. Also, Figure 15 shows that both RMSD and convergence

time values are lower for NMMD than for MD for the entire tested resolution and SNR ranges, meaning that NMMD is faster, more accurate, and has a better robustness to noise than MD over the entire tested resolution and SNR ranges.

III.4. Discussion and conclusion

In this chapter, I described NMMD, a new flexible fitting method for cryo-EM, which combines NMA and MD simulations. Given an atomic structure and a cryo-EM map to fit, NMMD simultaneously estimates global atomic displacements based on NMA and local atomic displacements based on MD simulation, by their simultaneous numerical integration.

In the tests described here, I compared NMMD with MD simulation fitting using a variety of maps (synthetic and experimental EM maps of six molecular complexes, with different noise levels and resolutions). These tests showed that the combination of MD simulation and NMA has generally better performance than MD simulation alone. Both NMMD and MD simulations were run with REUS sampling procedure to automatically adjust the value of the force constant.

The results showed that adding normal modes to MD-based fitting makes the fitting faster by around 40% (in average). Also, the results showed that adding normal modes to MD-based fitting improves the fitting accuracy (the obtained RMSD values were lower for NMMD than for MD in all cases except in one where the achieved RMSD value was almost the same for the two methods).

In contrast to an NMA-based fitting, which generally induces structural distortions, NMMD produces biochemically meaningful structures whose quality can, in some cases, even be slightly better than the quality of the structures obtained with an MD-based fitting (observed in this study using MolProbity scores). In this context, a simultaneous use of NMA and MD is an advantage of NMMD over a two-step approach in which an NMA-based fitting is followed by an MD-based fitting. In general, such two-step strategies are less likely to produce good quality structures, as the NMA-based fitting may create undesirable distortions of the structure that are impossible to correct with the MD-based flexible fitting.

Interestingly enough, adding normal modes to MD-based fitting improves the selection of the value of the force constant by REUS. Our NMMD approach could be used in a coarse-to-fine multiresolution scheme, which could further enhance the REUS-based sampling strategy. Such schemes do not necessarily reduce the computing time, but increase robustness against noise and minimize the risk of getting trapped into local minima. A multiresolution scheme could be to generate different resolutions of the same EM map by its low-pass filtering or down-sampling and, then, run NMMD to perform the EM map fitting at the different resolutions by propagating and

refining the solution from a coarser resolution level to the next finer level, to fit the map first globally than locally, in each of the different replicas.

In this study, no prior information was imposed on the contribution of normal modes, by using the same ”mass” value (m_q) for all the normal modes. Also, the value of this parameter was manually tuned to $m_q = 10$, which was the value that ensured good speed and stability for all the experiments presented here. Based on the diversity of the structures used in our experiments, it can be expected that this value gives good results in most cases. However, finding the optimal value of this parameter for each particular case may increase the NMMD efficiency further. Future developments could include an automated optimization of this parameter, for instance through REUS-based parameter estimation during the fitting.

NMMD uses the CC to measure the similarity between the EM map and the map simulated from the atomic structure. The lower the EM-map resolution (global and local), the higher the uncertainty of the EM fitting with an atomic structure will be, independently of the similarity measure used. Multiple local minima in the potential (the potential that guides the fitting based on the chosen similarity measure) can lead the fitting to a local-minimum conformation that is different from the global-minimum conformation. We addressed this issue by running multiple parallel replicas with REUS. The distribution of the final values of the similarity measure (here, CC) obtained by the replicas (Figure 9) is informative of the fitting uncertainty: large CC variance suggests high uncertainty (different conformations obtained by different replicas), whereas small CC variance suggests low uncertainty (similar conformations obtained by different replicas). Beside the use of the CC variance over the different replicas, the uncertainty of the fit could be evaluated by observing the distances between the structures from the different replicas in a PCA-based low-dimensional space (Figure 14). Another way to visualize distances between the structures from different replicas (in a low-dimensional space) could be to perform a multivariate analysis of a matrix of pairwise RMSD-based distances between these structures.

The GENESIS software includes an efficient parallelization strategy for MD-based fitting, based on a combination of MPI and OpenMP, which is well suited for computation over multiple nodes. NMMD in GENESIS offers the same parallelization strategy as the MD-based fitting. In this study, we decided not to use this parallelization scheme, but a parallelization over replicas, using a single core per replica. However, for more time demanding fitting tasks, one could use NMMD with a larger number of cores per replica.

NMMD software code is publicly available as part of ContinuousFlex plugin [198] (https://github.com/scipion-em/scipion-em-continuousflex) for Scipion [55], including a graphical user interface giving the user the opportunity to easily use NMMD on parallel systems.

Chapter IV. MDSPACE: a NMMD-based method developed for analyzing continuous conformational variability in single particle image

This chapter presents MDSPACE, a novel approach for analyzing continuous conformational variability in single particle images, which is based on NMMD flexible fitting of a given atomic model (the so-called reference model, which represents the initial conformation for the fitting) into individual particle images. MDSPACE was published in 2023 [25]and the Methods, Results, and Discussion-Conclusion sections of this chapter were extracted from the published manuscript [25] and adapted for inclusion in this PhD thesis manuscript. Section IV.3.3 is novel and is currently being written as a manuscript in which I will be co-first author.

IV.1. Introduction

The gain in speed obtained by NMMD (described in Chapter III) opens the door for MD-based analysis of large datasets such as heterogeneous sets of single particle images. However, the analysis of individual single particle images is still challenging due to the low SNR of the images. Also, contrary to the fitting of 3D EM maps, the particles are 2D images which therefore requires an additional 3D-to-2D projection step.

A first version of a method for flexible fitting of a given particle image using MD simulation of an initial atomic model was developed in GENESIS by Alex Mirzaei, a postdoc in my lab at Sorbonne University. This first method was analyzing the conformation only in one image, using solely MD simulation. Moreover, this code was slow and not optimized for parallel computing. In particular, the calculation of the biased potential energy and its gradient involved calculating several 2D projections of the 3D atomic model per MD iteration using full-length Gaussian functions at atomic centers to simulate the density. I entirely rewrote the code with new features allowing to analyze large set of particle images in parallel and considering their pre-determined particle pose (e.g. using SPA), which is further refined during the fitting. The new implementation also includes a fast and parallelized calculation of the biased potential energy and gradient, which truncate the Gaussian functions above a certain threshold, greatly decreasing the computational cost.

The developed method for 3D-to-2D flexible fitting could now be applied on large single particle datasets and produce conformational landscapes using dimensionality reduction methods

(e.g. PCA, UMAP, see section II.3.1) on the fitted atomic models. However, this 3D-to-2D flexible fitting method alone revealed to be suboptimal for accurate extraction of conformational landscapes due to a difficulty to individually fit some of the particle views. To tackle this issue, I developed an iterative approach that extracts the information about the principal motion directions from the ensemble of the fitted conformations (obtained for a set of particle images through individual 3D-to-2D flexible fitting of each image) and, then, uses this information for the next round of the individual 3D-to-2D flexible fitting by replacing normal modes by the principal motion directions obtained in the previous round. The iterative refinement of the principal motion directions extracted from the ensemble of the fitted conformations helps to iteratively refine the individual 3D-to-2D flexible fitting of the particle images, which in turn results in the refinement of the entire conformational landscape.

The new iterative method, referred to as MDSPACE (stands for Molecular Dynamics simulation for Single Particle Analysis of Continuous Conformational hEterogeneity) was published in [25] and allows to obtain conformational landscapes from highly heterogeneous image datasets containing continuous conformational heterogeneity, by analyzing individual low-SNR single particle images. MDSPACE requires a single input atomic structure and an initial particle pose, which are used for fitting each particle image. The initial particle pose is refined at each iteration of the process. MDSPACE was tested with a synthetic dataset of the heterodimeric ABC exporter TmrAB [4], an experimental dataset of yeast 80S ribosome-tRNA complexes from EMPIAR-10016 [89], and an experimental dataset of AAA ATPase p97. As shown in the experiments, MDSPACE produces continuous conformational landscapes connecting known metastable states with identified intermediate transition states accurately, even when the initial conformation is distant from the target conformations in the particle images. Furthermore, MDSPACE produces atomic-scale conformational landscapes and energy landscape of any dimension (1, 2, 3, etc.), from which conformational trajectories can be interpreted by animations on the atomic models or on the density maps reconstructed from the data.

MDSPACE was integrated into ContinuousFlex plugin [198] of Scipion [55]. The Scipion environment allows for an efficient input and output data management and includes i) a graphical user interface, which facilitates the use of methods and the reproduction of experiments and results; ii) the availability of many image processing methods, which have been developed over many years for single particle analysis such as Xmipp [54] and Relion [12]; and iii) parallel data processing on clusters and supercomputers.

The MDSPACE method is described in section IV.2. Section IV.3 shows the performance of the new method with synthetic data and experimental data. A discussion and conclusion are provided in section IV.4.

IV.2. Methods

IV.2.1. MDSPACE

MDSPACE is an iterative approach (Figure 16) for analyzing continuous conformational variability in single particle images based on MD simulation. The MD simulation is guided by a 2D biasing potential. MDSPACE iteratively refines an initial conformation of the particle in each image. The initial conformation is the same for all particle images and determined by a given atomic structure. Also, MDSPACE iteratively refines an initial particle orientation and position in each image (the initial rigid-body alignment obtained by classical approaches based on classification and projection matching or by other continuous conformational variability approaches prior to using MDSPACE). In the first iteration, a provided atomic structure is flexibly fitted to each particle image, independently of other images, using an original 3D-to-2D flexible fitting approach based on normal-mode empowered MD simulation (indicated as "NMMD step" in Figure 16) and the given initial particle orientation and position in the image. The obtained ensemble of fitted atomic models is then rigid-body aligned to the initial atomic structure and PCA is performed on the rigid-body aligned models. The principal components obtained by PCA represent the dominant conformational changes extracted from the ensemble of fitted atomic models in one iteration of MDSPACE. In the next iteration of MDSPACE, the previously extracted principal component vectors replace the normal mode vectors used in the "NMMD step", to incorporate the ensemble information in the fitting of individual particles (principal-component empowered MD simulation indicated as "PCMD step" in Figure 16s). Each new iteration of MDSPACE (involving a new round of 3D-to-2D flexible fitting) is based on principal-component empowered MD simulations that encourage the conformation to move along the principal component vectors representing the dominant motions, with the effect of refining the fitting. Rigid-body movements that occurred during the MD-based flexible fitting are measured at the end of each MDSPACE iteration and are then used to refine the parameters of the initial rigid-body alignment of the particle images for the next MDSPACE iteration. In each MDSPACE iteration, it is possible to visualize the PCA space (determined by the first few principal axes), individual atomic models, movies of atomic-model displacements along different directions in this space, or 3D reconstructions from the groups of images with similar particle conformations (close points in the PCA space) along these directions. These directions can be the principal axes or can be

determined by a path traversing the densest regions in this space. The information obtained from the PCA space can be compared between different iterations.

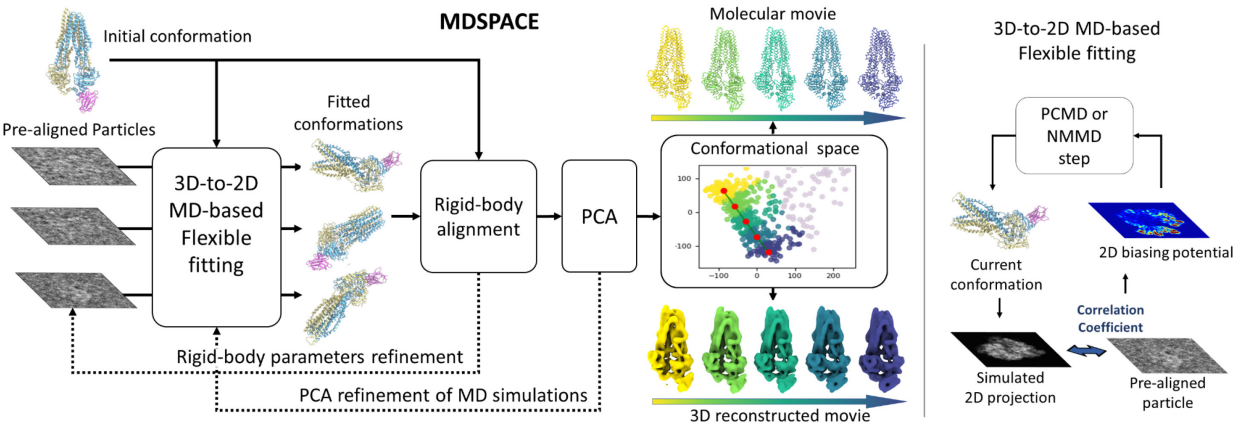

Figure 16 : Flowchart of the MDSPACE method (left) proposed for iterative continuous conformational analysis of single particle images, which is based on a 3D-to-2D flexible fitting approach (right) that can use normal-mode empowered MD simulation (indicated as "NMMD step" in this figure) or principal-component empowered MD simulation (indicated as "PCMD step" in this figure). The MD simulation is guided by a 2D biasing potential (right). The dotted lines represent the iterative process, which may be repeated several times to refine the conformational space.

IV.2.2. 3D-to-2D flexible fitting using MD simulations

MDSPACE perform the flexible fitting with biased MD (see section II.4.1) by using a CC-driven biasing potential (equation (II-9)), similar to standard MD-based flexible fitting approach that is normally used to fit 3D volumes [34]. However, instead of fitting a 3D volume, MDSPACE fits 2D images. Therefore, the CC is computed in 2D (equation (II-10)). To compute the CC between a 2D projection of a 3D atomic model (simulated image) and a given 2D image, the model is converted into density using Gaussian functions at the position of each atoms modified to account for i) the projection direction that is determined by a rotation matrix $\boldsymbol{R}$; and ii) for the shift that is determined by a translation vector $\boldsymbol{T}$, as follows:

$$\boldsymbol{\rho}_{sim}^{i,j}(\boldsymbol{r}) = \frac{1}{2\pi\sigma^2}\sum_{n=1}^{N}\exp\left(-\frac{1}{2\sigma^2}\left((i-r_n^x)^2+\left(j-r_n^y\right)^2\right)\right),$$
$$\begin{pmatrix} r_n^x \\ r_n^y \\ r_n^z \end{pmatrix} = \boldsymbol{R}^{-1}\ \boldsymbol{r}_n - \boldsymbol{T}, \qquad \text{(IV-1)}$$

where σ is the standard deviation of the Gaussian functions, N is the number of atoms in the model with the coordinates and r_n^x and r_n^y are the x and y coordinates of the n-th atom (see equation

(II-1)) rotated and translated with a 3 × 3 rotation matrix $\boldsymbol{R}$ and a 3 × 1 translation vector $\boldsymbol{T}$. Note that the projection depends on the pose ($\boldsymbol{R}$, $\boldsymbol{T}$) and on the provided atomic coordinates $\boldsymbol{r}$.

The biasing force of such 2D biasing potential is obtained by calculating the gradient of the CC (equation (II-11)) which involve calculating the gradient of the simulated projection with respect to the atomic coordinates $\boldsymbol{r}_n$ which can be written in the following form:

$$\frac{\partial \boldsymbol{\rho}_{sim}^{i,j}(\boldsymbol{r})}{\partial\, \boldsymbol{r}_n} = \frac{1}{\sigma^2}\boldsymbol{\rho}_{sim}^{i,j}(\,\boldsymbol{r}_n)\boldsymbol{R}\begin{pmatrix} i - r_x^n \\ j - r_y^n \\ 0 \end{pmatrix}.$$

IV.2.3. 3D-to-2D flexible fitting using normal-mode empowered MD simulations

In the first iteration of MDSPACE, 3D-to-2D flexible fitting is performed using the NMMD algorithm (simultaneous integration of the global displacement along normal modes and the MD-based local displacement), which accelerates flexible fitting of a given initial atomic conformation to target conformations in cryo-EM single particle images (the refinement of the atomic conformation against each particle image). In the next iterations of MDSPACE, the normal modes in NMMD are replaced by the principal component vectors that are extracted from the conformational ensemble obtained at the previous MDSPACE iteration, which iteratively refines the conformation and rigid-body alignment parameters of the particle in each image, as described in the subsection IV.2.6.

IV.2.4. Initial orientation and position of the particles by rigid-body pre-alignment

Although rigid-body motions are allowed in MD simulations, it is preferable for the initial conformation to be pre-aligned with the particle images in order to prevent the simulation to get trapped into local minima. The accuracy of the pre-alignment depends on a multitude of factors. For instance, as in the case considered here, where large continuous conformational heterogeneity is present in the particle images and rigid-body alignment methods are used for the pre-alignment (e.g., based on projection matching between the initial conformation and the particle images), the pre-alignment errors will be larger. Here, we show that MDSPACE can refine the initial rigid-body alignment. We show that a coarse rigid-body pre-alignment is sufficient to guide the MD simulation in the right direction, as rigid-body alignments are further performed during the MD-based 3D-to-2D flexible fitting (fine rigid-body movements of the atomic model during MD simulation to finely match the particle in the image), as explained in the next subsection.

IV.2.5. Refinement of the initial orientation and position of the particles

The MDSPACE method assumes that a set of rotations $\boldsymbol{R}_1$ and translations $\boldsymbol{T}_1$ (corresponding to $\boldsymbol{R}$ and $\boldsymbol{T}$ in equation (IV-1)) have been obtained by pre-alignment of the initial conformation with the particle images (regardless of the method used for this pre-alignment). During each MDSPACE iteration and for each particle, the atomic structure undergoes rigid body transformations, driven by the MD-biasing potential towards the target particle conformation and position in the image. Let the unknown rigid-body transformation that occurred during the biased MD simulation be $\boldsymbol{R}_2$ and $\boldsymbol{T}_2$. If this transformation is estimated, it allows to refine the initial transformation ($\boldsymbol{R}_1$ and $\boldsymbol{T}_1$) to obtain a refined set of rotation and translation ($\boldsymbol{R}_3$ and $\boldsymbol{T}_3$, respectively):

$$\begin{aligned} \boldsymbol{R}_3 &= \boldsymbol{R}_2 \cdot \boldsymbol{R}_1, \\ \boldsymbol{T}_3 &= \boldsymbol{R}_2 \cdot \boldsymbol{T}_1 + \boldsymbol{T}_2, \end{aligned} \tag{IV-2}$$

The $\boldsymbol{R}_3$ and $\boldsymbol{T}_3$ parameter values are then used as the refined values of the initial rotation and translation parameters for the next MDSPACE iteration (in place of $\boldsymbol{R}_1$ and $\boldsymbol{T}_1$, respectively).

To estimate the rotation $\boldsymbol{R}_2$ and the translation $\boldsymbol{T}_2$, the rigid-body motion between the initial conformation (atomic coordinates $\boldsymbol{r}_i$) and the finally fitted conformation (atomic coordinates $\boldsymbol{r}_f$) is measured by the minimizing the RMSD between these two conformations:

$$\min_{\boldsymbol{R}_2, \boldsymbol{T}_2} \sqrt{\frac{1}{N} \sum_{n=1}^{N} \| \boldsymbol{r}_f^n - (\boldsymbol{R}_2 \cdot \boldsymbol{r}_i^n + \boldsymbol{T}_2) \|^2},$$

where N is the number of atoms. The RMSD minimization is performed using an optimization algorithm based on singular value decomposition that is available in BioPython [199].

IV.2.6. PCA-based refinement of MD simulations

The 3D-to-2D flexible fitting accurately fits most particle images. However, some images are more difficult to fit, especially the images associated with specific particle views for which the conformational change is less detectable or ambiguous in the projection plane (e.g., a motion mainly along the projection axis will affect the 2D projection but the conformational change will be less detectable in the projection). In such cases, the fitting will most likely induce no displacement from the initial atomic positions (the initial conformation) or perform small-scale, local rearrangements. Most likely, it will not find the correct conformation in images with such views. This is a direct consequence of fitting each particle image individually (independently of other images) in the presence of high level of noise in the images.

To make the approach more robust to particle views and noise, MDSPACE involves an iterative approach that refines the MD-based 3D-to-2D flexible fitting of individual particle images by incorporating the ensemble conformational information into MD simulation, which encourages MD simulations of the particle images with "bad" views to follow the principal motion directions learned from the images with "good" views at the previous iteration. In this approach, the information about the principal motions learned at one iteration is used as a prior to reanalyze the dataset at the next iteration, which boosts the MD-based fitting of individual particle images making it robust to difficult views and noise. For this reason, at the end of each MDSPACE iteration, the conformations obtained by individually fitting different particle images are rigid-body aligned with respect to the initial conformation and then analyzed using PCA. The PCA results in principal component vectors that represent global motions seen in data at that MDSPACE iteration, which are then incorporated in a new round of the fitting at the next MDSPACE iteration. More precisely, the PCA-space conformational information is incorporated into normal-mode empowered MD-based flexible fitting by replacing normal-mode vectors by principal component vectors of the PCA space, which we refer to as PCA-empowered MD-based flexible fitting. This encourages the simulation to move along the principal component vectors found in the previous MDSPACE iteration. The PCA-empowered MD-based flexible fitting yields a new set of atomic structures, whose principal components can be obtained by PCA and then incorporated in a new round of the fitting. MDSPACE, which alternates the fitting with the PCA analysis, refines the PCA space over the iterations, making it closer to the target conformational landscape, as illustrated in Figure 16.

It is worth noting that the PCA-empowered MD-based flexible fitting in MDSPACE not only refines the flexible fitting (the finally fitted conformations) but also refines the initial rigid-body alignment of the particle images. More precisely, at the end of each iteration of MDSPACE, the rigid-body parameters that initiated this iteration are combined with the alignment parameters extracted from the MD simulation performed at the same iteration (see section IV.2.5) and the obtained refined rigid-body parameters are used to start the new MDSPACE iteration.

IV.2.7. General recommendation for running MDSPACE iteratively

To increase speed and accuracy of analyzing large datasets of single particle images, we generally recommend using normal-mode empowered MD simulations for the first MDSPACE iteration and using principal-component empowered MD simulations (PCA-based refinement) for all other iterations. Adding normal modes to MD simulation accelerates the fitting. Adding principal component vectors from the previous MDSPACE iteration to MD simulation in the next MDSPACE iteration improves robustness to difficult views and noise. Regarding the number of

principal component vectors to use for the PCA-based refinement, 3 principal component vectors may generally be enough, but possibly more will be required for some systems. The number of principal component vectors can be selected based on the observed decrease in the singular values of the PCA components. The computational cost of adding more principal component vectors is negligible.

A coarse-to-fine data processing scheme can be used to additionally speed up MDSPACE processing (e.g., in the case of large data sets of large complexes such as ribosomes). More precisely, at the first MDSPACE iteration (normal-mode empowered MD simulations), the principal components of the conformational variability can be learned by processing a small data subset, which will result in a coarsely estimated low-dimensional conformational landscape. At the next MDSPACE iteration (principal-component empowered MD simulation), the conformational landscape can be refined using a larger number of images and the principal components obtained at the previous MDSPACE iteration.

IV.2.8. Software implementation

IV.2.8.1. *MD simulation, Normal Mode Analysis, and PCA methods and software:*

The iterative MDSPACE method requires a large number of computationally costly MD simulations, as MD-based fitting is applied several times for each particle (*M* times per particle, for *M* iterations of MDSPACE). To reduce the computational cost, we choose a coarse-grained approach for MD simulations using off-lattice Cα Gō model as described by Clementi and collaborators [147] (see section II.1.2). The Cα Gō model can be extended to Phosphorus atoms, as done for the experiments with 80S ribosome-tRNA cryo-EM dataset [89]. Gō models can successfully capture native dynamics and conformational transitions of diverse systems [200] using much smaller computational resources than all-atom simulations. However, the non-local interaction in Gō-like models are determined from a "native state" corresponding to the experimental structure, which tends to bias the dynamics towards the experimental structure [146]. In the case of smaller systems (e.g., much smaller than the ribosome studied in this work), one could replace the Cα Gō model by all-atom simulation as it would avoid such bias. Gō models were obtained by SMOG2 software [201]. As shown here, short 30-picosecond MD simulations (used with both synthetic and experimental datasets) are sufficient for 3D-to-2D flexible fitting with MDSPACE. As in the case of NMMD (Chapter III), MD simulations and calculation of normal modes in our experiments were performed in GENESIS and ElNemo, respectively, which are currently also available in ContinuousFlex PCA is obtained by probabilistic principal component analysis implementation in Scikit-Learn [202, 203].

IV.2.8.2. *PCA space clustering and 3D reconstructions:*

To interpret continuous conformational variability from a low-dimensional space such as a PCA space, we performed clustering of close points in the PCA space and grouping of the corresponding particle images into 3D reconstructions to visualize conformational variability in terms of EM maps. A new clustering tool was developed to allow a manual or automated drawing of a trajectory of points in the PCA space and performing an automatic clustering of each particle image to the closest point on the trajectory. An automatically generated trajectory consists of a set of points regularly spaced on a line that can be drawn along a PCA axis or any other direction in the PCA space. The trajectory of points can be manually dragged to adjust to data distribution. The resulting clusters of particle images in the PCA space can be manually refined (by adding or removing data, including removal of outliers). The 3D reconstruction of the EM maps from the clusters is obtained by direct Fourier interpolation using Xmipp [54, 204].

IV.3. Results

IV.3.1. Experiment with synthetic cryo-EM data

In this subsection, I show the performance of MDSPACE using synthetic data of the heterodimeric ABC exporter TmrAB [4].

IV.3.1.1. *Synthetic dataset of TmrAB*

To assess the performances of MDSPACE in a controlled environment (with a known ground-truth solution), we synthesized a dataset by simulating experimental cryo-EM conditions as much as possible. The system studied is a heterodimeric ABC exporter, TmrAB, composed of two ABC proteins, TmrA and TmrB, in complex with a nanobody, Nb9F10 (Figure 17a). Multiple atomic models of the complex in different conformations, derived from cryo-EM maps, are available in the Protein Data Bank (PDB) [4] including an outward-facing conformation (PDB-6RAH, designated here as $TmrAB_{OF}$) and an inward facing conformation (PDB-6RAF, designated here as $TmrAB_{IF}$). $TmrAB_{IF}$ has a closed extracellular gate and an open intracellular gate (Figure 17a), whereas $TmrAB_{OF}$ has an open extracellular gate and a closed intracellular gate (Figure 17b, left).

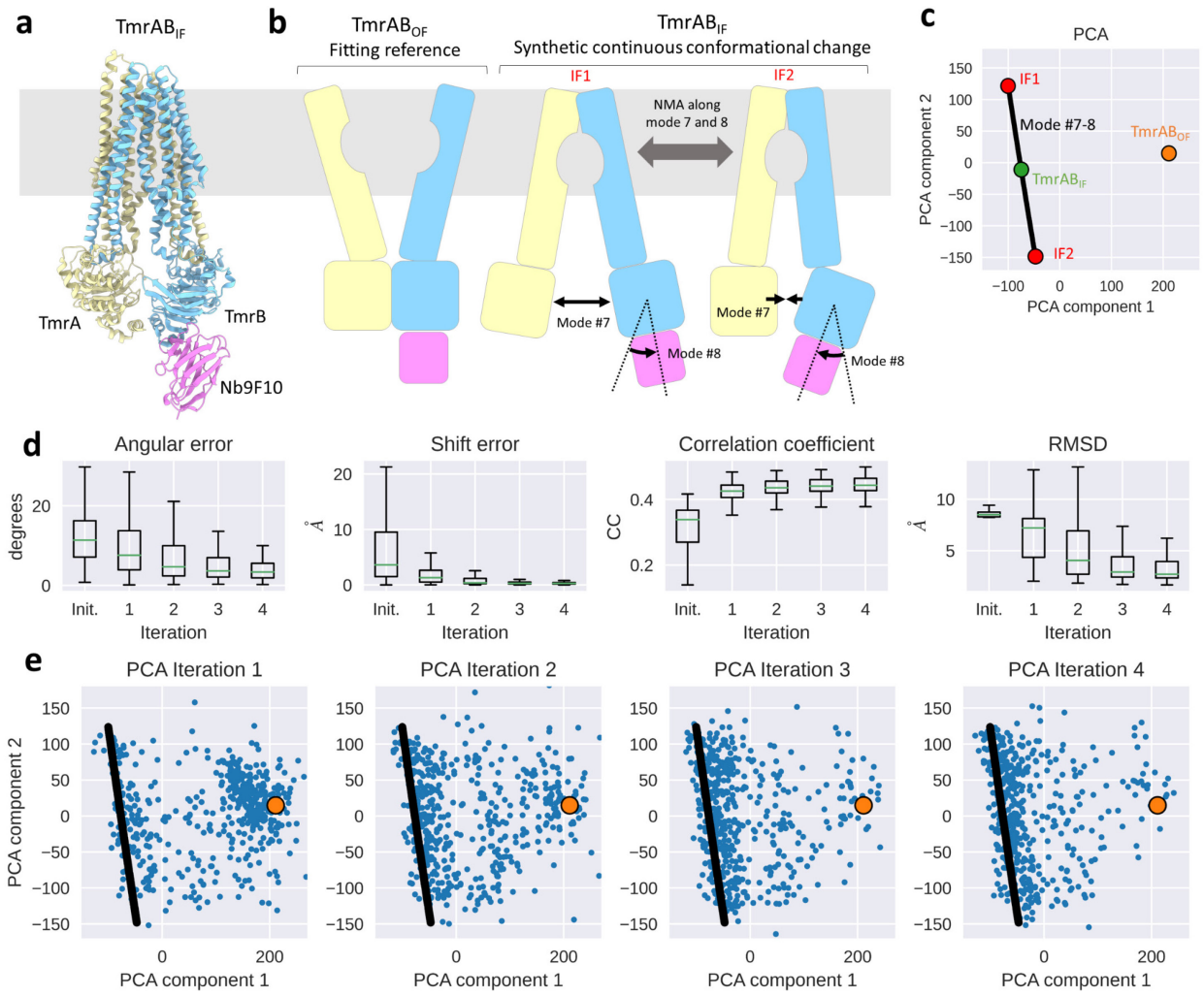

Figure 17 : MDSPACE analysis of the synthetic dataset of TmrAB. (a) Structure of TmrAB in inward-facing conformation (TmrABIF). (b) Diagram of TmrAB in outward-facing conformation (TmrABOF, the initial conformation for fitting) and the synthetic continuous conformational change simulated from TmrABIF using modes 7 and 8. IF1 corresponds to the conformation with negative mode amplitudes and IF2 with positive mode amplitudes. (c) Two-dimensional PCA space of the ground-truth synthetic conformations (black line: synthetic conformational transition trajectory), the initial conformation for the fitting (orange point), and the conformation used to generate the conformational variability (green point). (d) Accuracy of MDSPACE analysis of 500 particles measured at each MDSPACE iteration shown as box plots (black) and median (green). From left to right: errors between the ground-truth and estimated angles; errors between the ground-truth and estimated shifts; correlation coefficients between the images and the projections of the estimated (fitted) conformations; and RMSDs between the ground-truth and estimated (fitted) conformations. (e) Evolution of the principal component space over the iterations (from left to right: iteration 1 to 4); blue points represent 500 fitted conformations, each of which was obtained by fitting the initial conformation (orange color) to one of the images synthesized from 500 ground-truth conformations (black color).

In this experiment, $TmrAB_{OF}$ was used as the initial conformation for 3D-to-2D flexible fitting against synthetic images, whereas the $TmrAB_{IF}$ was used with its normal modes 7 and 8 to generate multiple synthetic conformations from which images were generated, by randomly sampling a continuous trajectory of the conformational transition defined by these two normal modes. We note here that normal modes are ordered according to their frequency and that the first 6 lowest-frequency normal modes (modes 1-6) are never used, as representing rigid-body displacements. The $TmrAB_{IF}$'s mode 7 describes opening and closing of the intracellular gate and mode 8 describes a rotation of the intracellular part of TmrB together with the nanobody, as shown in Figure 17b. To synthesize multiple conformations of the complex, we performed random atomic displacements of $TmrAB_{IF}$ along modes 7 and 8, using the same normal-mode amplitude for both normal modes, where this normal-mode amplitude was sampled from a random uniform distribution in the range [-100,100]. This procedure was used to generate 500 different conformations along the trajectory. Figure 17c shows the ground-truth synthetic continuous conformational variability that is projected onto the first two principal component axes (ground-truth two-dimensional PCA space), together with the projections of $TmrAB_{IF}$ and $TmrAB_{OF}$ on the same PCA space. By observing the relative distances between the points in the ground-truth PCA space (Figure 17c), one can notice that the initial conformation for 3D-to-2D fitting against the synthetic images ($TmrAB_{OF}$) is significantly far from the target conformations (ground-truth trajectory). More precisely, the average root mean square deviation (RMSD) between $TmrAB_{OF}$ and the synthetic conformations generated from $TmrAB_{IF}$ is 8.1 Å.

The synthetic conformations were converted into high-resolution density maps using a method based on scattering factors [205] and projected (1 projection per conformation) using ray casting in real space to obtain a set of 2D particle images of 128 × 128 pixels and a pixel size of 2 Å × 2 Å. The projection orientations followed a uniform random angular distribution around a 3D sphere and the particles in the images were shifted following a uniform random distribution in the [-5,5] pixel range (i.e., [-10,10] Å). We simulated the effect of the electron microscope by applying a CTF with noise added before and after the CTF, using the image formation model described elsewhere [50]. We synthesized realistically looking images using a simulated 200 kV microscope with a spherical aberration of 2 mm, a defocus of -0.5 μm, and Gaussian noise distributed during the image formation so as to simulate a SNR of 0.1 in the images.

Both the low SNR of the synthetic images and the large RMSD between the initial and target conformations (average RMSD of 8.1 Å) are challenges for MDSPACE. For each synthetic image, the 3D-to-2D flexible fitting of $TmrAB_{OF}$ should close the extracellular gate and precisely fit the opening of the intracellular gate and the rotation of the nanobody, in the presence of high noise in the images and using 2D biasing potential.

IV.3.1.2. *Rigid-body pre-alignment and MDSPACE iterations*

To rigid-body pre-align the initial conformation with the particle images, we performed 4 iterations of projection matching between the images and the density map from the initial conformation, in Xmipp [54, 204], using the angular sampling rate of 7°, 5°, 3°, and 2° over the 4 iterations as well as no limit on the angular search in the first iteration and a limit on the angular search range of 10°, 6°, and 4° in the other three iterations. As the initial conformation is very different from the target conformations in the images, the projection matching gives only a rough alignment. In this article, we demonstrate that MDSPACE is able to refine such rough initial rigid-body alignment while also iteratively refining deciphering of the conformational variability.

We run 4 iterations of MDSPACE using 30-picosecond MD simulations and the force constant of 2000 kcal/mol on the dataset of 500 synthetic particle images. Although our general recommendation is to run the first iteration of MDSPACE using normal-mode empowered MD-based fitting and all other iterations using principal-component empowered MD-based fitting (as described in the section IV.2.6), the fitting here was performed without normal modes to avoid bias as the images were synthesized using normal modes. More precisely, flexible fitting using only MD simulations was used in the first iteration and PCA-based refinement in the other 3 iterations of MDSPACE. Starting from the second iteration, the PCA-based refinement at each iteration was performed using the first 3 principal components (3 most dominant motions) obtained at the previous iteration. As shown below, this protocol refines both the initial rigid-body alignment (pre-alignment) and the conformational variability retrieval (fitted conformations).

IV.3.1.3. *Performance regarding recovery of the ground-truth rigid-body parameters*

The MDSPACE analysis of the synthetic dataset shows that the rotation and translation parameters obtained at the pre-alignment step get refined over the iterations (the first two plots of Figure 17d, from left to right). Figure 17d shows the angular and shift error distributions over the iterations, i.e., per-iteration statistics on the errors between the ground-truth and estimated angles (the first plot from the left) and shifts (the second plot from the left). Recall here that the estimates of the angles and shifts at each iteration are obtained by combining the initial angles and shifts at that iteration and the angles and shifts extracted from the MD simulation at the same iteration (equation (IV-2)).

The median angular error of the pre-alignment is 11.3 degrees and the median shift error is 3.6 Å, meaning that the projection matching resulted in a relatively poor rigid-body pre-alignment. Even with such a poor pre-alignment, the recovery of the ground-truth angles and shifts at the 4th iteration of MDSPACE is satisfactory (the median angular and shift errors dropped to 3.3 degrees and 0.25 Å, respectively).

IV.3.1.4. *Performance regarding recovery of the ground-truth continuous conformational transition*

The MDSPACE analysis of the synthetic dataset shows that the conformational space gets refined over the iterations (Figure 17e). Figure 17e shows the evolution of the conformational space estimation over four MDSPACE iterations, together with the ground-truth conformations (the samples of a line-form trajectory in the normal mode space in this experiment) and the initial conformation used for the 3D-to-2D flexible fitting. For simplicity of the comparison over the iterations, we here show the conformational space reduced to two dimensions (2D PCA space), whereas the first three principal components were used for the PCA-based refinement (from the second MDSPACE iteration on). Also, it should be recalled that the PCA is recalculated at each MDSPACE iteration (PCA components change over the iterations) and that the PCA components calculated at one iteration are used to empower the MD-based flexible fitting at the next iteration. However, for the purpose of a visual comparison between the fitted and target (ground-truth) conformations, Figure 17e shows the fitted conformations in the same space as the ground-truth conformations. The principal components actually used to empower the MD-based flexible fitting (in the PCA-based refinement) differ from those shown in Figure 17e as they do not comprise the ground-truth information.

At the first MDSPACE iteration, we observe that a part of the fitted conformations is close to the ground-truth conformations, but many are far (approximatively 50 % are close to the initial conformation used for the 3D-to-2D flexible fitting and they correspond to the images with the projection direction with poor conformational variability information) (Figure 17e). After the second MDSPACE iteration (first PCA refinement), we observe a clear improvement of the conformational space as the fitted conformations get closer to the target conformations (ground-truth trajectory). This tendency continues over the iterations and, at the last iteration, most fitted conformations are close to the target conformations and only a few remain close to the initial conformation.

The analysis of the CCs between the particle images and the projections of the fitted conformations and the analysis of the RMSDs between the fitted and target (ground-truth) conformations, shown in the last two plots of Figure 17d (from left to right), confirm that the conformations get refined over the iterations. The CC distribution shows a significant CC increase from the initial pre-alignment iteration to the first iteration (the median CC value increased from 0.33 to 0.42), then the CC increases slowly over the following iterations (the median CC values of 0.42, 0.43 and 0.44 for iterations 2, 3 and 4). The RMSD distribution shows that the RMSD decreases significantly up to the third iteration (the median RMSD values of 8.4 Å and 2.6 Å at iterations 1 and 3) and then changes slowly.

IV.3.1.5. *Recovery of the ground-truth continuous conformational transition in terms of animations and 3D reconstructions from the conformational space*

We analyzed the conformational distribution in the PCA space obtained at the 4th iteration of MDSPACE in terms of 3D reconstructions from clusters of close points in this space, i.e., from the corresponding synthetic particle images, in order to check whether 3D reconstructions along the point distribution in the PCA space follow the ground-truth conformational transition trajectory.

The 3D reconstructions were performed from clusters determined in the PCA space as described in the section IV.2.8.2 (Figure 18a). First, we interactively defined a 5-point linear trajectory that approximatively fits the point distribution in the PCA space. Then, we performed an automated clustering based on the closest points to each of the 5 points and interactively removed the outlier points. Finally, we performed a 3D reconstruction from each of the 5 clusters without outliers, using the rigid-body parameters obtained after the 4th iteration of MDSPACE, which resulted in five reconstructed EM maps.

Figure 18b shows two atomic conformations corresponding to two extremums of the ground-truth synthetic trajectory (designated as IF1 and IF2) and Figure 18c shows the EM maps obtained from the first and last clusters along the 5-point linear trajectory (Figure 18a). By comparing these two 3D reconstructed maps (Figure 18c) with the corresponding ground-truth synthetic atomic conformations (Figure 18b), we observe that the method was able to capture the ground-truth synthetic conformational change.

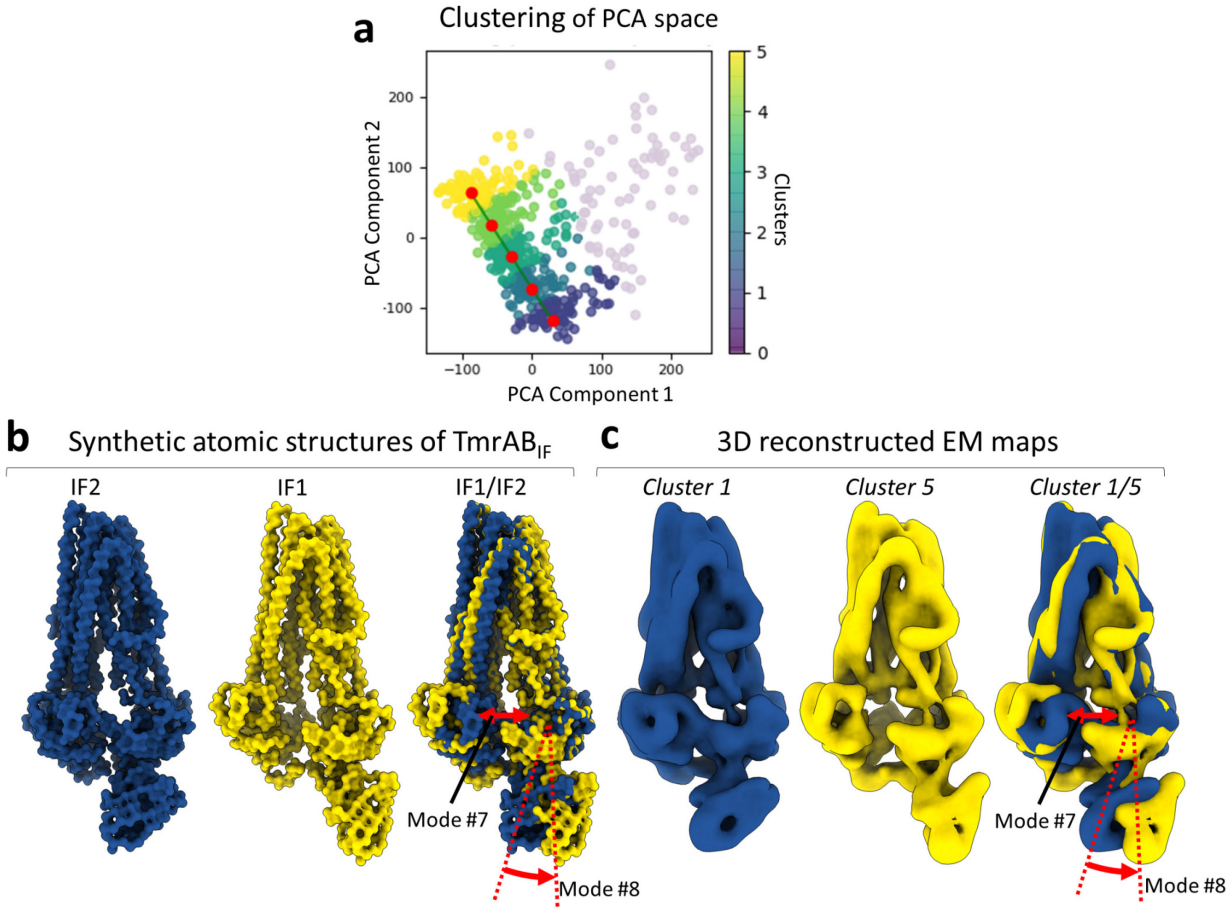

Figure 18 : Recovery of the ground-truth synthetic conformational transition trajectory of TmrAB based on PCA space clustering and 3D reconstructions from the clusters. (a) Five PCA space clusters obtained automatically along an interactively defined trajectory (red points) and colored from yellow to blue. (b) Two synthetic atomic conformations of TmrAB in inward-facing conformations (TmrABIF) corresponding to the two extremums of the ground-truth synthetic trajectory, denoted as IF1 and IF2. (c) 3D reconstructed EM maps from the yellow and blue clusters shown in (a). The conformational transition is visible on the superposed yellow and blue atomic structures (b) and EM maps (c). The color code is the same for the clusters in the PCA space (a), the atomic structures (b), and EM maps (c).

IV.3.2. Experiment with cryo-EM data from EMPIAR

In this subsection, we show the performance of MDSPACE using experimental cryo-EM data of yeast 80S ribosome-tRNA complexes available in EMPIAR database under the code EMPIAR-10016 [89].

IV.3.2.1. *EMPIAR-10016 dataset (yeast 80S ribosome-tRNA complexes)*

During protein synthesis, tRNAs are translocated from the A (aminoacyl) to P (peptidyl) to E (exit) sites of the ribosome. During this process, two tRNAs adopt hybrid A/P and P/E states while progressing from the A-A and P-P states to the P-P and E-E states, which also involves a rotation between the two subunits of the ribosome [206-208].

In the original study of yeast 80S ribosome-tRNA complexes, resulting in the EMPIAR-10016 dataset publication [89], two cryo-EM maps were obtained using the FREALIGN likelihood-based image classification [14, 209] into 5 classes from an initial set of 86,866 particle images, and deposited in the EMDB (accession codes: EMD-5976 and EMD-5977). The EMPIAR-10016 dataset contains a stack of particle images (image size: 360 × 360 pixels; pixel size: 1.05 Å × 1.05 Å) and 5 metadata files containing the orientation and translation parameters for 5 image classes. Two of the metadata files contain the parameters of 23,726 and 22,369 images, which correspond to the classes that yielded the two reconstructed cryo-EM maps (EMD-5976 at the resolution of 6.2 Å and EMD-5977 at the resolution of 6.3 Å, respectively). The EMD-5976 map corresponds to the rotated conformation (the ribosome inter-subunit rotation of around 9°) with one tRNA in a hybrid P/E state (designated as 80S-tRNA). The EMD-5977 map corresponds to the nonrotated conformation with two tRNA in the classical P-P and E-E states (designated as 80S-2tRNA). Two atomic models were derived from these cryo-EM maps (Figure 19) and deposited in the PDB (accession codes: PDB-3J77 and PDB-3J78) [89].

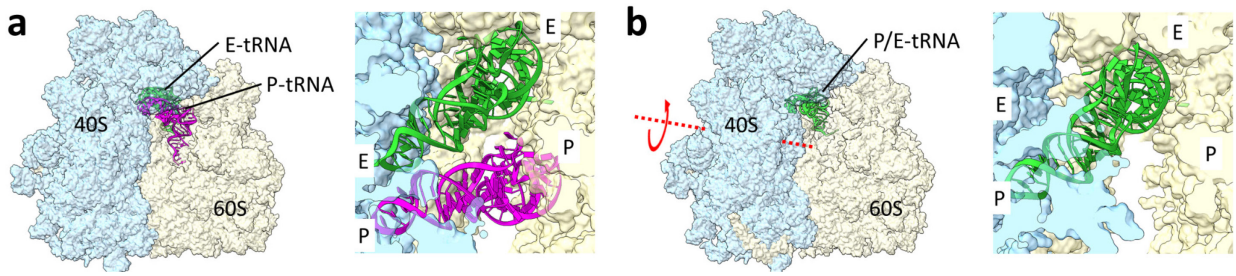

Figure 19 : Atomic models of yeast 80S ribosome-tRNA complexes derived from cryo-EM maps obtained with FREALIGN likelihood-based image classification of EMPIAR-10016 set of single particle images [89]. (a) Nonrotated conformation with two tRNAs in the classical E/E and P/P states (80S-2tRNA). (b) Rotated conformation with one tRNA in a hybrid P/E state (80S-tRNA). The 40S and 60S subunits are represented as light blue and light-yellow surfaces, whereas tRNAs are displayed as ribbons. The boxes at the right in (a) and (b) show close-up top views of the tRNAs.

IV.3.2.2. *MDSPACE analysis of EMPIAR-10016 dataset*

For the conformational variability analysis with MDSPACE, we used the particle images that yielded the EMD-5976 and EMD-5977 maps (all available 46,095 particle images), and we used their available orientation and translation parameters (determined by FREALIGN) as rigid-body pre-alignment parameters. We down-sampled the original particle images by a factor of two in each of the two dimensions, yielding images of size 180 × 180 pixels with a pixel size of 2.1 Å × 2.1 Å. The other 3 classes identified in the original publication [89] in the EMPIAR-10016 dataset appear less clean and less populated, and were therefore not used in our analysis with MDSPACE.

We performed 2 iterations of MDSPACE using 30-picosecond MD simulations, a force constant value of 10,000 kcal/mol, and the coarse-to-fine data processing scheme described in the section IV.2. As the initial conformation, we used the atomic model of the 80S-tRNA conformation (the rotated conformation with one tRNA in a hybrid P/E state) (Figure 19b). In the first iteration, we analyzed 10,000 particles using normal-mode empowered MD-based flexible fitting (10 lowest-frequency normal modes were added to MD simulations as in NMMD). In the second iteration, we analyzed the entire set of 46,095 particles using principal-component empowered MD-based flexible fitting (PCA refinement) and the first 3 principal components (3 most dominant motions) obtained at the previous iteration.

The CC distribution shown in Figure 20c follows the same behavior as in the experiment with synthetic data. We observe a strong increase of the CC from the rigid-body pre-alignment to the first iteration (the median CC increases from 0.121 to 0.149), which is followed by a light increase at the next iteration (the median CC increases from 0.149 to 0.161 between the first and second iterations). This indicates that the conformations obtained at the first iteration get refined at the second iteration of MDSPACE.

IV.3.2.3. *MDSPACE recovers 80S-tRNA and 80S-2tRNA conformational states*

The analysis of the conformational space obtained with MDSPACE shows a continuum of conformational states (Figure 20a). The singular values of the principal components (Figure 20b) decrease by 20% between the first and third components. To observe the extracted conformational variability in 3D, we performed clustering of the conformations in the PCA space, using 5 clusters linearly distributed along the first principal component (Figure 20a). Figure 20d shows the 3D reconstructions from the first and last clusters along the first principal component. These two clusters contain a relatively low number of particles (approximatively 3,000 particles per cluster) as they correspond to the extremums of the trajectory along the principal axis, which may explain a low resolution of the two reconstructed maps. Additionally, the maps in Figure 20d were low-pass filtered for the sake of reducing noise for visualization (low-pass cutoff frequency: 10 Å). We identified the first and last clusters (cluster 1 and 5) to correspond to the 80S-tRNA and 80S-2tRNA states, respectively, based on the presence of a single tRNA in a hybrid P/E state in the 3D reconstruction from cluster 1, two tRNA in the E/E and P/P states in the 3D reconstruction from cluster 5, and the inter-subunit rotation in the 3D reconstruction from cluster 1 (pre-translocation state) with respect to the 3D reconstruction from cluster 5 (post-translocation state). The remaining clusters (clusters 2 to 4) correspond to intermediate states of the inter-subunit rotation.

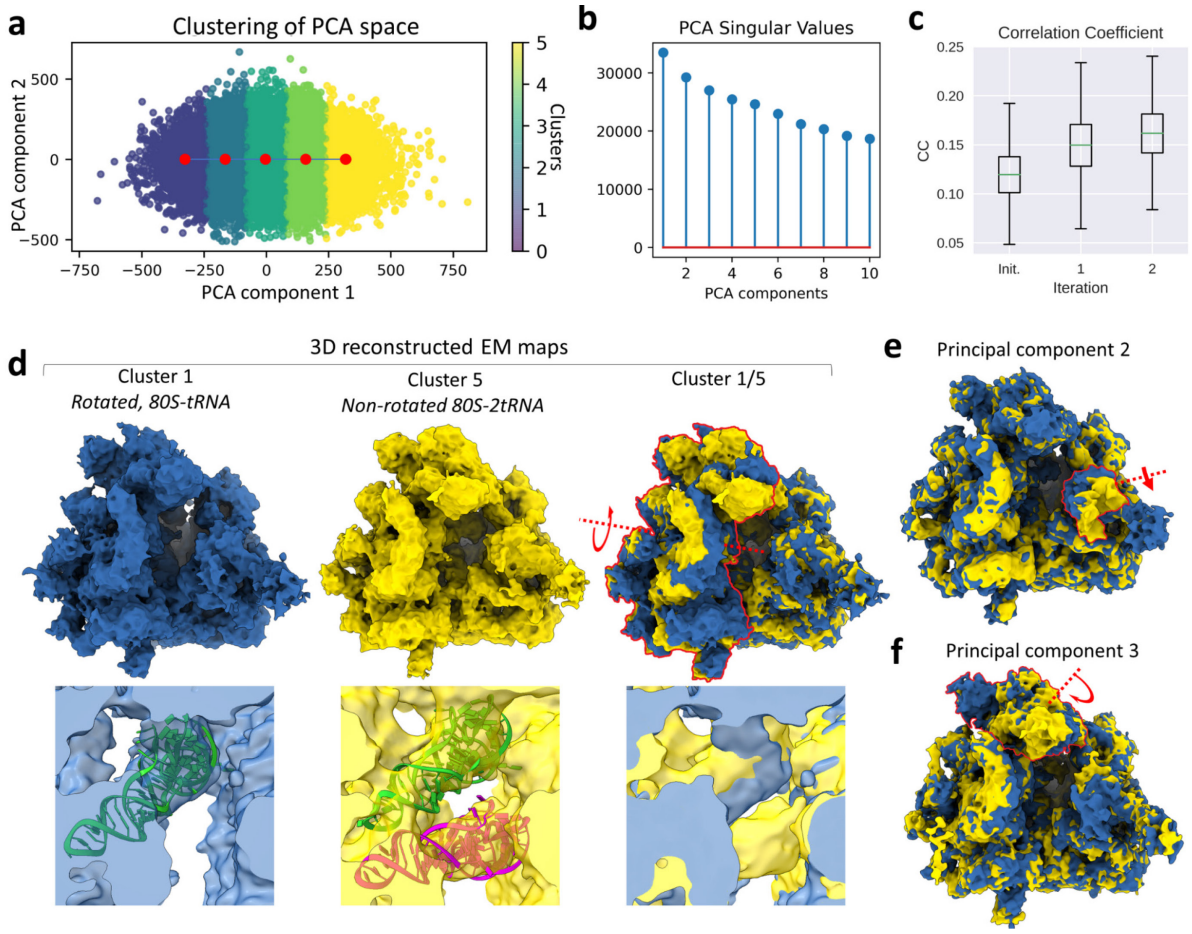

Figure 20: MDSPACE analysis of 80S ribosome-tRNA complexes from EMPIAR-10016 cryo-EM dataset [89]. (a) Conformational space obtained by MDSPACE, described by the first two PCA axes. Five PCA-space clusters colored from blue to yellow were automatically obtained along an interactively defined trajectory along the first PCA axis (red points). (b) Singular values of the PCA components. (c) Distribution of the correlation coefficient over the MDSPACE iterations. (d) 3D reconstructed EM maps from cluster 1 (blue points in (a)) and cluster 5 (yellow points in (a)). At the bottom, the first two panels from left to right show close-up top views of the E and P sites on the 3D reconstructed EM maps from cluster 1 and cluster 5 overlapped with the tRNAs (ribbon representation) of the 80S-tRNA model (rotated) and the 80S-2tRNA model (non-rotated), respectively. The remaining panel at the bottom shows a close-up top view of the overlapped EM maps from clusters 1 and 5. The red arrow shows the rotation of the 40S subunit. (e) 3D reconstructed EM maps from clusters 1 and 5 along the second PCA axis. (f) 3D reconstructed EM maps from clusters 1 and 5 along the third PCA axis. The red arrows in (e) and (f) indicate the conformational changes. The clustering procedure used for the first PCA axis was repeated for the second (e) and third (f) PCA axes.

IV.3.2.4. *MDSPACE reveals multiple conformational changes of the 80S ribosome*

Additional conformational changes were observed by clustering along the second and third principal components (in 5 clusters by repeating the clustering procedure that was used for the first principal component). The conformational change along the second principal component shows a displacement of ribosomal protein P0 (Figure 20e), which with other P-stalk ribosome proteins

plays a role in the interaction with translation initiation factors and promotion of translation initiation [210]. The conformational change along the third principal component shows a rotation of the head of the small subunit (Figure 20f), playing a role in facilitating the translocation process [211]. Other, but finer, conformational changes have been observed, as described in the next subsection.

IV.3.2.5. *MDSPACE reveals gradual transitions with many intermediate states at atomic scale*

MDSPACE yielded the conformational space at atomic scale. In this subsection, we analyze this space in terms of atomic models along the first principal component. We selected a 9-point linear trajectory along this principal component (Figure 21a) and performed an inverse PCA for these points, which allowed us to visualize the atomic coordinate displacements along this 9-point trajectory. The 9 obtained atomic models were compared with the 80S-tRNA and 80S-2tRNA models, which indicated a continuum of conformational states and agreement with the expected conformational changes (Figure 21c-e). Figure 21c shows the conformational change of the L1 stalk (25S rRNA single loop) which is known to interact with the E-site tRNA [5, 89, 212]. Figure 21d shows the conformational change of ribosomal protein uS12, located on the head of the small subunit, emphasizing the inter-subunit rotation. Figure 21e shows the conformational change of the section of 18S rRNA at the P-site of the small subunit. We can note that MDSPACE resolves the 80S-tRNA and 80S-2tRNA models, which approximately correspond to the 4^{th} and 9^{th} points of the trajectory, respectively (Figure 21c-e).

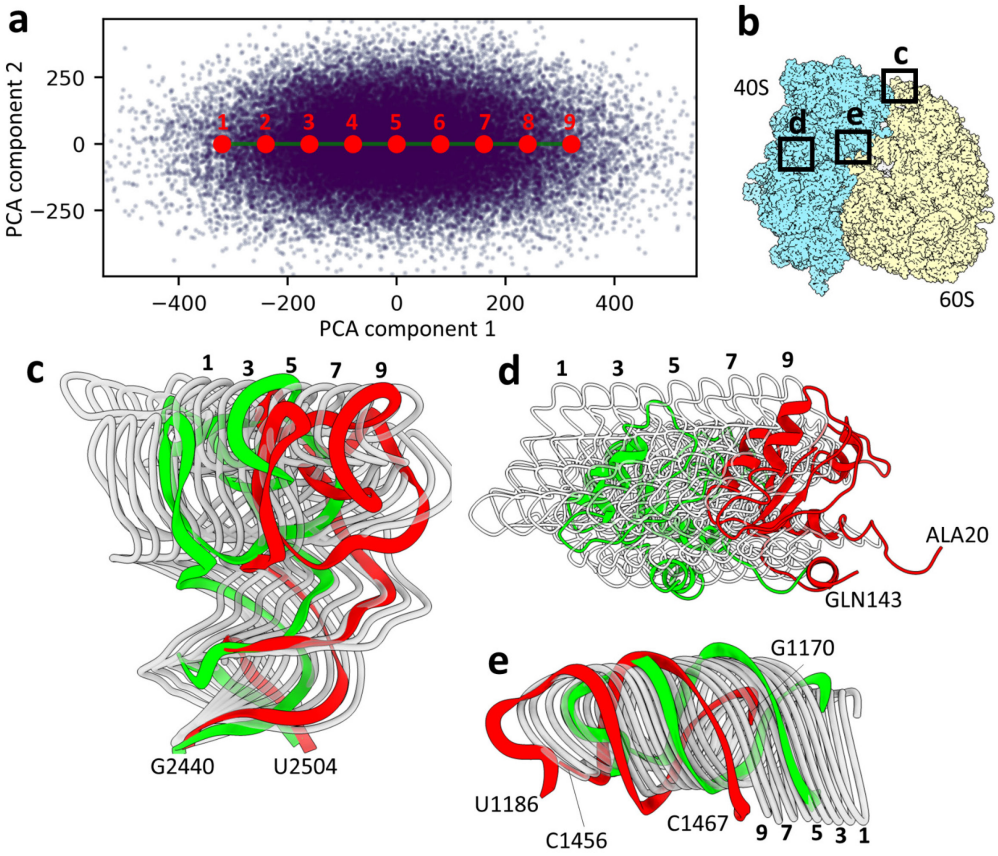

Figure 21 : Continuous conformational trajectory at atomic scale identified in the PCA space. (a) Conformational space obtained with MDSPACE (also shown in Figure 20a). The red points correspond to a 9-point linear atomic-coordinate trajectory along the first PCA axis, selected interactively in the PCA space. (b) Approximate locations of the 9 atomic models shown in (c)-(e) corresponding to the 9 red points in (a). (c) Motion of 25S rRNA single loop of the large-subunit L1 stalk (white) along the 9-point trajectory shown in (a). (d) Motion of ribosomal protein S12 (white), part of the head of the small subunit, along the 9-point trajectory shown in (a). (d) Motion of the section of the small-subunit 18S rRNA at the P-site (white) along the 9-point trajectory shown in (a). The models in (c)-(e) are superposed with the 80-tRNA and 80S-2tRNA models (green and red, respectively).

IV.3.3. Analysis of p97 conformational heterogeneity using MDSPACE

IV.3.3.1. *Experimental dataset of AAA ATPase p97*

In this study, I aimed at processing the conformational changes of AAA ATPase p97. P97 is an essential unfoldase involved in many cellular processes including protein quality control, a crucial process involved in cancer cell survival, making it an attractive target for chemotherapy [213-215]. P97 also plays a role in viral, bacterial and parasitic infections and mutation of p97 can cause neurodegenerative diseases, making it an important therapeutic protein.

Structurally, p97 assembles a homo-hexamer with two concentric stacked rings composed of two ATPase domains (D1 and D2), surrounded by the N-terminal domains, and the disordered C-terminal domains (Figure 22a). Cryo-EM studies of p97 have long shown that the N-domain is highly flexible [216, 217], and high-resolution cryo-EM analysis resulted in a poor resolution of this domain [3]. The N-domains bind different cofactors [33, 218] to initiate its protein unfolding activity [219-221].

The position of the N-domains of p97, either co-planar to the D1 ring (referred to as the "down-conformation") or above the D1 ring (referred to as the "up-conformation") is a determinant of the co-factors that bind to p97 and target it to different functions in the cell [222, 223]. The N-domains of p97 were observed predominantly in the "up" conformation (Figure 22a) when both bound to ATPγS and in a coplanar conformation when bound to ADP [3]. These domains were resolved at relatively low resolution indicating that the discrete classification conducted during image processing did not well characterize the conformational heterogeneity of the domain.

For this study, I analyzed a cryo-EM data of a mutant of p97 (R155P) in which the N-domains were supposedly stabilized in the up-conformation conformation. Sample preparation and data collection was performed by Sepideh Valimehr, a previous PhD student in my lab at the University of Melbourne [224]. In that previous work, a traditional discrete-classification algorithm using cryoSPARC [13] was performed by Sepideh Valimehr and Mohsen Kazemi (University of Melbourne), and resulted in an EM map reconstruction with a lower local resolution in the regions corresponding to the N-domains (Figure 22b), suggesting that the N-domain is highly flexible. The flexibility of the N-domain both in terms of variation range, and in terms of possible correlated motion between N-domains could inform on how the N-domain interact with other cofactors. Therefore, I used MDSPACE to further this p97 study.

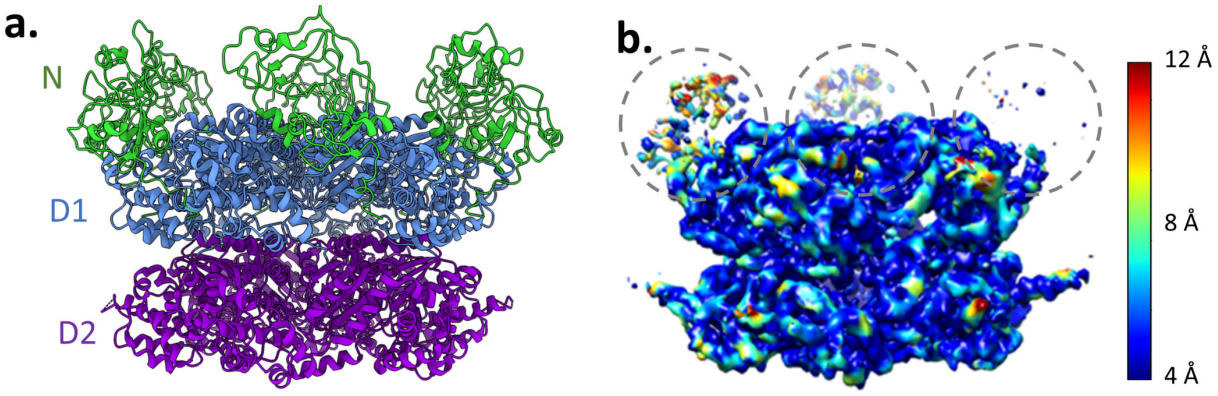

Figure 22: Structure of p97. (a) Atomic model of p97 with the N domains in the "up" conformation (PDB 5FTN [3]) (b) The cryo-EM map resulting from the studied dataset obtained with traditional, discrete-

classification methods. The colormap shows the local resolution. The local resolution map was provided by Sepideh Valimehr and Mohsen Kazemi.

IV.3.3.1. *Analysis of conformational variability of p97 using MDSPACE*

MDSPACE was applied on the original set of 274,640 particles, binned by 2 to a pixel size of 2.62 Å². I used PDB 5FTN [3] as the initial conformation for the fitting, which corresponds to the conformation with the N-domains in the "up" conformation. The particle pose obtained by homogeneous reconstruction in cryoSPARC [13] was used as the initial rigid-body alignment of the particles for MDSPACE. The MD simulations were performed using the Cα-Gō model for 50 picoseconds with the time step of 1 femtosecond. Two MDSPACE iterations were performed. The first iteration was restricted to the analysis of only 5,000 particles. In the second iteration, the entire dataset was used (274,640 particles). For the first iteration, a subset of five lowest frequency normal modes (modes 7-11) of the initial conformation was selected for the fitting. Only the first 10 principal components of the conformational space obtained at the end of the first iteration were used as the normal modes in the second iteration. A set of 242,130 fitted conformations of p97 was obtained at the end of the second iteration. The flexible fitting for the remaining 32,510 particles failed due to instability in the MD simulation that is sometimes due to incorrect parameters used (e.g. too long time step or too large force constant) or image artifacts such as caused by incorrectly picked particles (e.g. ice contamination or edge of the carbon hole). UMAP was applied to the set of fitted structures to project the data onto a low-dimensional conformational state. A further analysis of the conformational variability was performed at the level of the monomers, by extracting the coordinates corresponding to each monomer in the set of the fitted conformations, resulting in a set of 1,452,780 monomer structures. This new set was rigid-body aligned to the D1 and D2 domains (excluding the N-domain) of one monomer of the initial conformation (PDB 5FTN). A new dimension reduction suing UMAP was performed on the set of the aligned monomers.

IV.3.3.2. *Continuous conformational variability at the hexameric level*

MDSPACE analysis revealed a high variability of the N-domains. The atomic coordinates fitted to the particle images were analyzed using UMAP to obtain a low-dimensional conformational landscape in ten dimensions, where the two first dimensions are shown in Figure 23a. The landscape reveals two distinct clusters, one large cluster that contains 98% of the particles and another smaller cluster containing 2% of the particles. The averaged structures for the first and second clusters are presented in Figure 23b and Figure 23c respectively. The coloramp in Figure 23b-c shows the root mean square fluctuation (RMSF), which measures the average variation of each residue with respect to the cluster average structure. The RMSF of the D2 and D1 rings for the large cluster is low (below 2 Å) indicating strong stability of D1 and D2 rings whereas the

RMSF of the N-domains is higher than 10 Å indicating a large conformational variability. The small cluster have a similar variability of N-domains (RMSF above 10 Å), but the D1 and D2 rings are less stable (RMSF above 5 Å). The averaged structure in Figure 23c shows an opening between D1 and D2 rings.

The large cluster contains multiple local minima in the 10-dimensional space that are associated with conformational changes of the N domains, from which six particular conformational states were encircled in Figure 23a. The fitted structures associated to the six regions were averaged and the position of their N domain with respect to the average is presented in Figure 23d. The different conformational states of the N domains in Figure 23d reveal that each N domain is mobile and can adopt a particular inclination along a specific rotation axis which will be further analyzed in the next section. These N domain rotations can be different for each of the six N domains.

However, 3D reconstruction of those sub-clusters still contains N domains that are visible only partially, which indicates heterogeneity in the clusters. Indeed, if we consider each of the six N domains evolving independently of the others along its rotation axis, this results in a very large number of conformational states. Although UMAP was able to well separate evident features such as the motions of D1 and D2 rings and to indicate some of the transitions of the N domain, the high order of the embedding (10 dimensions) makes the clustering a difficult task.

Additionally, a very large number of clusters should be used to fully describe the heterogeneity in the case of exploring the 10 dimensions, which would reduce the number of particles in each cluster and result in very noisy 3D reconstructions (the six shown clusters contain about 10,000 particles, which is sufficient to obtain 3D reconstructions with a reasonable amount of noise).

Another consideration is that the UMAP embedding does not account for rotations along the pseudo-symmetry (C6) axis. Indeed, one particular conformational state could be repeated 6 times in the data (n x 60° rotations along the C6 symmetry axis, n={1,...,6}) which increases even more the complexity of the UMAP embedding. This caused by the initial particle pose that align the particles mainly on the D1 and D2 without considering the slight variation of the N-domain (because classification algorithm cannot this large number of N-domain variations).

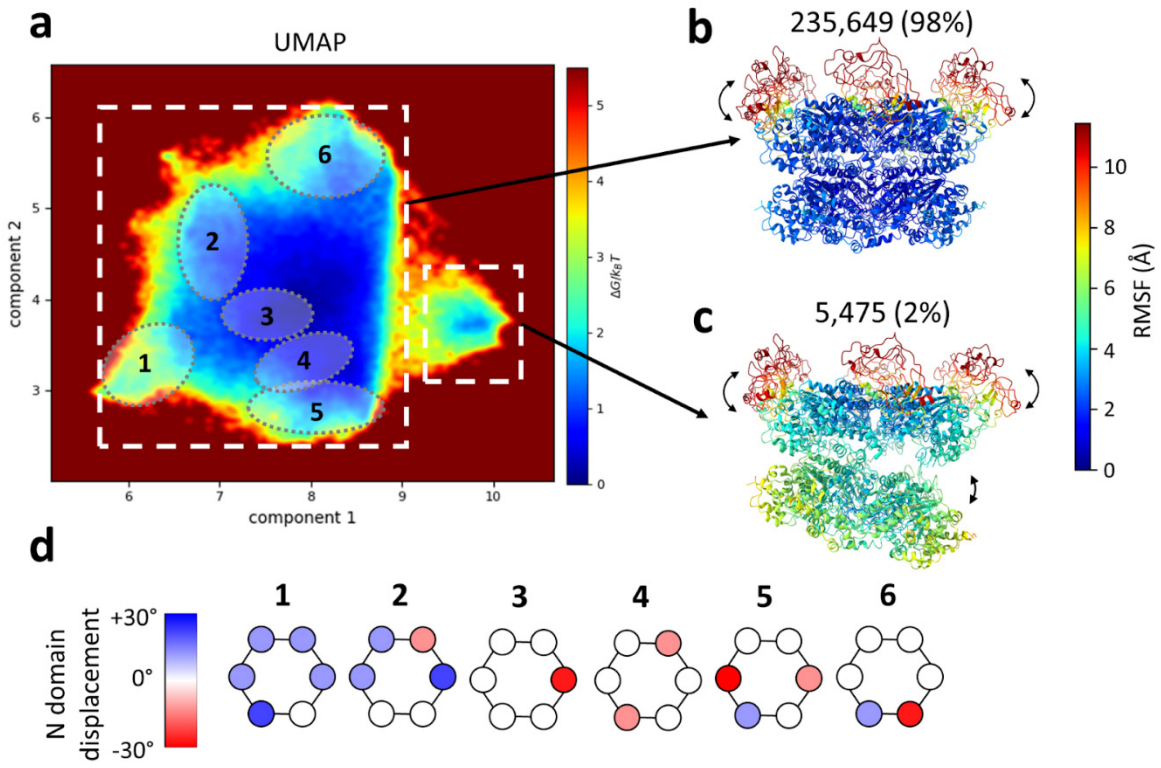

Figure 23: Conformational landscape obtained by MDSPACE at the hexameric level (a) Conformational space obtained by MDSPACE using UMAP. Two clusters can be distinctly separated (two white squares) and are presented in (b) and (c). Six subclusters showing examples of N domain variation are encircled and presented in (d). (b-c) Average atomic structures of the two clusters designated by white squares in (a). The colormap shows the average variation of each residue (RMSF) with respect to the averaged structure. The first cluster (b) shows a strong stability of D1 and D2 combined with variations of the N domain, the second (c) shows variations of the N domains, combined with a gap opening between D1 and D2 rings. (d) N domain displacement around the average structure for the six sub-clusters shown in (a).

IV.3.3.3. *Continuous conformational variability at the monomeric level*

Although the analysis of the conformational variability of the structures fitted by MDSPACE allowed us to observe particular arrangements of the N domains at the hexameric level, it did not allow us to accurately track the range of variations of the N domain. To obtain a more detailed description of the N domain conformations, each hexameric structure was split into six monomeric structures, which were aligned and embedded onto a new conformational landscape using UMAP, shown in Figure 24a. The conformational landscape of the monomers reveals a single dominant motion (half circle trajectory in Figure 24a) with a central low energy region (high density of particles). By selecting regions along this trajectory, as shown with the black boxes in Figure 24a, I identified several positions of the N domain. The fitted structures in the black encircled regions in Figure 24a were averaged to produce the atomic structures presented in Figure 24b. The central region in Figure 24a match the initial conformation of p97 with the N domain up (PDB 5FTN) and

was therefore annotated as ‘0°’ tilt. The left and right parts of the trajectory are associated with downward and upward rotations of the N domain, respectively, with a maximum of approximately 30° in each rotation direction. The averaged atomic structures in Figure 24b can be compared with the atomic structure in the coplanar conformation (PDB 5FTM), which is shown in grey color at the right side of Figure 24b. As it can be noticed, the rotation of the N domain in the coplanar conformation is outside the rotation range observed in the other conformations shown in Figure 24b (the rotation of the N domain is about -45° in the coplanar conformation). The particles corresponding to the black boxes in Figure 24a were used to reconstruct the EM maps that are shown in Figure 24c. The obtained EM maps show the N-domain rotation range of about 60° between the map at the left side and the map at the right side of Figure 24c. Also, these two maps now show a well-defined density of the N domain.

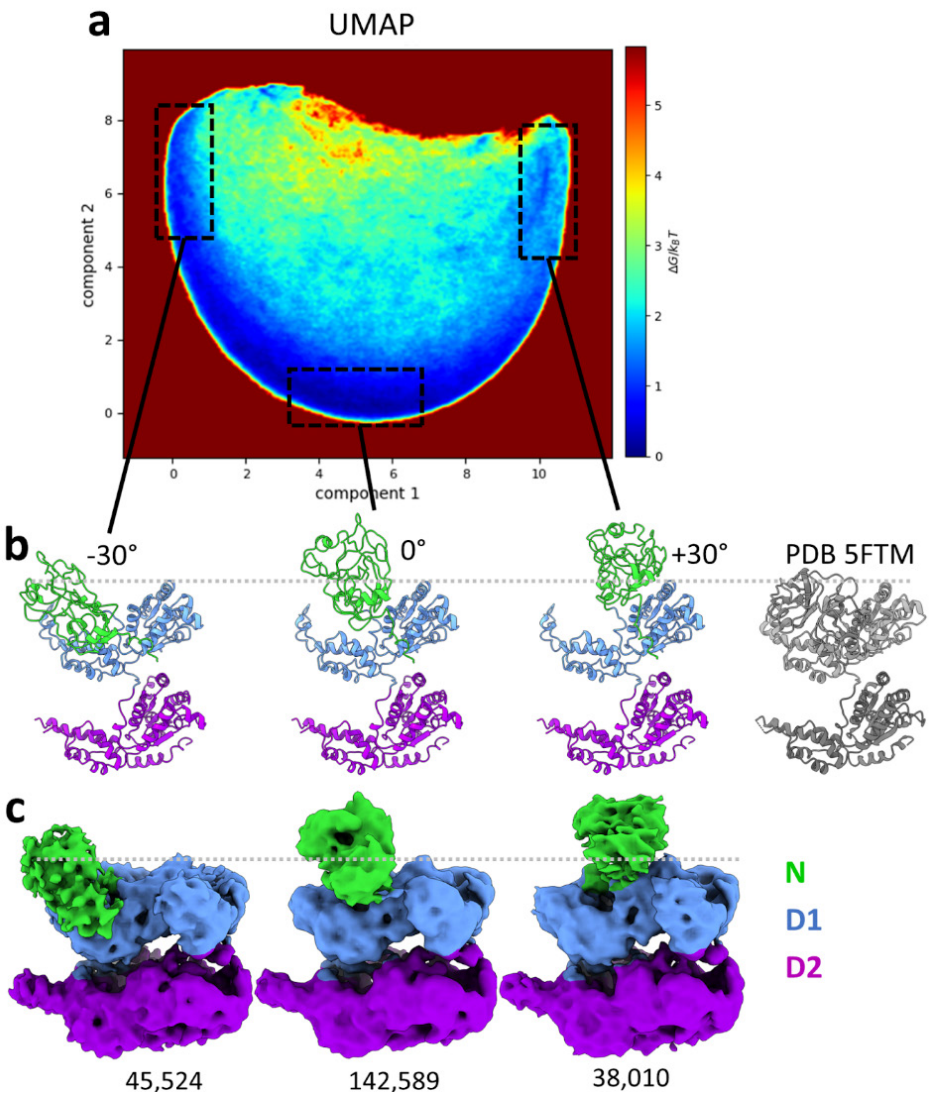

Figure 24: Conformational landscape obtained by MDSPACE at the monomeric level. (a) Conformational landscape of the two first components obtained by UMAP, showing a half circle trajectory along the first component axis. Three regions, encircled in black, correspond to the edges and the middle of the trajectory. (b) The averaged atomic structures of three regions designated in black in (a) with a N-domain angular displacement of -30°, 0° and +30° respectively, and the PDB 5FTM, structure of p97 in coplanar conformation (gray). (c) 3D reconstructions of three regions designated in black in (a) with a N-domain angular displacement of -30°, 0° and +30° respectively.

IV.3.3.4. *Retrieving hexameric conformations from the analysis at the monomeric level*

The conformational trajectory of the rotation of the N domain presented in the previous section allowed us to accurately define a conformational state for each monomer in each particle. This conformational state was reported on its corresponding location on the hexameric structures. For this purpose, the trajectory in Figure 24a was split into three regions considering to have globally homogeneous conformation (i.e. N domain with an angular displacement of approximately -30°,

0° and +30°), as shown in Figure 25a. Then, each monomeric state (-30°, 0°, or +30°) was reported on its location on the hexamer structure, as shown in Figure 25b, creating a particular combination of N domain conformation for each of the 242,130 particles. There is a total of 130 unique combinations of states, and for each combination, the size of the population was counted and reported in the histogram in Figure 25c. This histogram shows the distribution of the conformational states present in the data at the hexameric level and reveals that some relative positions between the N domains are more favored than the others. The most present conformational states (with more than 4,000 particles per state) show five N domains in the same position and one in another position, or four adjacent N domains in one position and two in another position. On the contrary, the least probable conformations are the states where each adjacent N domain is in a different position, which accounts for less than 10 particles per state.

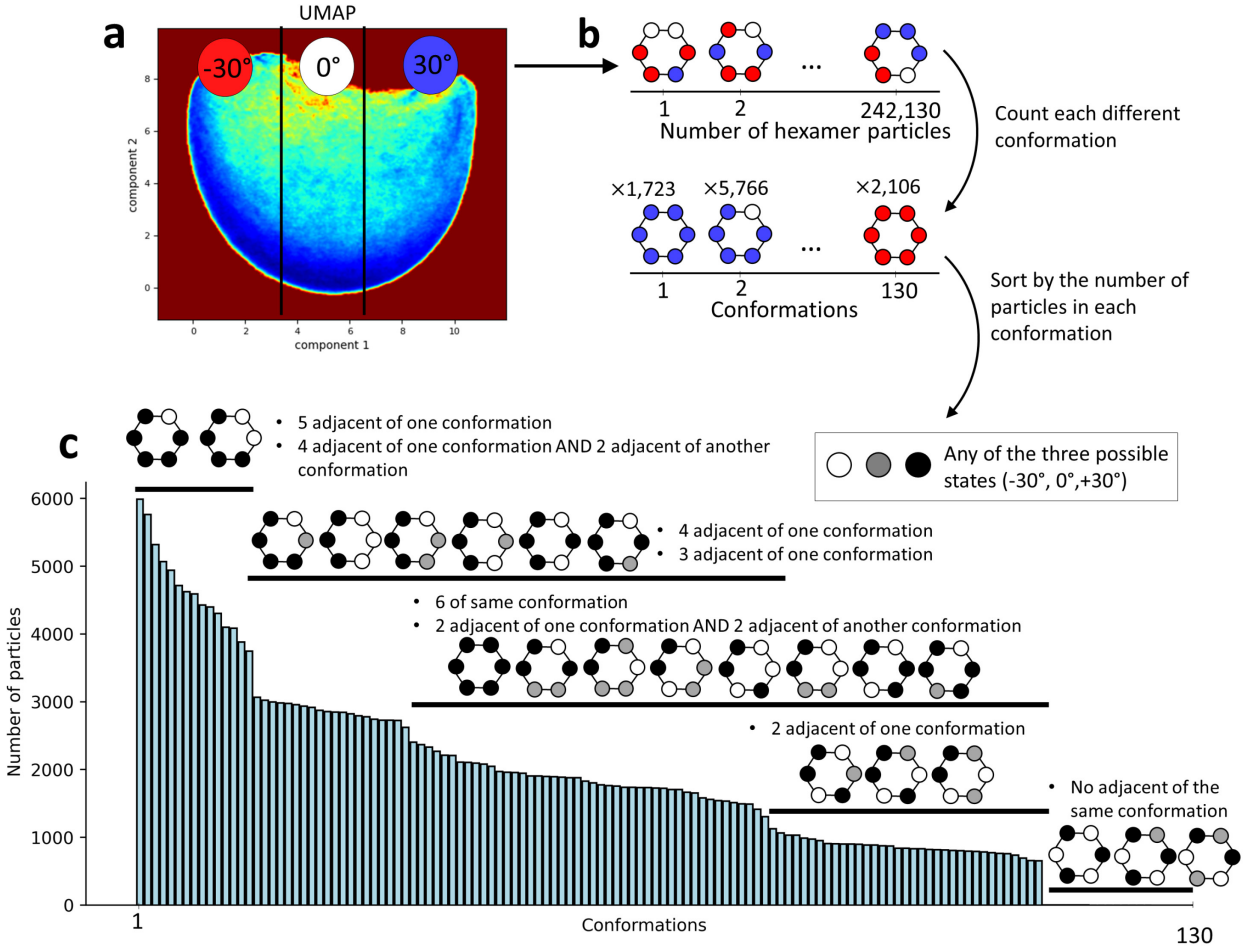

Figure 25: Determination of the hexameric conformations of p97-R155P based on the monomeric conformational states. (a) Conformational space at the monomeric level obtained by UMAP, split in three regions that are considered to be globally homogeneous regarding the N domain conformational state (-30°, 0°, +30°) (b) Schematic of the process done for reporting the conformation of each monomer to its

corresponding hexamer structure. (c) Histogram of all the possible hexameric conformational states given on the monomeric conformational states defined in (a).

IV.4. Discussion and Conclusion

In this chapter, I presented a new approach for analyzing continuous conformational variability in cryo-EM single particle images, MDSPACE, and showed its performance with synthetic and experimental data. The used experimental data are the public dataset of 80S ribosome-tRNA complexes (EMPIAR-10016) and an unpublished dataset of the ATPase p97 complexes studied in the team of I. Rouiller at the University of Melbourne. MDSPACE uses, to the best of our knowledge, the first 3D-to-2D MD-based flexible fitting method and estimates the conformational landscape of the cryo-EM data at atomic scale, using a single input atomic model.

MDSPACE obtains one atomic conformation per particle image by displacing the atomic coordinates of the given atomic model to align the model with the conformation in the given particle image. It uses normal-mode empowered MD simulations in the first iteration and principal-component empowered MD simulations in the remaining iterations. The normal modes speed up the MD-based fitting of the initial conformation to the target conformation in the particle image. The principal-components extracted from the ensemble conformational information at the end of each MDSPACE iteration refine the flexible fitting at the next MDSPACE iteration, especially for the images with difficult particle views and low SNR. The obtained conformational space allows obtaining atomic-resolution molecular movies and 3D reconstructions along its principal axes or along other directions (“trajectories”) identified through the densest regions, which may correspond to the most probable conformational transitions.

Each iteration of MDSPACE requires running one MD simulation per particle image. The MDSPACE code is MPI parallelized to process one particle image per CPU thread. However, MD simulations are still computationally expensive and their computational costs depends on multiple factors, such as simulation length, biomolecular size, and model type (all-atom or coarse-grained). Therefore, the large size of cryo-EM single particle datasets is a challenge. MDSAPCE tackle the speed issue of by using short-length simulations with coarse-grained models combined with NMMD to speed up the fitting.

Even with a short simulation and a coarse-grained model, the simulations were computationally costly. For instance, one simulation of 80S ribosome-tRNA complex takes one hour of CPU time on one core of Intel Xeon 6248 processor, meaning that one iteration of MDSPACE with the whole dataset (46,095 particles) would require about 46,000 CPU hours on one core of Intel Xeon 6248 (note that the corresponding wall-clock time is much shorter, as the processing is distributed over

multiple CPU cores in parallel). The set of 46,095 particles was processed using 16 nodes with two Intel Xeon 6248 processors per node (20 CPU cores per processor), leading to approximatively 3 days of computation on 640 CPU cores used without multithreading (i.e., 1 thread was used on each core). To further reduce the computing time in the large experimental data case, we used only a reduced dataset in the first iteration of MDSPACE (10,000 for 80S ribosome and 5,000 for p97). This subset of images ("training" set) was sufficient to estimate the global aspect of the conformational space, which was then refined in the next iteration using the entire dataset. This approach will be more investigated in the future, by further reducing the "training" data subset, as it may further increase the computational efficiency of MDSPACE.

The conformational space depends on the chosen initial conformation. Different initial conformations will result in different conformational spaces in the sense that the same conformational state will be visible in different regions of the different conformational spaces. However, the principal motions that can be extracted from these different conformational spaces will be similar, unless the initial conformations are totally unrelated to the particles in the given dataset. If several potential initial conformations are available and all of them are coherent with the dataset, any of them can be used as the initial conformation. In case of doubts, the conformation that is the closest to the global average conformation (reconstructed from all the images) should be chosen as the initial conformation.

Chapter V. MDTOMO: a NMMD-based method developed for analyzing continuous conformational variability in cryo electron subtomograms

This chapter presents MDTOMO, a novel approach for analyzing continuous conformational variability in subtomograms, which is based on extracting atomic models from subtomograms using NMMD (Chapter III). MDTOMO was published in 2023 [26] and this chapter describes the method and its performance with synthetic and experimental data. The Methods, Results, and Discussion-Conclusion sections of this chapter were extracted from the published manuscript [26] and adapted for inclusion in this PhD thesis manuscript.

V.1. Introduction

MDTOMO analyzes each individual subtomogram by performing NMMD-based flexible fitting to extract the atomic model that matches the conformational state in the subtomogram. MDTOMO allows interpreting datasets in the form of atomic conformational landscapes and their further analysis in terms of free-energy landscapes and animations of the atomic models along selected directions in these landscapes. Moreover, MDTOMO allows obtaining subtomogram averages from localized regions in the energy landscape. While other MD-based flexible fitting methods are used for fitting cryo-EM maps (e.g., NMMD method described in Chapter III) or SPA images (e.g., MDSPACE method described in Chapter IV), MDTOMO is the first method that uses MD-based flexible fitting in the context of analyzing subtomograms.

I tested MDTOMO on a synthetic dataset of an ABC exporter [4] and *in situ* dataset of a SARS-CoV-2 spike protein (S) [225]. The synthetic dataset provides a controlled environment where the results of the method can be compared with the ground-truth data, which allows assessing the method performance and robustness to noise and missing-wedge-induced deformations. Although MDTOMO does not currently correct for the missing wedge, it was able to produce meaningful results even for high amounts of noise. MDTOMO applied to *in situ* data indicates its ability to decipher the conformational landscape as a continuum of intermediate states.

V.2. Methods

MDTOMO is a method for extracting information about continuous conformational variability of macromolecular complexes from subtomograms. More precisely, by NMMD flexible fitting of an initial atomic model to each subtomogram, MDTOMO can estimate an atomic-scale model of

the conformational variability landscape, from which it is possible to obtain the energy landscape. Moreover, MDTOMO can visualize the conformational changes along any selected direction in the landscape, in terms of the displacements of atomic coordinates or the displacements of densities of the subtomogram averages calculated in localized regions along the selected direction.

NMMD flexibly displaces the initial atomic coordinates to match the underlying conformational state in each subtomogram. It extracts the conformational state and refines the rigid-body pre-alignment by estimating the rigid-body rearrangement that occurred during the simulation. Finally, the obtained fitted atomic models are projected onto a low-dimensional space (representing an essential conformational space), which allows visualizing conformational distribution and deciphering potential trajectories of conformational changes.

To speed-up MDTOMO analysis of large sets of subtomograms, MD simulations of NMMD can be performed using the Cα Gō model [147], which can be obtained with SMOG2 software [201], as was the case in the experiments shown here. The rigid-body pre-alignment of the subtomograms with the initial model prevents the simulation to get trapped into local minima and it is refined during the fitting. For the synthetic ABC exporter experiments, the rigid-body pre-alignment was obtained by STA based on the fast rotational matching algorithm (FRM) [226] implemented in Scipion [55]. Concerning the *in situ* dataset of SARS-CoV-2 spikes, we used the STA rigid-body alignment that had been obtained in the original work [225].

MDTOMO refines the initial particle pose using an estimate of the rigid-body rearrangement that occurred during the MD simulation. As in MDSPACE (see section IV.2.5), the estimation of the rigid-body motion during the MD simulation is obtained by optimizing the rotation and translation parameters to minimize the root mean square deviation (RMSD) between the initial conformation and the fitted conformation, using an optimization algorithm based on singular value decomposition in BioPython [199]. The refined particle pose is used in to calculate the subtomogram average of local regions of the energy landscape presented in the analysis of the SARS-CoV2 spike dataset.

MDTOMO is implemented as part of ContinuousFlex [198] available as a plugin of Scipion [55]. The Scipion back-end allows a user-friendly graphical interface for MDTOMO. Also, it allows integrating the data analysis results from most of the recent cryo-ET methods through the new ScipionTomo suite [80].

V.3. Results

In this section, I show the performance of MDTOMO using a set of synthetic subtomograms of an ABC exporter [4] and a set of experimental *in situ* subtomograms of a SARS-CoV-2 spike protein (S) [225].

V.3.1. Experiment with a synthetic dataset of ABC exporter

V.3.1.1. *Synthetic dataset of ABC exporter*

To test the performance of MDTOMO, we synthesized a dataset as close as possible to experimental cryo-ET conditions. The dataset simulates a highly heterogeneous cryo-ET dataset of a heterodimeric ABC exporter, TmrAB, along the substrate translocation cycle. During the translocation cycle, the ABC exporter changes its conformation from outward-facing to inward-facing, allowing the substrate to enter the intracellular cavity. ATP-binding first induces closing of the intracellular gate (an occluded conformation of the exporter, where both intracellular and extracellular gates are closed) and, then, opening of the extracellular gate (an outward-facing conformation, where the extracellular gate opens while the intracellular gate remains closed), which allows a release of the substrate in the extracellular environment. The structure of TmrAB in inward-facing, outward-facing, and occluded conformations were derived from cryo-EM maps [4] and are available in the Protein Data Bank (PDB) under the accession codes 6RAF, 6RAH, and 6RAK, respectively. Figure 26a shows the PDB: 6RAK structure of TmrAB in the occluded conformation while Figure 26b shows a sketch of the translocation cycle, with all three mentioned conformational states of TmrAB during the cycle.

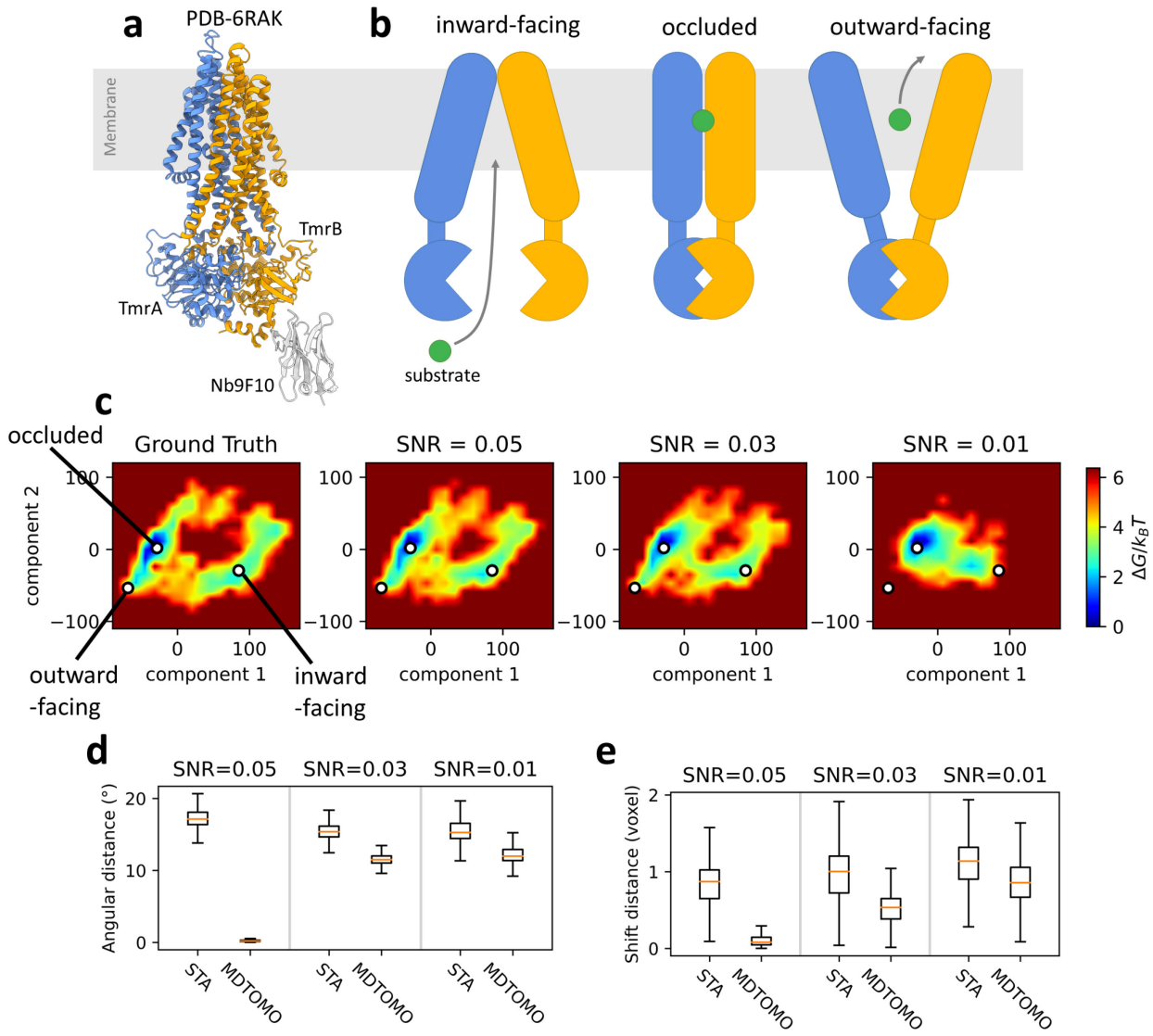

Figure 26: MDTOMO analysis of a synthetic dataset of TmrAB. (a) Structure of TmrAB in the occluded conformation. (b) Sketch of the substrate translocation process, from left to right: inward-facing, occluded, and outward-facing conformations of TmrAB. (c) Free-energy landscapes along the first two PCA components, from left to right: the ground-truth data used to synthesize the subtomograms and the MDTOMO results for the datasets with the SNR of 0.05, 0.03, and 0.01. The location of the three given PDB structures (6RAF, 6RAH, and 6RAK), used to target the MD simulations during the data synthesis, are also shown (white discs) in the free-energy landscapes. (d-e) Comparison of the initial and final rotational (d) and translational (e) alignment errors, for the three datasets (three SNR values).

The conformational heterogeneity was generated by standard MD-based flexible fitting (without normal mode acceleration) to obtain a dataset with realistic conformational variability, composed of a number of more populated conformational states (representing more stable states), close to the three mentioned states of TmrAB during the translocation cycle, and a number of less populated, intermediate states between these more stable states. The MD simulation, starting from

the occluded conformation, was directed towards one of the two other conformational states at a time (inward-facing or outward-facing), alternating between these two states, by choosing the biasing potential that forces the simulation to go to one of the two conformational states. The biasing potential was defined by a simulated EM map of the corresponding conformational state which was changed every 10 picoseconds, resulting in a simulated MD trajectory of the exporter alternating between the outward-facing and inward-facing conformations while passing through a large number of intermediate states, including the occluded conformation. The MD simulation was performed for 1 nanosecond using Cα-Gō model. One can notice that a 10-picosecond period may not be long enough for reaching the target conformation for the fitting, meaning that the fitted model at the end of each 10-picosecond MD-based flexible fitting run may be slightly different from the target conformation. Nevertheless, the dataset consisting of a continuum of conformational states visited during this MD simulation is rich and representative of the translocation cycle enough for performing the tests of the method that is proposed here. It should be noted that using the identical method for generating the synthetic dataset and for the dataset fitting (i.e., NMMD method and Cα-Gō model) would be a problem, as it could reduce the confidence in the results of fitting. That would be especially problematic if normal modes were used in both cases (generating the data and solving the problem), as they describe motions that would be easy to retrieve by NMMD. Therefore, we did not include normal modes (through NMMD) in the MD simulations used for generating the data. Nevertheless, the same MD model (Cα-Gō model) was used in the generation of data and problem solving, but we consider that this model allows enough degrees of freedom of the structure to avoid the bias.

To simulate the sub-tomograms, we followed a previously published method [24, 183]. We started by collecting 3000 atomic models from the simulated trajectory and we converted each model into a volume of size 100^3 voxels and a voxel size of $(2\ \text{Å})^3$ using a method based on scattering factors [205]. The obtained volumes were low-pass filtered at a cutoff frequency of 1/(6 Å) to take into account the impact of various imperfections (e.g., average radiation damage per image and tilt-series alignment errors). These effects were approximated here by limiting the maximum resolution of the subtomograms. However, more realistic synthetic data would account for accumulation of the radiation damage over the tilt series acquisition and would model the motion of the particles within the tilt series. Then, the volumes were rotated and shifted randomly, and projected with a tilt angle from -60° to +60° using an angular step of 2°. The tilt-series images were modulated by a CTF that corresponds to a 200 kV microscope, a defocus of -0.5 μm, and a spherical aberration of 2 mm. In practice, tilt series have large defocus variations within and between subtomograms. Here we simulate a constant defocus for all subtomograms but more realistic synthetic data would require to account for those defocus variations. Following a method

that adds Gaussian noise before and after the CTF[50], three sets of images were obtained corresponding to a SNR of 0.05, 0.03, and 0.01. The synthetic subtomograms were obtained from the simulated images using weighted back-projection in Fourier space[54]. The subtomograms were aligned using STA based on the FRM algorithm[226] in Scipion. It should be noted that the subtomogram synthesis uses a different method for converting atoms into density (scattering factors) than the NMMD fitting (3D Gaussian kernels on atomic positions), making the fitting more difficult.

V.3.1.2. *Recovery of the ground-truth conformational space*

We applied MDTOMO to each of the synthetic datasets (one for each SNR value) composed of 3000 subtomograms using 100-ps MD simulations starting from the PDB:6RAK model (occluded conformation). Figure 26c shows the free-energy landscape along the first two PCA components for the synthetic MD trajectory (ground truth) and the free-energy landscapes obtained with MDTOMO for each of the three datasets (subtomograms obtained from the images with the SNR of 0.05, 0.03, and 0.01). We observe that MDTOMO was able to recover the conformational landscape very accurately for the datasets with the SNR of 0.05 and 0.03. Indeed, Figure 26c shows that the densest regions of the conformational space (i.e., the regions of the lowest free energy) are those that are the closest to the three given PDB structures (6RAF, 6RAH, and 6RAK) that were used to target the MD simulations during the data synthesis. In the case of the dataset with the highest level of noise (SNR of 0.01), the outward-facing conformation was detected less accurately (some densities are missing for the outward-facing state) than the two other states (occluded and inward-facing).

V.3.1.3. *Recovery of the rigid-body alignments*

The final rigid-body alignment, obtained by MDTOMO, was compared with the ground truth rigid-body displacement (the rotations and shifts used to synthesize the data). The initial rigid-body alignment, done by STA, was also compared with the ground truth rigid-body displacement. The initial and final alignment errors are compared in Figure 26d and Figure 26e, which show respectively the rotational and translational errors for the three datasets (three SNR values). We observe that, for both rotational and translational parameters and each value of the SNR, MDTOMO allows a refinement of the rigid-body alignment (the final errors are smaller than the initial errors).

V.3.2. Experiment with a dataset of SARS-CoV-2 spike protein *in situ*

V.3.2.1. *In situ dataset of SARS-CoV-2 spike protein*

The SARS-CoV-2 spike (S) protein is a trimer of three identical glycoproteins (Figure 27a). It is involved in the cell infection as it mediates the viral entry by binding to the angiotensin-converting enzyme 2 (ACE2) receptor on the cell surface [227, 228]. The S-protein head includes three receptor binding domains (RBDs) that are located at the top of the spike and exists in different conformations on the viral capsid. We used MDTOMO to analyze a set of experimental subtomograms of the S protein *in situ* that had previously been obtained and analyzed using STA and 3D classification [225]. In that previous publication [225], the analysis of 20,830 subtomograms resulted in several classes, out of which two classes were further described. They correspond to two distinct conformational states, out of which one has three closed RBDs (fully closed state) and the other has one RBD opened (the classes containing 8,273 and 4,321 subtomograms, respectively). The tilt-series data are available in the EMPIAR database (accession code: EMPIAR-10453 [225]) and the subtomograms were provided by Beata Turoňová (Max Planck Institute of Biophysics, Frankfurt am Main, Germany).

MDTOMO was used to analyze the entire set of 20,830 subtomograms, which were first downsampled from their original size (256^3 voxels) to the size of 128^3 voxels and then pre-aligned using the rigid-body alignment parameters obtained by STA in the previous publication[225]. The initial conformation for the NMMD fitting with MDTOMO was PDB 6VXX, which corresponds to the fully closed state. The simulations included an initial energy minimization, followed by 100-ps NMMD simulations incorporating the 10 lowest frequency normal modes, with a force constant of 7000 kcal/mol.

V.3.2.2. *Continuous conformational landscape of the S protein*

The conformational variability was extracted by PCA from the set of atomic models obtained by MDTOMO. This PCA revealed multiple asymmetric conformations of the S protein. The plot in Figure 27b shows the variance described by each PCA component, and indicates that the first six principal components are responsible for more than 60% of the total variability. By exploring the conformational changes in different directions in the PCA space for the first six components, we could observe that the first three principal components (the most dominant motions) describe asymmetric opening and closing motions of the RBDs, whereas the next three principal components (less dominant motions) describe asymmetric up and down motions of the N-terminal domains (NTDs). Here, “asymmetric motion” means that each domain (RBD or NTD) can move independently from other domains.

Figure 27d shows a space determined by the first three principal components. Each of the three black arrows in this space, denoted by "RBD-A open", "RBD-B open", and "RBD-C open" (Figure 27d), shows a direction of motion that mainly describes a single RBD opening (one RBD is opening while the other two RBDs remain closed), for each of the three RBDs denoted in Figure 27a by "RBD-A", "RBD-B" and "RBD-C", respectively. Figure 27f shows the conformational change along the "RBD-A open" direction (opening of the RBD-A domain). In other directions of this space (Figure 27d) we could observe other conformational changes. In particular, when going from low-energy regions to high-energy regions in other directions than those indicated by black arrows (Figure 27d) we could observe opening of more than one RBD. For instance, the red arrow denoted by "All RBDs open" in Figure 27d indicates a direction along which all three RBDs are opening together, and Figure 27g shows the conformational changes along this direction (from the conformation with all three RBDs closed to a conformation with all three RBDs open).

Although the original study (which used STA and discrete classification) revealed only a fully closed state and a state with a single RBD open[225], simultaneous opening of more than one RBD has been discussed in the literature[229]. However, considering the symmetry (C3) of the conformation with all three RBDs fully closed and our results showing that each of the three RBDs can undergo opening motion independently of the other two RBDs, we explored the possibility that a large number of the conformational models obtained by MDTOMO could be identical up to a rotation by 120° or 240° around the C3 pseudo-symmetry axis, which could have caused a difficulty in well separating the principal components by the PCA. Therefore, we designed a method to reduce the variability in order to simplify the PCA. This method is based on aligning the conformational models obtained by MDTOMO with respect to 5 conformations along the "RBD-A open" direction (Figure 27d), which describe opening of the RBD-A domain. Three out of the five conformational models used for this alignment are shown in Figure 27f. This method, which simplifies conformational variability analysis of the S protein, is explained next.

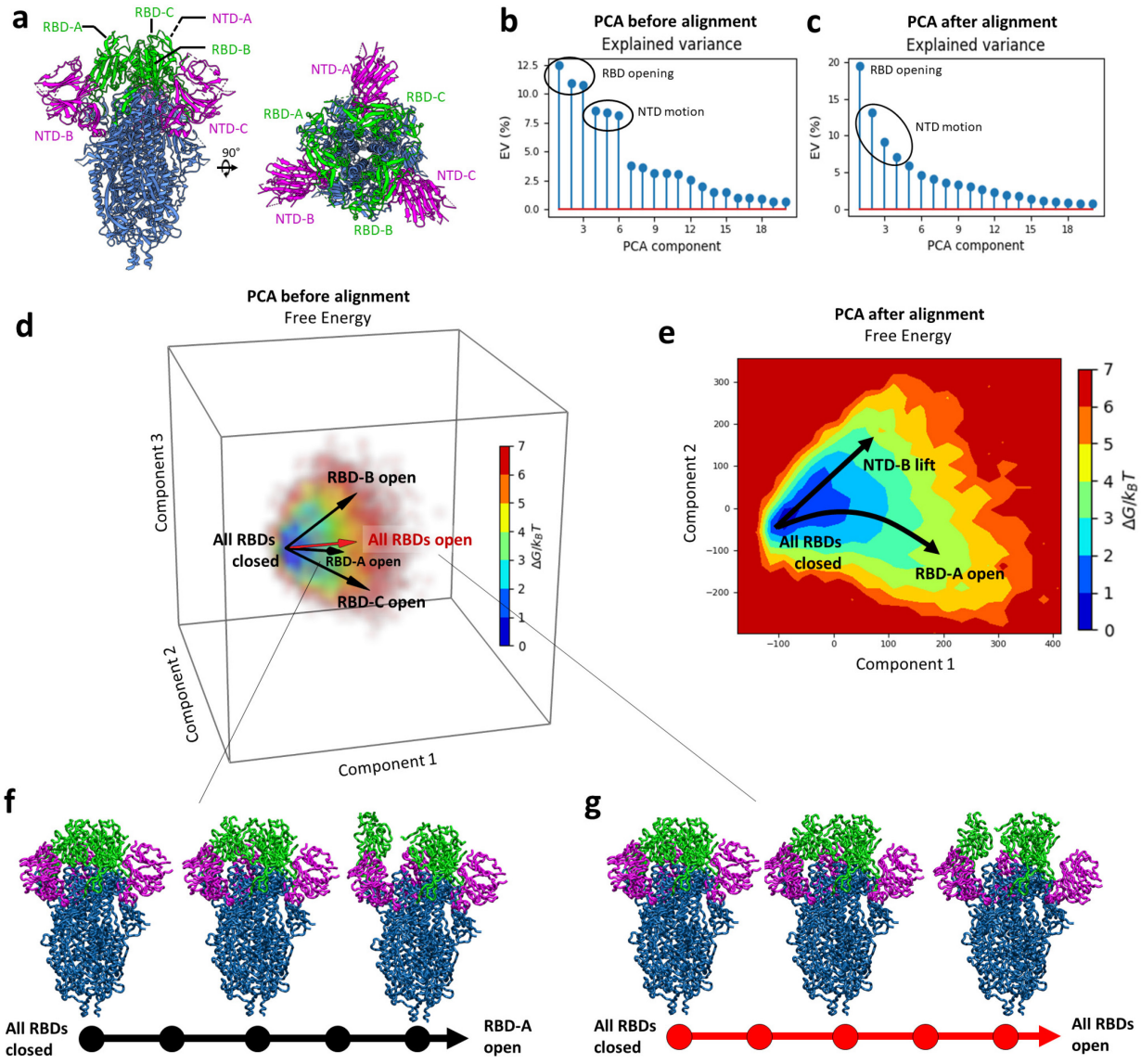

Figure 27: PCA analysis of MDTOMO results with a cryo-ET dataset of SARS-CoV-2 S protein. (a) Structure of the S protein (PDB 6VXX). (b-c) Explained variance of the PCA space obtained before (b) and after (c) reducing the number of principal components to describe opening of RBDs. (d) Free-energy landscape determined by the first three principal components, obtained before reducing the number of components to describe opening of RBDs. (e) Free-energy landscape determined by the first and second principal components obtained after reducing the number of components to describe opening of RBDs. The arrows show the directions associated with the RBD and NTD motions discussed here. (f) RBD-A opening trajectory following the direction "RBD-A open" shown in (d). (g) Trajectory of opening of all three RBDs together following the direction "All RBDs open" shown in (d).

V.3.2.3. *Simplifying PCA of conformational variability based on pseudo-symmetries of conformations*

As introduced in the previous subsection, we reduced the number of principal components describing the variability due to RBD motions by aligning the conformational states obtained by

MDTOMO with the states of a selected trajectory of a single RBD opening extracted from the initial PCA space (Figure 27d), which we here refer to as reference trajectory. The reference trajectory was selected as a five-state trajectory that corresponds to the most dominant single-RBD motion in the space determined by the first three PCA components, which is the opening of RBD-A (Figure 27d,f). The alignment was done by rotating the MDTOMO-derived conformations around the C3 pseudo-symmetry axis (here the z-axis) by multiples of 120° and matching the rotated conformations with the states on the reference trajectory. Note that we chose to split the trajectory in five states as this number of states was large enough for a comprehensive description of the different intermediate states.

More precisely, the set of 20,830 MDTOMO-derived conformational models was extended three times, by applying $n \times 120°$ rotation around the z-axis with $n = \{0,1,2\}$, resulting in 62,490 models (20,830 triplets, each triplet composed of the model rotated by $n \times 120°$). Each of the three models in a triplet was compared by RMSD with the five models from the reference trajectory resulting in 15 RMSD values (five RMSD values per model of a triplet). The lowest value of the 15 RMSD values was identified and the associated rotation factor n was collected to determine the angle for the realignment of the MDTOMO-derived conformational model.

A new PCA was performed on the set of models obtained after the alignment, resulting in a new conformational variability space, described mostly by two principal components corresponding to a motion of RBD-A opening and a motion of NTD-B lifting, as shown in Figure 27e. Based on Figure 27e, it can be noted that the number of principal components necessary to describe RBD motions was reduced from 3 to 1. Also, in the new PCA space, the NTD motions are described by lower principal components (Figure 27c) than in the initial PCA space (components 2-4 in the new space vs. components 4-6 in the initial space).

The set of MDTOMO-derived conformational models aligned with the selected trajectory of single RBD opening (the trajectory shown in Figure 27f) was further analyzed using UMAP, which allowed a better separation of the different conformational states and identification of more populated distinct conformations. UMAP was obtained with a reduction dimension of 5, an Euclidean metric for computing the distance in the ambient space, a neighborhood size of 15 and a minimum distance between points of 0.1.

V.3.2.4. *UMAP representation of the conformational landscape simplified using pseudo-symmetries of conformations*

The conformational landscape obtained after reducing the number of principal components to describe opening of RBDs (Figure 27e) shows a motion of RBD-A opening and a motion of NTD-B lifting. The set of aligned MDTOMO-derived conformational models was analyzed with UMAP

to better separate different conformational states. The first two UMAP components are presented in terms of free-energy in Figure 28a. The conformational models and the density maps shown in Figure 28 are the conformational-model and subtomogram averages calculated from the four most populated regions (regions of lowest energy), manually selected and depicted in Figure 28a.

The most populated region, designated as (b) in Figure 28a, corresponds to a conformer with all 3 RBDs closed. The corresponding conformational-model average and subtomogram average are shown in Figure 28b. The most distant low-density region from the region (b) along the first UMAP component, designated as (c) in Figure 28a, corresponds to a conformer with a single open RBD, i.e., RBD-A (Figure 28c). The central region, designated as (d) in Figure 28a, corresponds to intermediate states of a single RBD opening (Figure 28d). The bottom region along the second UMAP component, designated as (e) in Figure 28a, corresponds to a conformer with the NTB-B lifted (Figure 28e).

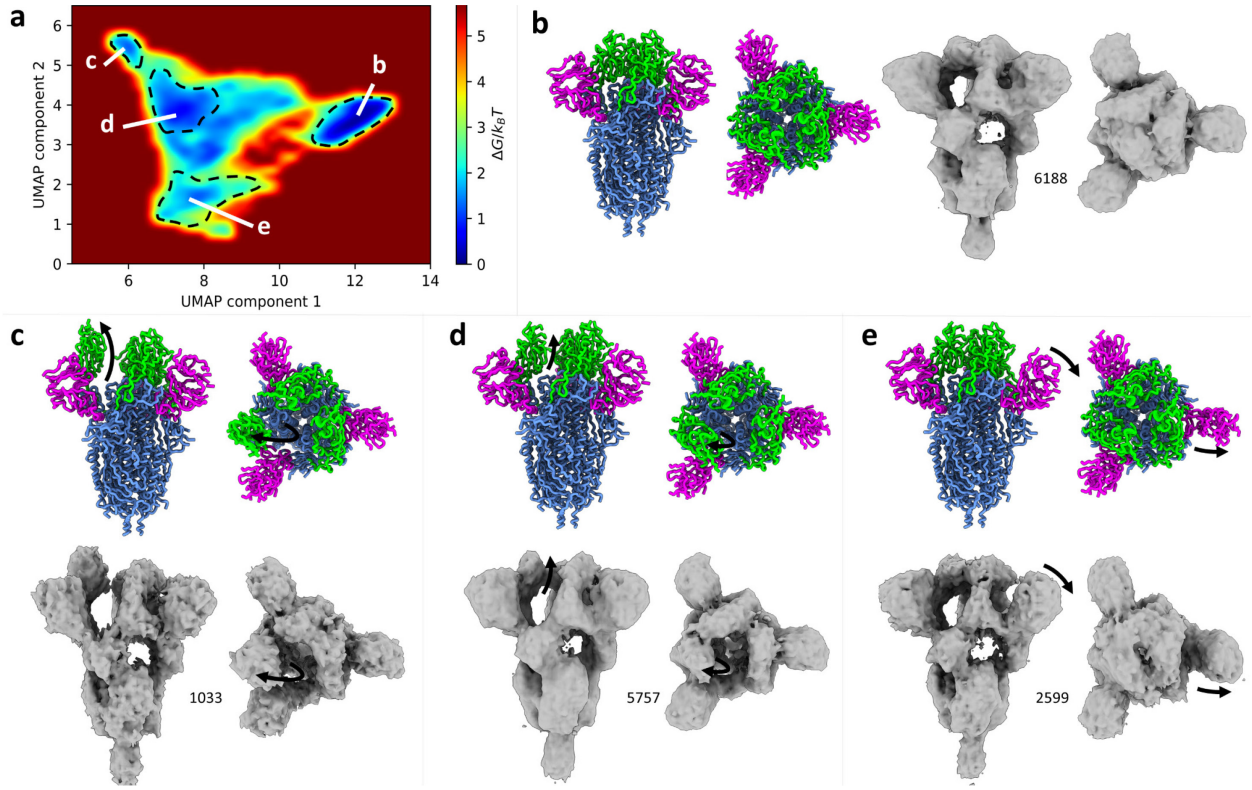

Figure 28 : UMAP analysis of MDTOMO results with a cryo-ET dataset of SARS-CoV-2 S protein, after simplifying description of the variability due to opening of RBDs. (a) Free-energy landscape along the UMAP components 1 and 2. (b-e) Conformational-model and sub-tomogram averages from four manually selected regions in the free-energy landscape (denoted by b, c, d, and e). The colored model corresponds to the average of the conformational models in the selected region. The density maps are the averages of the subtomograms from the selected regions in the energy landscape. The number of the averaged particles is indicated near each map. Black arrows represent the conformational change from the

initial structure used for the fitting within MDTOMO (PDB-6VXX, which is a conformation with all RBD closed).

V.4. Discussion and Conclusion

In this chapter, I presented MDTOMO, a novel approach to explore continuous conformational landscapes of biomolecular complexes by analyzing subtomogram datasets. MDTOMO uses efficient, coarse-grained and normal-mode-empowered MD simulations to extract the conformational states from subtomograms, which are typically averaged out during classical STA, and allows reconstructing low-dimensional (usually 2D or 3D) energy profiles of the structures from the experimental subtomogram data. MDTOMO is the first method for analyzing subtomograms using MD-based flexible fitting.

MDTOMO starts by extracting an atomic model from each individual subtomogram, by NMMD flexible fitting of a given reference atomic model into the subtomogram. The extracted atomic models are then projected onto a low-dimensional space, which can be considered as an essential conformational space. In this space, closer points represent more similar conformations, and denser regions of points represent the regions containing more likely or more stable (low-energy) conformations in the given dataset. The density of points in this space can be interpreted in terms of the free energy of the complex or by computing and analyzing the averages of the subtomograms and their associated atomic models in the regions of the space characterized by the lowest free energy of the complex. I showed that MDTOMO produces expected results using synthetic data (ABC exporter). Also, I showed that the results produced using experimental SARS-CoV-2 spike protein data are coherent with previously published results regarding the existence of conformations with all three RBDs closed and those with a single RBD open[225]. Additionally, they revealed gradual conformational transitions related to the opening of single RBDs, a lifting motion of a single NTB, and additional conformations with more than one fully or partially open RBD.

Although this study gives two examples of application of MDTOMO, the method can be applied on various proteins and biological macromolecular complexes of various size, with the only limitation of the computation time, which scales with the number of atoms included in the model and the size of the subtomograms. For instance, I previously showed that MDSPACE [25] was able to successfully study large complexes as 80S ribosome using NMMD-based fitting. Therefore, MDTOMO could be applied to such large complexes as well. Note that MDTOMO can also be applied on cryo-ET datasets with smaller voxel size and larger volumes (the subtomograms of SARS-CoV-2 spikes were 128^3 voxels of 2.65 $Å^3$).

MDTOMO was performed by applying a single NMMD fitting compared to the iterative process proposed in MDSPACE that uses PCA to refine the conformational space at each iteration. Here I this iterative process was not useful as the data analyzed are volumes instead of images in SPA. Nevertheless, this iterative process is implemented for MDTOMO and might be useful in the future for datasets with particularly low SNR.

PCA was used in the analysis of the synthetic dataset of ABC exporter. PCA is a method for dimensionality reduction by decomposing the data variability into a linear combination of principal component vectors (see section II.3.1.1). For the experiments of SARS-CoV2 spikes, we used UMAP for non-linear dimensionally reduction (see section II.3.1.2), as it allowed a better separation of the different conformational populations than PCA. Especially, it allowed a clear clustering of the two conformational states that were observed in the original study using STA (close RBD and one-open RBD, b and c in Figure 28) and several other intermediate metastable states (d and e in Figure 28). However, as UMAP extracts non-linear features from the data, it produces a conformational landscape with non-linear parametrization in which the interpretation of energy densities must be done carefully. In this conformational embedding, equivalent distances are not necessarily equivalent conformational difference. For instance, the distance between the centers of regions c and d and between the centers of regions d and b in the UMAP embedding in Figure 28 are 2 and 4.5, respectively, whereas these two distances in terms of the conformational difference both correspond to a rotation of the RBD of approximatively 30°. It also implies that local minima observed in the energy landscape are not necessarily meaningful stable conformational states (e.g., many local minima are visible in the central region of Figure 28a), but rather close intermediate states grouped together during the embedding. To prevent such misinterpretations, future developments will include the improvement of the conformational landscape interpretation using methods that could preserve local point densities [230].

In MDTOMO, the correlation coefficient was used as the metric for fitting the conformation in subtomograms. In standard methods, the alignment and classification of subtomograms uses correlation coefficient combined with a compensation for the missing wedge (e.g., constrained correlation coefficient[114], wedge-masked differences[231]) because of the sensitivity of the correlation coefficient to anisotropic deformations of subtomograms caused by the missing wedge. MDTOMO uses correlation-driven MD-based fitting without compensating for the missing wedge. The use of missing wedge-compensated metric such as constrained correlation coefficient (CCC) to drive the MD fitting was considered. However, prototype implementation of the gradient of the CCC required for the energy calculation shown to be too computationally expensive for running MD fitting and would requires further developments in order to be efficient. Nevertheless, MDTOMO was able to produce expected results with synthetic and experimental data without

compensating for the missing wedge, which can be explained by the fact that the NMMD flexible alignment in MDTOMO is not only driven by the force that depends on the correlation coefficient (derivatives of the correlation coefficient with respect to the atomic positions), but also by the classical MD force field (equation (II-7)). The weight given to the potential defined by the correlation coefficient, with respect to the classical MD potential, is a free parameter (the so-called force constant; *k* equation (II-7)) whose value is set so as to drive the fitting toward the conformation in the subtomogram by preserving structural constraints, which reduces the risk of overfitting the noise and the missing-wedge-induced deformations in the subtomograms.

Compared to the majority of other methods for analyzing conformational flexibility in cryo-EM (see section I.5.2), MDTOMO uses an available atomic structure, whose coordinates are displaced (by performing NMMD simulation) to match the conformations in the subtomograms. Such atomic structure is not available for all datasets, but those datasets for which an atomic structure is available can efficiently be analyzed using MDTOMO. Indeed, MDTOMO incorporates structural constraints into the analysis via MD simulation, which is particularly important for analyzing cryo-ET subtomograms that suffer from noise and missing-wedge-induced deformations, as these structural constraints limit the motions to those that are feasible (underlying biochemical processes of the complex) and, therefore, they reduce the risk of overfitting mentioned previously. Also, the use of an available atomic structure in MDTOMO allows obtaining a conformational landscape in which the conformational transitions can be visualized directly on the atomic structure, contrary to the majority of other methods that visualize the conformational transitions on density volumes.

Another method that can use an available atomic structure to analyze conformational flexibility in cryo-ET datasets is HEMNMA-3D[183]. More precisely, HEMNMA-3D can use either an available atomic structure or an available density volume (e.g., subtomogram average or cryo-EM map)[183], which is flexibly fitted into subtomograms by solely using normal mode analysis. Normal modes are useful for their simplicity and computational efficiency. However, they tend to induce distortions of the structure in case of large amplitudes of conformational changes. Also, they restrict the conformational exploration to a set of collective motions that must be defined in advance and that may be insufficient to describe the full range of conformational changes. MDTOMO overcomes these limitations by employing coarse-grained MD simulations to allow a full exploration of the conformational landscape without inducing structural distortions, and by empowering these MD simulations using normal modes to make simulations fast.

The MDTOMO approach requires running one MD simulation per subtomogram, which can be computationally expensive. The MDTOMO software is parallelized so that multiple simulations

can run on multiple CPU cores simultaneously, with one simulation running only on one CPU core. Furthermore, the use of coarse-grained simulation coupled with normal-mode-based acceleration allows performing MDTOMO in a reasonable amount of time. The total execution time on 8 nodes of 2x20 Intel 6248 2.50 GHz for the two presented datasets is shown in Table 3. However, it is important to note that the computational cost scales with the molecular weight of the complex (the higher the number of atoms, the higher the cost of the MD simulation).

Table 3: Execution time of MDTOMO for the two datasets analyzed in this article.

Complex	**Weight**	**100 ps-simulation MDTOMO on Intel 6248 2.50 GHz**	
		one CPU core – one particle	*320 CPU cores – full dataset*
ABC exporter	150.85 kDa	372.3 seconds	0.9 hours (3000 particles)
SARS-CoV2 spike	438.26 kDa	987.5 seconds	17.8 hours (20830 particles)

Chapter VI. Discussion, Conclusion and Future Directions

Throughout this thesis, I worked on novel hybrid methods that combine MD simulations with image analysis methods to extract information on biomolecular conformational variability from cryo-EM and cryo-ET data. This led to the development of three competitive methods, NMMD for flexible fitting of cryo-EM maps, and MDSPACE and MDTOMO for the analysis of conformational heterogeneity of single particle cryo-EM and cryo-ET datasets, respectively. In this final chapter, I start by discussing the newly developed methods and describe how they compare with other methods. Then, I conclude and give an overlook of the possible directions for future work.

VI.1. Discussion

VI.1.1. NMMD as a new hybrid method for structure modeling

Compared to methods solely based on MD, NMMD incorporates Normal Mode Analysis (NMA) into the simulation to increase the speed of the global-dynamics atomic displacements allowing a better efficiency of the fitting than standard MD-based methods. In NMMD, a set of low-frequency normal modes, representing the most collective motions (the motions that move the largest number of atoms synergistically), is incorporated in the MD integration scheme to encourage the MD simulation to move along the normal mode directions. The normal modes allow a fast displacement along the global dynamics and the MD simulation ensures the structural stability of the system and corrects for the distortions that can be induced by normal modes. The incorporation of normal modes typically reduces the length of the simulation required to fit large-amplitude conformational changes and therefore allows reducing the overall simulation length and the computational cost of the simulation. NMMD has been used to fit atomic models into cryo-EM maps and proved to be, on average, 40% faster than the flexible fitting solely based on MD simulation [38].

A previous attempt to combine MD and NMA, MDeNM-EMfit (see section II.4.4) was based on an alternation between a stochastic estimation of a linear combination of normal modes (the combination that moves the structure in the direction of the CC increase) and a short MD simulation initiated with this normal-mode combination [182]. This normal-mode combination was updated periodically every 2 ps and each update required a new calculation of normal modes and a new stochastic estimation of their amplitudes. NMMD, is the first method that combines the

displacements based on MD and NMA via their simultaneous integration. In NMMD, the integration of normal-mode amplitudes at each time step is based on an analytical expression for the gradient of the potential energy, which ensures accurate and fast results. Furthermore, our NMMD experiments have shown that normal modes can remain constant during the fitting while normal-mode amplitudes are updated at each step with no delay, which saves computing time as multiple NMA runs are not required.

VI.1.2. MDSPACE and MDTOMO as new hybrid methods for conformational heterogeneity analysis

VI.1.2.1. *Reference-based approaches*

One of the key features of MDSPACE and MDTOMO compared to other methods for analyzing conformational heterogeneity resides in the incorporation of prior structural information, through a known atomic model. This atomic model is used by simulation methods to predict feasible conformational displacements. This model is often referred to as a reference model or initial conformation.

When analyzing particle images or subtomograms, a method that directly estimates the conformation from the particle data without prior structural knowledge is prone to errors due to a high amount of noise in the data (for instance, the conformation that best describes the data can be a non-feasible conformation with broken covalent bonds, superimposed atoms, etc.). There are multiple approaches to prevent this kind of error. An evident approach would be to restrict the conformational exploration to a set of known feasible conformations. However, the set of known conformations should be continuous, as the approach would otherwise be a discrete-state approach, similar to multireference classification. In the approaches proposed in this thesis work, the continuous distribution of conformational states is generated by simulating molecular mechanics of a given atomic model. This approach can be referred to as "reference-based" as it depends on the chosen input atomic model (reference).

In most continuous-state methods, the predicted conformations are represented by density maps [18, 20, 22-24, 27, 101, 105, 116-118, 126], and only a few use atomic models representations [19, 28, 29, 123, 124, 183], including MDSPACE and MDTOMO. Among the methods based on density representation, some methods describe the possible conformational states by applying smooth deformations on a reference density map to preserve the structure locally [22, 24, 27]. Although these methods are not strictly "reference-based", they describe the conformational space by elastic deformation around a reference structure. This reference density map is typically initiated with an EM map (e.g. obtained by 3D reconstruction without heterogeneity analysis) that is refined during the analysis. For instance, Tomoflow [24] deforms a reference map using a vector

field of deformation (optical flows) and attempts to preserve the structure by forcing the optical flows to respect smoothness locally. Another example is 3DVA [22], where the conformational space is described by a sum of a reference map and a weighted sum of principal volumes that are shared by all the particles. Finally, another example is 3DFlex [27], which tries to locally conserve the "mass" of the density of a reference map by penalizing the conformational exploration by a regularization function that enforces a local rigidity of the density.

In all of these examples, a reference map is used to model the possible conformations. However, such methods assume that a smooth deformation or locally rigid deformation of a reference map produces a physico-chemically valid conformation. Although this can be true within a short range of deformation amplitudes around the reference, avoiding non-feasible or energetically improbable conformations is generally not guaranteed.

A better reference than a density map is an atomic model, if it can be obtained for the system under study. Currently, for a large number of experimental cryo-EM studies, the related atomic models are already available (e.g. from X-ray crystallography or from another cryo-EM study), or can be predicted with increasing accuracy with AI-based methods such as AlphaFold [31]. An atomic model provides more than the 3D atomic positions in the Cartesian space (comparable to using a density map as reference), but also gives information about the atomic interactions (e.g., for each atom, an element, a charge, a set of covalent bonds), which can be used to define a forcefield and perform simulations, as seen in Chapter II. Such simulations allow producing a large set of feasible conformations that can be confronted with the experimental data to infer the underlying conformational space (as in MDTOMO and MDSPACE).

The first method that used an available atomic structure to analyze conformational flexibility in cryo-EM and cryo-ET datasets is HEMNMA [19]. Contrary to MDSPACE and MDTOMO, HEMNMA (and the cryo-ET extension and HEMNMA-3D [183]) uses solely normal modes for the conformational exploration, which are useful for their simplicity and computational efficiency but tend to induce distortions of the structure in case of large amplitudes of conformational changes. Also, HEMNMA restricts the conformational exploration to a set of collective motions described by a subset of normal modes, which must be defined in advance and may be insufficient to describe the full range of conformational changes. It is worth noting that HEMNMA can also use a reference density map (instead of an atomic model). In the case of a reference density map, HEMNMA converts this reference map into a collection of 3D Gaussian functions whose centers are used (instead of the atomic centers) to calculate normal modes. It has been shown that normal modes calculated from density maps give similar results to those calculated from atomic models, especially if the number of Gaussian functions is comparable to the number of Cα atoms [232].

Normal modes are calculated using ENM of the system (described by atomic or Gaussian centers), which means that they are calculated based on the shape (3D) of the system. NMA produces global deformations (by a linear combination of harmonic-oscillator motions) and discard the atomic interaction information (atomic-level degrees of freedom). In that sense, the principle of the normal-mode-based methods like HEMNMA is similar to the principle of the linear variability decomposition methods like 3DVA (both types of methods find a linear displacement in the conformational space around a reference, and normal modes in HEMNMA are similar to principal component volumes in 3DVA).

To fully exploit the structural information of the atomic model for analyzing the conformational heterogeneity, other methods use MD simulations including cryoBIFE [28], Ensemble Reweighting [29] and the methods developed in this thesis, MDSPACE and MDTOMO. Contrary to NMA, MD simulations account for atomic-level degrees of freedom, which allows modeling more complex dynamics and alleviates the limitations of NMA.

Recently, an atomic model was used as a reference in deep learning-based methods [123, 124]. Here the model is trained to learn the atomic positions of a given particle using a regularization function that incorporates a force field term to preserve the structure given a reference atomic model. These methods are very promising as they could potentially alleviate the computational cost using MD simulations while using complex dynamics with atomic-level degrees of freedom. However, they were only tested on small systems with synthetic data and showed difficulties to train the model especially when combining the estimation of the conformation with the estimation of the particle pose (which are known to be difficult to decouple). Furthermore, they are still limited to a very basic force field model by considering only the lengths of bonds between residues along the backbone (probably because these methods are still in an early development stage) [123, 124].

However, there are limitations to using a prior reference in the analysis. The major limitation is when this reference differs compositionally from the studied system. This can be neglected when only a few residues are modified or when a small ligand is present or absent. However, the analysis gets uncertain when the reference has larger modifications compared to the data such as a truncated domain, absence or presence of a cofactor, or different assembly states. Therefore, reference-based methods are sensitive to compositional heterogeneity, i.e., in the context where the studied system is not only conformationally heterogeneous but also contains compositional heterogeneity. Nevertheless, in the 80S ribosome case studied in section IV.3.2, MDSPACE could capture the binding of the second tRNA (even if this tRNA was not present in the reference model) because it is associated with a conformational change of the 80S ribosome (inter-subunit rotation). In some

cases of datasets, MDSPACE could help to identify a disassembly of large subunits of the complexes because the images containing single subunits may look like outliers in the resulting low-dimensional conformational space (under the condition that the entire complex is used as the reference and that the fraction of the entire complexes is much larger than the fraction of the single subunits). However, if such strong compositional heterogeneity is suspected, it is recommended to use classical discrete-classification methods before using MDSPACE or MDTOMO to filter out the compositional heterogeneity.

Also, I found difficulties fitting datasets where the initial reference is dramatically far from the conformational populations of the data. For the p97 data, the position and movement of the N-domains could not be detected with the same accuracy using a reference where all the N-domains were co-planar to the D1 ring, as the overall conformational population of the data was closer to the conformation where all the N-domain are "up". It can be explained as it would require a longer simulation to fit conformations that are more different, which increases the chance of getting trapped in local minima on the way. Nevertheless, in the same dataset, we were still able to detect the presence of a small set of proteins (estimated at 2%) adopting a conformation very different from the main population (RMSD = 5.3 Å). As a gain, it would be worth assessing how divergent the conformations can be from the reference model to be detected.

VI.1.2.1. *Atomic-scale conformational space*

Although more and more methods are developed to extract the conformational space from cryo-EM/ET datasets (including MDSPACE and MDTOMO), the interpretability of the produced landscapes is questionable. Depending on the method, the conformational space can represent an abstract space that is not guaranteed to deliver interpretable features. In that sense, MDSPACE and MDTOMO have the advantage, over most of the other methods, as they produce atomic-scale conformational spaces.

In continuous-state approaches, the processing is typically performed in the following order i) the method assigns a conformation to each particle ii) the conformational space is reduced to a low-dimensional space with dimensionality reduction methods for visualization purposes and the regions of interest in the low-dimensional space are interpreted by calculating their 3D structure representation. However, the available methods differ in the type of representation used (density map or atomic model) in these two stages. In this section, I discuss how these can affect the interpretability of the produced conformational space.

First, the methods differ in the way they assign a conformation to each particle, which can be by assigning a set of fitted atomic models (e.g., MDTOMO and MDSPACE), a set of normal mode amplitudes (e.g. HEMNMA [19], HEMNMA-3D (Harastani et al, 2021), DeepHEMNMA

(Hamitouche and Jonic, 2022), see section II.5.1), a set of coefficients of a linear combination of principal volumes (e.g. 3DVA [22], see section I.5.2.1), a latent variable encoded in a neural network (e.g. cryoDRGN [23], 3DFlex [27], see section I.5.2.1), a set of vectors of the deformation field (Tomoflow [24], see section I.5.2.2). Therefore, some representations are directly based on an atomic structure (MDSPACE and MDTOMO), some can retrieve an atomic structure (HEMNMA) or a density map (3DVA, Tomoflow), and others are purely abstract (e.g. deep learning methods). Note that a few methods such as cryoBIFE [28] or Ensemble Reweighting [29] produce an energy landscape but do not directly assign a conformation to each particle, which does not allow to map a certain region to a 3D density reconstructed from the data (3D reconstruction or subtomogram average).

Having a physical representation of each particle, such as an atomic model is a great advantage as it allows to perform preliminary analysis before projecting on a low-dimensional space. For example, in the experiments done with the cryo-ET data of the SARS-CoV2 spikes with MDTOMO, the conformational space first obtained was incorrectly attributing three dimensions to the opening/closing of one RBD, due to the C3 symmetry of the structure (see section V.3.2). However, as the atomic structure was available for each particle, I was able to correct the conformational space by performing a posterior alignment on the fitted structures in order to merge the three dimensions of RBD opening into one. This kind of operation would be impossible to perform with an abstract representation such as a latent variable in deep learning methods. For other representations using a density map (e.g. with optical flows), although it would be theoretically possible, it would be more challenging than simply aligning atomic structures. Another example is the analysis of p97 with MDSPACE, in which case the conformational space at the level of the hexamers could not be interpreted directly due to the high variability of the six N-domains that engendered a complex multi-dimensional conformational space. Here, the atomic structure representation was used to extract the variability on a local region of the space (one monomer) and reconstruct a local conformational space. The local conformational spaces obtained for the six monomers were then merged. This allowed to finely observe the conformational change of the N-domain and even estimate the conformational distribution at the hexametric level. Such analysis could only be performed with methods with an atomic model representation.

In the second step, once the conformation is estimated for each particle, most of the time the dimension of the resulting conformational space is still too high, and dimensionality reduction methods such as PCA or UMAP are applied (note that the manifold embedding method directly applies this step from the particles [18]). The obtained low-dimensional space is then interpreted by retrieving (or calculating) the 3D structure corresponding to points or regions of interest in the landscape. Here, the methods differ by the type of representation of the distribution of

conformation. Most methods map the conformational space to 3D densities, for instance, a point in the latent variable in deep learning methods is passed to the decoder network that learned to convert the point to a density map. Another example is in Tomoflow, the point in the PCA space corresponds to an optical-flow that is applied to deform the subtomogram average and obtain the density map of that conformation. In MDSPACE and MDTOMO, the conformational space directly maps to an atomic model which has the great advantage of visualizing 3D trajectories at the atomic level. On the contrary, methods using density representation only allow visualizing trajectories on the density-map level. Although the resulting densities could be analyzed a posteriori by flexible fitting methods to interpret the atomic positions in the density, that operation would be highly time-consuming as the conformational space can be composed of thousands of different conformations (which would result in applying thousands of times a flexible fitting, similarly as in MDSPACE or MDTOMO).

Besides the atomic-level representation of the conformational space, MDSPACE and MDTOMO also produce density maps of selected regions in the conformational space directly from the data (3D reconstructions in MDSPACE and subtomogram averages in MDTOMO). Comparing the density maps and the fitted atomic models from the same region of the conformational space is useful for validating the fitted models.

VI.1.2.2. *Computational efficiency*

As previously discussed, the use of MD simulations has several advantages in the context of analyzing conformational heterogeneity in cryo-EM and cryo-ET datasets. However, it comes with a large computational cost. The computational cost of a method is an important factor for a broader use of the method in the EM field. Therefore, in MDSPACE and MDTOMO, I adopted a strategy to maximally reduce the computational cost of the simulation, which is based on i) combining MD simulations with normal modes through NMMD; ii) using coarse-grained models for the simulations; and iii) optimizing the length and force constant of the simulations.

The first point of this speed-up strategy is the combination of MD simulations with normal modes through NMMD, which I showed to accelerate the fitting of EM maps on average by 40%, when compared to the flexible fitting method that uses only MD simulations and that served as the basis for the development of NMMD. I showed that NMMD insures simulating accurate dynamics and that it can even be more robust to noise in some cases. Although NMMD is much more computationally expensive than the fitting methods that solely use NMA, it allows conserving a high number of degrees of freedom, which are necessary to maintain the integrity of the structure and to avoid distortions induced by NMA.

The second point of the speed-up strategy is the use of a coarse-grained model. In the presented experiments, a Cα Gō model was employed for the simulations. This model has several advantages, it is fast as it reduces the representation by Cα atoms (one atom per residue) and it is simple to parametrize. However, the Gō-like models are approximating the non-bonded interaction by prior knowledge of the initial conformation (see section II.1.2) and therefore the model is biased towards the reference. Furthermore, representing the atomic model only by the Cα trace is a high level of coarse-graining, which models only the dynamics of the backbone and discards the side-chain interactions. Although it is possible to reverse the coarse-graining operation [233] (approximate the all-atom model from the Cα trace) to recover the all-atom model after performing the simulation, the recovered all-atom model is still less realistic than the model that accounts for the side-chain dynamics during the simulations. In MDSPACE and MDTOMO, the Cα Gō model can be easily replaced by another model depending on the available computational resources, such as an all-atom model (CHARMM is already implemented) or a more realistic coarse-grained model such as MARTINI [141]. Contrary to Gō-like models, physically-based models such as MARTINI, are not biased by the initial conformation as they translate the classical all-atom force field to the coarse-grained model. Furthermore, MARTINI uses a representation that incorporates side-chains, therefore allowing the modeling of more realistic environments than Cα Gō. However, MARTINI has a more detailed representation of the structure, which induces larger computational costs of simulations.

However, it is important to keep in mind that the choice of a proper force field model in an unconstrained MD simulation (classical MD simulation without biasing potential) is crucial as the goal of such simulation is to predict conformational events as they would happen in the real world, and therefore, a bias by the force field model may induce inaccurate dynamics which bias the interpretation of the results. In the context of flexible fitting, the simulation is by nature “biased” as the main driving force is the biasing potential (the experimental data). In that case, the role of the force field model is more to maintain structure during the fitting, while the biasing potential drives the prediction of the conformational state. Therefore, the choice of the force field model is less prone to bias the results.

The last point of the speed-up strategy is the optimization of the length and force constant of simulations. The two parameters should be adjusted to ensure the fastest convergence to the correct conformation, which may be challenging. Their values depend on the factors such as the molecular system or the chosen force field. A high value of the force constant speeds up the fitting, as it helps the simulation to go faster through energy barriers, but the value of the force constant must be tuned with caution as too high values result in structural distortions. Concerning the length of the simulation, if the simulation is too short, the fitting gets closer but does not have enough time to

converge. On the contrary, if the simulation is too long, the computational cost increases without improving the results. Currently, there is no satisfying measure for assessing automatically the convergence. In the experiments described for MDSPACE and MDTOMO, the force constant and the simulation length were determined based on preliminary experiments on a small number of images or subtomograms (around ten) with a trial-and-error process. First, the optimal force constant was determined by running long simulations (up to 1 ns) with different values of the force constant and selecting the value that induces the fastest conformational changes (the fastest increase in the *CC*) without inducing structural distortions. Then, the optimal simulation length for the obtained optimal value of the force constant was determined as the time after which the *CC* does not increase anymore in all the simulations.

VI.1.3. Validating the methods

In the experiments presented, the developed methods were tested on synthetic and experimental datasets. Synthetic data are helpful for method development as they provide a controlled environment where the results of the method can be compared with the ground-truth data. This allows to validate the method with quantitative measurements. However, when validating the method on experimental data, it gets challenging to find quantitative metrics to evaluate the performance of the methods.

In the context of experimental data, heterogeneity analysis methods (particularly for discrete-state approaches) typically evaluate the performance of the analysis by the resolution of the reconstructed maps. The spatial resolution of the maps is typically calculated by the Fourier Shell Correlation (FSC) [234]. The principle is to split the set of particle images of a given map into two halves. The two halves are reconstructed separately and the correlation between the two maps is calculated for each spatial frequency. The closest the two maps are, the higher the correlation for high-frequencies and therefore the higher the resolution. However, the spatial resolution alone is not suited to assess the ability of a method to distinguish different conformations. For instance, methods that distinguish more different conformational states might not improve the resolution compared to methods that find a few high-resolution states.

In most continuous-state approaches, quantifying the different conformations present in a conformational space is typically done by visually identifying regions of interest in the conformational space and comparing these regions by visualizing the corresponding 3D structures. For instance, in MDSPACE and MDTOMO, a graphical interface is available to semi-automatically draw trajectories and clusters in the conformational space and visualize the corresponding atomic models and EM maps. Note that this approach can get particularly difficult when the conformational space is larger than 2 dimensions. However, there is currently a lack of

standardized metrics for the conformational heterogeneity that could be used for comparison with other datasets in different experimental conditions, or for comparison with other methods on the same dataset. The development of new quantitative measurements is therefore of high interest for the field. In [125], the authors proposed an idea for a new metric, assuming that the method can obtain a density volume for each particle (in the case of MDSPACE and MDTOMO, that would be an atomic model converted to a volume), the volumes are classified with hierarchical clustering by calculating the FSC between each volume pair. The resulting hierarchy tree is a function of the spatial frequency, so that, for a given spatial frequency, the clustering regroups only the volumes that are indistinguishable at that resolution.

As the methods developed in this thesis work are producing atomic models (compared to most other methods that are based on densities), several metrics are available to assess the quality of the produced models. MolProbity [196] is a structure validation tool that was employed, for instance, in NMMD experiments to quantitatively compare the produce models between NMMD and MD simulations (see section III.3). This measurement could be employed in the future in MDTOMO and MDSPACE to evaluate the structure in different locations of the conformation space.

Another way to assess the quality of the produced atomic models (that was used in the experiments presented in this thesis work) is to assess the agreement between the fitted atomic models and the reconstructed EM map in each location of the conformational space. Supposedly, the fitted models should agree well with the EM maps, however, it might be difficult to obtain this EM map for conformations that are shared with a small number of particles as the reconstruction gets noisy (e.g. less than 1,000 particles).

Furthermore, as the field of heterogeneity analysis is very active, more and more methods become available and there is an urgent need for proper benchmarks between methods. Unfortunately, there is not currently a common database for conformational heterogeneity such as Imagenet [236] for the field of computer vision. Recently a cryo-EM heterogeneity challenge was proposed with two heterogeneous datasets on which the method developers are invited to test their own methods (www.simonsfoundation.org/heterogeneity-in-cryo-electron-microscopy). This type of initiative can be beneficial to the field in the future, as comparing methods may show on which applied system each method is bests suited.

VI.1.4. Biological importance of conformational studies and possible applications

The developed methods provide a better insight into the different conformational states that could be present in cryo-EM datasets. This is a true advantage for understanding the dynamics of

macromolecules and identifying the mechanism of how a protein functions. An example of application of such conformational studies is drug discovery, for instance, understanding how a chemical compound binds to a complex [1, 237-239]. In structure-based drug design (SBDD), by analyzing the structure of a target macromolecule, new ligands can be discovered *in silico* that binds with high affinity to the target and modify its function (e.g. inhibit the target protein). This analysis is possible by analyzing the atomic-resolution structure of the target and calculating a set of binding measurements at different locations on the structure with a large database of small molecules. In that sense, having access to a wide range of conformational states at atomic-scale (e.g. produced by MDSPACE and MDTOMO) is highly valuable. For instance, different conformational states may form different binding sites, which as a result could allow identifying a wider range of potential ligands. Besides, many active sites are observed to be structurally flexible [240]. This flexibility is often required to allow the ligand to enter the binding site. Therefore, a better identification of the flexibility of the binding site can be helpful for SBDD methods to properly estimate the protein-ligand interactions. Furthermore, conformational studies of protein-ligand interactions could be done by comparing a conformational study with and without a ligand. The produced conformational space could be compared to investigate if that ligand actually induces a conformational change in the target macromolecule and affect its function.

VI.2. Conclusion

Cryo-EM is an experimental technique particularly suited to studying the conformational variability of macromolecules. The particle images in cryo-EM are snapshots of the studied complex in different conformations, offering a unique potential for conformational studies. However, extracting the conformational states from the individual cryo-EM particles is challenging due to the complexity of disentangling rotational and translational heterogeneity from the conformational heterogeneity, considering the high level of noise of the recorded images. Traditional approaches for heterogeneity analysis are classifying the particles into discrete classes that are globally homogeneous in terms of conformation. The major limitations of this approach are that a few classes are most often not sufficient to describe the continuous variability of macromolecules and that most of the conformationally heterogeneous particles are filtered out (as they are difficult to classify) which results in a largely incomplete overview of the conformational behavior of the studied system. Contrary to single particle cryo-EM, cryo-ET allows analyzing the molecular complexes in their cellular environment. Subtomograms of these complexes extracted from cryo-ET tomograms can be analyzed similarly to single particle images to extract the conformational variability of the complexes in the cellular environment.

This thesis aimed at developing new image processing methods for continuous conformational analysis of heterogeneous cryo-EM and cryo-ET datasets through the use of MD simulations. The work resulted in the development of three methods. The first method is NMMD, which efficiently combines MD simulations with NMA for flexible fitting of cryo-EM maps. NMMD allowed to speed up MD-based flexible fitting and it was exploited for application in heterogeneity analysis. This led to the other two contributions of this thesis, namely MDSPACE and MDTOMO methods, which are the first methods using flexible fitting that combines MD simulation and normal modes for analyzing large continuous conformational heterogeneity in single particle cryo-EM and cryo-ET, respectively. MDSPACE and MDTOMO are competitive methods that have the advantage of exploiting prior structural information to produce an atomic-scale conformational landscape. The methods were tested on synthetic and experimental datasets of various systems and shown to be effective at exploring the conformational landscapes of highly heterogeneous datasets.

VI.3. Future directions

Through the years, the field of cryo-EM image processing has been constantly improving in terms of quality and efficiency. The rise of GPUs combined with notable algorithm improvements led to efficient image processing pipelines treating millions of particles in just a few hours on single workstations. The methods presented in this thesis for heterogeneity analysis, MDTOMO and MDSPACE, are novel algorithms bringing MD simulations into cryo-EM image processing. Despite their advantages, MD-based methods have the drawback of being highly computationally expensive and MDTOMO and MDSPACE still require accessing a large amount of CPU resources. In that case, supervised deep learning methods can bring a large speed up, for instance, HEMNMA [19] (another hybrid method for continuous conformational analysis that uses NMA instead of NMMD), was recently extended with a supervised deep neural network trained to learn conformational and alignment parameters produced by HEMNMA [102]. The deep learning-based extension, DeepHEMNMA [102], greatly speeds up the computation. A similar supervised approach could be highly beneficial to MDTOMO and MDSPACE, as MD simulations are even more costly than NMA. Yet, the ability of such a method to learn conformations produced by MD simulations is still to be validated.

Many recent cryo-EM developments were focusing on conformational studies *in situ* with cryo-ET. Indeed, the potential for conformational studies in cryo-ET through methods such as MDTOMO is of great biological interest. Recently, method developments in cryo-ET have been focusing on refining the tilt-series alignment together with the subtomogram alignment in a process referred to as “per tilt per particle” [16, 17, 81, 88]. This process accounts for misalignments of the tilt-series that limit the resolution of the reconstructed tomogram. Therefore, rather than

considering the subtomogram fixed as in STA, the per-tilt-per-particle refinement refines the alignment parameters at the level of subtilt-series (the sub-images of the tilt-series corresponding to a subtomogram). Although this process is typically performed posterior to STA, recent methods attempted to avoid aligning subtomograms (which is computationally expensive) but directly aligning subtilt-series [241]. Subtilt-series were used instead of subtomograms for structural heterogeneity analysis in cryo-ET data in a very recent deep learning-based TomoDRGN [126]. However, TomoDRGN does not refine the alignment together with the structure. In the future, MDTOMO could be extended to analyze continuous conformational variability from subtilt-series by replacing the 3D-to-3D flexible fitting with a 3D-to-2D (such as in MDSPACE). Such a method would have the potential to refine the tilt-series alignment together with the conformation estimation, with other advantages such as removing the need to account for the missing-wedge.

Finally, the methods developed in this thesis are hybrid methods that exploit available structural information from previous structural studies. More and more of these structures are solved each year with cryo-EM which represent an enormous amount of data that increase our knowledge of macromolecular structures. Therefore, some systems have already several available models in different conformations that can be used as additional prior information for conformational studies. For MDSPACE and MDTOMO, that could mean using multiple atomic models as multiple initializations for the MD simulations. That would have the advantages of being more robust to particle orientation and provide an effective measure of the convergence, for instance, by assessing that the fitting from multiple initializations converges to the same structure.

References

1. Renaud, J.-P., et al., *Cryo-EM in drug discovery: achievements, limitations and prospects.* Nature reviews Drug discovery, 2018. **17**(7): p. 471-492.
2. Kühlbrandt, W., *The resolution revolution.* Science, 2014. **343**(6178): p. 1443-1444.
3. Banerjee, S., et al., *2.3 A resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition.* Science, 2016. **351**(6275): p. 871-5.
4. Hofmann, S., et al., *Conformation space of a heterodimeric ABC exporter under turnover conditions.* Nature, 2019. **571**(7766): p. 580-583.
5. Khatter, H., et al., *Structure of the human 80S ribosome.* Nature, 2015. **520**(7549): p. 640-5.
6. Hoffmann, P.C., et al., *Structures of the eukaryotic ribosome and its translational states in situ.* Nature Communications, 2022. **13**(1): p. 7435.
7. Coureux, P.-D., et al., *Cryo-EM study of an archaeal 30S initiation complex gives insights into evolution of translation initiation.* Communications biology, 2020. **3**(1): p. 58.
8. Singer, A. and F.J. Sigworth, *Computational methods for single-particle electron cryomicroscopy.* Annual review of biomedical data science, 2020. **3**: p. 163-190.
9. Serna, M., *Hands on methods for high resolution cryo-electron microscopy structures of heterogeneous macromolecular complexes.* Frontiers in molecular biosciences, 2019. **6**: p. 33.
10. Sorzano, C.O.S., et al., *Survey of the analysis of continuous conformational variability of biological macromolecules by electron microscopy.* Acta crystallographica. Section F, Structural biology communications, 2019. **75**(Pt 1): p. 19-32.
11. Tang, W.S., et al., *Conformational heterogeneity and probability distributions from single-particle cryo-electron microscopy.* Current Opinion in Structural Biology, 2023. **81**: p. 102626.
12. Scheres, S.H., *RELION: implementation of a Bayesian approach to cryo-EM structure determination.* J Struct Biol, 2012. **180**(3): p. 519-30.
13. Punjani, A., et al., *cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.* Nature methods, 2017. **14**(3): p. 290-296.
14. Lyumkis, D., et al., *Likelihood-based classification of cryo-EM images using FREALIGN.* J Struct Biol, 2013. **183**(3): p. 377-88.
15. Bharat, T.A. and S.H. Scheres, *Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION.* Nature protocols, 2016. **11**(11): p. 2054-2065.

16. Castaño-Díez, D., M. Kudryashev, and H. Stahlberg, *Dynamo Catalogue: Geometrical tools and data management for particle picking in subtomogram averaging of cryo-electron tomograms.* Journal of structural biology, 2017. **197**(2): p. 135-144.

17. Himes, B.A. and P. Zhang, *emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging.* Nature methods, 2018. **15**(11): p. 955-961.

18. Dashti, A., et al., *Trajectories of the ribosome as a Brownian nanomachine.* Proc Natl Acad Sci U S A, 2014. **111**(49): p. 17492-7.

19. Jin, Q., et al., *Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes.* Structure, 2014. **22**(3): p. 496-506.

20. Tagare, H.D., et al., *Directly reconstructing principal components of heterogeneous particles from cryo-EM images.* Journal of structural biology, 2015. **191**(2): p. 245-262.

21. Lederman, R.R., J. Andén, and A. Singer, *Hyper-molecules: on the representation and recovery of dynamical structures for applications in flexible macro-molecules in cryo-EM.* Inverse Problems, 2020. **36**(4): p. 044005.

22. Punjani, A. and D.J. Fleet, *3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM.* Journal of Structural Biology, 2021. **213**(2): p. 107702.

23. Zhong, E.D., et al., *CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks.* Nature Methods, 2021. **18**(2): p. 176-185.

24. Harastani, M., et al., *TomoFlow: Analysis of continuous conformational variability of macromolecules in cryogenic subtomograms based on 3D dense optical flow.* Journal of molecular biology, 2022. **434**(2): p. 167381.

25. Vuillemot, R., et al., *MDSPACE: Extracting Continuous Conformational Landscapes from Cryo-EM Single Particle Datasets Using 3D-to-2D Flexible Fitting based on Molecular Dynamics Simulation.* J Mol Biol, 2023: p. 167951.

26. Vuillemot, R., I. Rouiller, and S. Jonić, *MDTOMO method for continuous conformational variability analysis in cryo electron subtomograms based on molecular dynamics simulations.* Scientific Reports, 2023. **13**(1): p. 10596.

27. Punjani, A. and D.J. Fleet, *3DFlex: determining structure and motion of flexible proteins from cryo-EM.* Nature Methods, 2023: p. 1-11.

28. Giraldo-Barreto, J., et al., *A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments.* Scientific Reports, 2021. **11**(1): p. 13657.

29. Tang, W.S., et al., *Ensemble Reweighting Using Cryo-EM Particle Images.* The Journal of Physical Chemistry B, 2023.

30. Chen, M. and S.J. Ludtke, *Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM.* Nature Methods, 2021. **18**(8): p. 930-936.

31. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. 596(7873): p. 583-589.

32. Allen, M.P., *Introduction to molecular dynamics simulation.* Computational soft matter: from synthetic polymers to proteins, 2004. 23(1): p. 1-28.

33. Mirzadeh, A., et al., *In silico prediction, characterization, docking studies and molecular dynamics simulation of human p97 in complex with p37 cofactor.* BMC Molecular and Cell Biology, 2022. 23(1): p. 1-12.

34. Orzechowski, M. and F. Tama, *Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations.* Biophysical journal, 2008. 95(12): p. 5692-5705.

35. Trabuco, L.G., et al., *Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics.* Structure, 2008. 16(5): p. 673-683.

36. Miyashita, O., et al., *Flexible fitting to cryo-EM density map using ensemble molecular dynamics simulations.* Journal of computational chemistry, 2017. 38(16): p. 1447-1461.

37. Kulik, M., T. Mori, and Y. Sugita, *Multi-scale flexible fitting of proteins to cryo-EM density maps at medium resolution.* Frontiers in molecular biosciences, 2021. 8: p. 631854.

38. Vuillemot, R., et al., *NMMD: Efficient cryo-EM flexible fitting based on simultaneous Normal Mode and Molecular Dynamics atomic displacements.* Journal of Molecular Biology, 2022. 434(7): p. 167483.

39. Bernal, J.D. and D. Crowfoot, *X-ray photographs of crystalline pepsin.* Nature, 1934. 133(3369): p. 794-795.

40. Adrian, M., et al., *Cryo-electron microscopy of viruses.* Nature, 1984. 308(5954): p. 32-36.

41. Noble, A.J., et al., *Routine single particle CryoEM sample and grid characterization by tomography.* Elife, 2018. 7: p. e34257.

42. Frank, J., *Averaging of low exposure electron micrographs of non-periodic objects*, in *Single-Particle Cryo-Electron Microscopy: The Path Toward Atomic Resolution: Selected Papers of Joachim Frank with Commentaries*. 1975, World Scientific. p. 69-72.

43. Saxton, W. and J. Frank, *Motif detection in quantum noise-limited electron micrographs by cross-correlation.* Ultramicroscopy, 1976. 2: p. 219-227.

44. Henderson, R., et al., *Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy.* Journal of molecular biology, 1990. 213(4): p. 899-929.

45. Wu, S., J.-P. Armache, and Y. Cheng, *Single-particle cryo-EM data acquisition by using direct electron detection camera.* Microscopy, 2016. 65(1): p. 35-41.

46. Brilot, A.F., et al., *Beam-induced motion of vitrified specimen on holey carbon film.* Journal of structural biology, 2012. 177(3): p. 630-637.

47. Scheres, S.H., et al., *Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization.* Nat Methods, 2007. 4(1): p. 27-9.

48. Liao, M., et al., *Structure of the TRPV1 ion channel determined by electron cryo-microscopy.* Nature, 2013. **504**(7478): p. 107-112.

49. Velázquez-Muriel, J., et al., *A method for estimating the CTF in electron microscopy based on ARMA models and parameter adjustment.* Ultramicroscopy, 2003. **96**(1): p. 17-35.

50. Sorzano, C.O., et al., *Fast, robust, and accurate determination of transmission electron microscopy contrast transfer function.* J Struct Biol, 2007. **160**(2): p. 249-62.

51. Rohou, A. and N. Grigorieff, *CTFFIND4: Fast and accurate defocus estimation from electron micrographs.* Journal of structural biology, 2015. **192**(2): p. 216-221.

52. Zhang, K., *Gctf: Real-time CTF determination and correction.* Journal of structural biology, 2016. **193**(1): p. 1-12.

53. Turoňová, B., et al., *Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4 Å.* Journal of structural biology, 2017. **199**(3): p. 187-195.

54. Strelak, D., et al., *Advances in Xmipp for Cryo–Electron Microscopy: From Xmipp to Scipion.* Molecules, 2021. **26**(20): p. 6224.

55. de la Rosa-Trevín, J.M., et al., *Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy.* Journal of Structural Biology, 2016. **195**(1): p. 93-99.

56. Tang, G., et al., *EMAN2: an extensible image processing suite for electron microscopy.* Journal of structural biology, 2007. **157**(1): p. 38-46.

57. Zheng, S.Q., et al., *MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy.* Nature methods, 2017. **14**(4): p. 331-332.

58. Zivanov, J., T. Nakane, and S.H. Scheres, *A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis.* IUCrJ, 2019. **6**(1): p. 5-17.

59. Huang, Z. and P.A. Penczek, *Application of template matching technique to particle detection in electron micrographs.* Journal of Structural Biology, 2004. **145**(1-2): p. 29-40.

60. Scheres, S.H., *Semi-automated selection of cryo-EM particles in RELION-1.3.* Journal of structural biology, 2015. **189**(2): p. 114-122.

61. Wang, F., et al., *DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM.* Journal of structural biology, 2016. **195**(3): p. 325-336.

62. Wagner, T., et al., *SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM.* Communications biology, 2019. **2**(1): p. 218.

63. Yang, Z., et al., *Iterative stable alignment and clustering of 2D transmission electron microscope images.* Structure, 2012. **20**(2): p. 237-247.

64. Jakobi, A.J., M. Wilmanns, and C. Sachse, *Model-based local density sharpening of cryo-EM maps.* Elife, 2017. **6**: p. e27131.

65. Sanchez-Garcia, R., et al., *DeepEMhancer: a deep learning solution for cryo-EM volume post-processing.* Communications biology, 2021. **4**(1): p. 874.

66. Frenz, B., et al., *RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps.* Nature methods, 2017. **14**(8): p. 797-800.

67. Terwilliger, T.C., et al., *Cryo-EM map interpretation and protein model-building using iterative map segmentation.* Protein science, 2020. **29**(1): p. 87-99.

68. Lopéz-Blanco, J.R. and P. Chacón, *iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates.* Journal of structural biology, 2013. **184**(2): p. 261-270.

69. Mahamid, J., et al., *Visualizing the molecular sociology at the HeLa cell nuclear periphery.* Science, 2016. **351**(6276): p. 969-972.

70. Turoňová, B., et al., *Benchmarking tomographic acquisition schemes for high-resolution structural biology.* Nature Communications, 2020. **11**(1): p. 876.

71. Hagen, W.J., W. Wan, and J.A. Briggs, *Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging.* Journal of structural biology, 2017. **197**(2): p. 191-198.

72. Turk, M. and W. Baumeister, *The promise and the challenges of cryo-electron tomography.* FEBS letters, 2020. **594**(20): p. 3243-3261.

73. Dubochet, J., et al., *Cryo-electron microscopy of vitrified specimens.* Quarterly reviews of biophysics, 1988. **21**(2): p. 129-228.

74. Wagner, F.R., et al., *Preparing samples from whole cells using focused-ion-beam milling for cryo-electron tomography.* Nature protocols, 2020. **15**(6): p. 2041-2070.

75. McDonald, K., *A review of high-pressure freezing preparation techniques for correlative light and electron microscopy of the same cells and tissues.* Journal of microscopy, 2009. **235**(3): p. 273-281.

76. Střelák, D., et al., *FlexAlign: An accurate and fast algorithm for movie alignment in cryo-electron microscopy.* Electronics, 2020. **9**(6): p. 1040.

77. Zheng, S., et al., *AreTomo: An integrated software package for automated marker-free, motion-corrected cryo-electron tomographic alignment and reconstruction.* Journal of Structural Biology: X, 2022. **6**: p. 100068.

78. Mastronarde, D.N. and S.R. Held, *Automated tilt series alignment and tomographic reconstruction in IMOD.* Journal of structural biology, 2017. **197**(2): p. 102-113.

79. Kremer, J.R., D.N. Mastronarde, and J.R. McIntosh, *Computer visualization of three-dimensional image data using IMOD.* Journal of structural biology, 1996. **116**(1): p. 71-76.

80. Jiménez de la Morena, J., et al., *ScipionTomo: Towards cryo-electron tomography software integration, reproducibility, and validation.* Journal of Structural Biology, 2022. **214**(3): p. 107872.

81. Chen, M., et al., *A complete data processing workflow for cryo-ET and subtomogram averaging.* Nature methods, 2019. **16**(11): p. 1161-1168.

82. Tegunov, D. and P. Cramer, *Real-time cryo-electron microscopy data preprocessing with Warp.* Nature methods, 2019. **16**(11): p. 1146-1152.

83. Hrabe, T., et al., *PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis.* Journal of structural biology, 2012. **178**(2): p. 177-188.

84. Chen, M., et al., *Convolutional neural networks for automated annotation of cellular cryo-electron tomograms.* Nature methods, 2017. **14**(10): p. 983-985.

85. Moebel, E., et al., *Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms.* Nature methods, 2021. **18**(11): p. 1386-1394.

86. Martinez-Sanchez, A., et al., *Robust membrane detection based on tensor voting for electron tomography.* Journal of structural biology, 2014. **186**(1): p. 49-61.

87. Scheres, S.H.W., et al., *Averaging of Electron Subtomograms and Random Conical Tilt Reconstructions through Likelihood Optimization.* Structure, 2009. **17**(12): p. 1563-1572.

88. Pyle, E. and G. Zanetti, *Current data processing strategies for cryo-electron tomography and subtomogram averaging.* Biochemical Journal, 2021. **478**(10): p. 1827-1845.

89. Svidritskiy, E., et al., *Structures of yeast 80S ribosome-tRNA complexes in the rotated and nonrotated conformations.* Structure, 2014. **22**(8): p. 1210-1218.

90. Zhou, A., et al., *Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM.* Elife, 2015. **4**: p. e10180.

91. Bai, X.C., et al., *Sampling the conformational space of the catalytic subunit of human gamma-secretase.* Elife, 2015. **4**.

92. Abeyrathne, P.D., et al., *Ensemble cryo-EM uncovers inchworm-like translocation of a viral IRES through the ribosome.* Elife, 2016. **5**.

93. Nakane, T., et al., *Single-particle cryo-EM at atomic resolution.* Nature, 2020. **587**(7832): p. 152-156.

94. Kato, K., et al., *High-resolution cryo-EM structure of photosystem II reveals damage from high-dose electron beams.* Communications Biology, 2021. **4**(1): p. 382.

95. Klaholz, B.P., A.G. Myasnikov, and M. Van Heel, *Visualization of release factor 3 on the ribosome during termination of protein synthesis.* Nature, 2004. **427**(6977): p. 862-5.

96. Simonetti, A., et al., *Structure of the 30S translation initiation complex.* Nature, 2008. **455**(7211): p. 416-20.

97. Klaholz, B.P., *Structure sorting of multiple macromolecular states in heterogeneous cryo-EM samples by 3D multivariate statistical analysis.* Open J Stat, 2015. **5**(7): p. 820-836.

98. Loerke, J., J. Giesebrecht, and C.M. Spahn, *Multiparticle cryo-EM of ribosomes.* Methods Enzymol, 2010. **483**: p. 161-77.

99. Fischer, N., et al., *Structure of the E. coli ribosome-EF-Tu complex at <3 A resolution by Cs-corrected cryo-EM.* Nature, 2015. **520**(7548): p. 567-70.

100. Penczek, P.A., J. Frank, and C.M. Spahn, *A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation.* Journal of structural biology, 2006. **154**(2): p. 184-194.

101. Dashti, A., et al., *Retrieving functional pathways of biomolecules from single-particle snapshots.* Nat Commun, 2020. **11**(1): p. 4734.

102. Hamitouche, I. and S. Jonic, *DeepHEMNMA: ResNet-based hybrid analysis of continuous conformational heterogeneity in cryo-EM single particle images.* Front Mol Biosci, 2022. **9**: p. 965645.

103. Haselbach, D., et al., *Structure and conformational dynamics of the human spliceosomal Bact complex.* Cell, 2018. **172**(3): p. 454-464. e11.

104. Moscovich, A., et al., *Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes.* Inverse Problems, 2020. **36**(2): p. 024003.

105. Katsevich, E., A. Katsevich, and A. Singer, *Covariance Matrix Estimation for the Cryo-EM Heterogeneity Problem.* SIAM J Imaging Sci, 2015. **8**(1): p. 126-185.

106. Jonić, S., *Computational methods for analyzing conformational variability of macromolecular complexes from cryo-electron microscopy images.* Curr Opin Struct Biol, 2017. **43**: p. 114-121.

107. Valle, M., et al., *Cryo-EM reveals an active role for aminoacyl-tRNA in the accommodation process.* The EMBO journal, 2002. **21**(13): p. 3557-3567.

108. Gao, H., et al., *Dynamics of EF-G interaction with the ribosome explored by classification of a heterogeneous cryo-EM dataset.* Journal of structural biology, 2004. **147**(3): p. 283-290.

109. Sorzano, C.O.S., et al., *A clustering approach to multireference alignment of single-particle projections in electron microscopy.* Journal of structural biology, 2010. **171**(2): p. 197-206.

110. Kimanius, D., et al., *New tools for automated cryo-EM single-particle analysis in RELION-4.0.* Biochemical Journal, 2021. **478**(24): p. 4169-4185.

111. Hashem, Y., et al., *Structure of the mammalian ribosomal 43S preinitiation complex bound to the scanning factor DHX29.* Cell, 2013. **153**(5): p. 1108-1119.

112. Nakane, T., et al., *Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION.* elife, 2018. **7**: p. e36861.

113. Navarro, P.P., H. Stahlberg, and D. Castaño-Díez, *Protocols for subtomogram averaging of membrane proteins in the Dynamo software package.* Frontiers in molecular biosciences, 2018. **5**: p. 82.

114. Förster, F., et al., *Classification of cryo-electron sub-tomograms using constrained correlation.* Journal of structural biology, 2008. **161**(3): p. 276-286.

115. Ni, T., et al., *High-resolution in situ structure determination by cryo-electron tomography and subtomogram averaging using emClarity.* Nature protocols, 2022. **17**(2): p. 421-444.

116. Liao, H.Y., Y. Hashem, and J. Frank, *Efficient estimation of three-dimensional covariance and its application in the analysis of heterogeneous samples in cryo-electron microscopy.* Structure, 2015. **23**(6): p. 1129-1137.

117. Andén, J., E. Katsevich, and A. Singer. *Covariance estimation using conjugate gradient for 3D classification in cryo-EM*. in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. 2015. IEEE.

118. Marshall, N.F., et al., *Fast Principal Component Analysis for Cryo-EM Images.* arXiv preprint arXiv:2210.17501, 2022.

119. Sorzano, C.O.S. and J.M. Carazo, *Principal component analysis is limited to low-resolution analysis in cryoEM.* Acta Crystallographica Section D: Structural Biology, 2021. **77**(6): p. 835-839.

120. Coifman, R.R., et al., *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps.* Proceedings of the national academy of sciences, 2005. **102**(21): p. 7426-7431.

121. Sztain, T., et al., *A glycan gate controls opening of the SARS-CoV-2 spike protein.* Nature chemistry, 2021. **13**(10): p. 963-968.

122. Mashayekhi, G., et al., *Energy landscape of the SARS-CoV-2 reveals extensive conformational heterogeneity.* Current Research in Structural Biology, 2022. **4**: p. 68-77.

123. Zhong, E.D., et al., *Exploring generative atomic models in cryo-EM reconstruction.* arXiv preprint arXiv:2107.01331, 2021.

124. Rosenbaum, D., et al., *Inferring a continuous distribution of atom coordinates from cryo-EM images using VAEs.* arXiv preprint arXiv:2106.14108, 2021.

125. Donnat, C., et al., *Deep generative modeling for volume reconstruction in cryo-electron microscopy.* Journal of Structural Biology, 2022: p. 107920.

126. Powell, B.M. and J.H. Davis, *Learning structural heterogeneity from cryo-electron sub-tomograms with tomoDRGN.* bioRxiv, 2023: p. 2023.05. 31.542975.

127. Zhong, E.D., et al. *CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images*. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

128. Zlatanova, J., et al., *The nucleosome family: dynamic and growing.* Structure, 2009. **17**(2): p. 160-171.

129. Eswar, N., et al., *Protein structure modeling with MODELLER.* Structural proteomics: high-throughput methods, 2008: p. 145-159.

130. Schwede, T., et al., *SWISS-MODEL: an automated protein homology-modeling server.* Nucleic acids research, 2003. **31**(13): p. 3381-3385.

131. Jamali, K., D. Kimanius, and S. Scheres, *ModelAngelo: Automated Model Building in Cryo-EM Maps.* arXiv preprint arXiv:2210.00006, 2022.

132. ben-Avraham, D. and M.M. Tirion, *Normal modes analyses of macromolecules.* Physica A: Statistical Mechanics and its Applications, 1998. **249**(1-4): p. 415-423.

133. Huang, J. and A.D. MacKerell Jr, *CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data.* Journal of computational chemistry, 2013. **34**(25): p. 2135-2145.

134. Wang, J., et al., *Development and testing of a general amber force field.* Journal of computational chemistry, 2004. **25**(9): p. 1157-1174.

135. Levitt, M. and A. Warshel, *Computer simulation of protein folding.* Nature, 1975. **253**(5494): p. 694-698.

136. Kolinski, A., et al., *An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side group centers of mass.* The Journal of Physical Chemistry B, 1998. **102**(23): p. 4628-4637.

137. Liwo, A., et al., *A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions.* Journal of molecular modeling, 2014. **20**: p. 1-15.

138. Brand, L. and M.L. Johnson, *Numerical Computer Methods, Part D*. 2004: Elsevier.

139. Koliński, A., *Protein modeling and structure prediction with a reduced representation.* Acta Biochimica Polonica, 2004. **51**.

140. Kmiecik, S., et al., *Coarse-grained protein models and their applications.* Chemical reviews, 2016. **116**(14): p. 7898-7936.

141. Marrink, S.J., et al., *The MARTINI force field: coarse grained model for biomolecular simulations.* The journal of physical chemistry B, 2007. **111**(27): p. 7812-7824.

142. Clark, L.A. and H.W. van Vlijmen, *A knowledge-based forcefield for protein–protein interface design.* Proteins: Structure, Function, and Bioinformatics, 2008. **70**(4): p. 1540-1550.

143. Go, N., *Theoretical studies of protein folding.* Annual review of biophysics and bioengineering, 1983. **12**(1): p. 183-210.

144. Cieplak, M., T.X. Hoang, and M.O. Robbins, *Folding and stretching in a Go-like model of titin.* Proteins: Structure, Function, and Bioinformatics, 2002. **49**(1): p. 114-124.

145. Sułkowska, J.I. and M. Cieplak, *Mechanical stretching of proteins—a theoretical survey of the Protein Data Bank.* Journal of Physics: Condensed Matter, 2007. **19**(28): p. 283201.

146. Hills, R.D., Jr. and C.L. Brooks, 3rd, *Insights from coarse-grained Gō models for protein folding and dynamics.* Int J Mol Sci, 2009. **10**(3): p. 889-905.

147. Clementi, C., H. Nymeyer, and J.N. Onuchic, *Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route"*

intermediates for protein folding? An investigation for small globular proteins. Journal of molecular biology, 2000. **298**(5): p. 937-953.

148. Tirion, M.M., *Large amplitude elastic motions in proteins from a single-parameter, atomic analysis.* Physical review letters, 1996. **77**(9): p. 1905.

149. Tama, F. and Y.-H. Sanejouand, *Conformational change of proteins arising from normal mode calculations.* Protein engineering, 2001. **14**(1): p. 1-6.

150. Tama, F., W. Wriggers, and C.L. Brooks III, *Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory.* Journal of molecular biology, 2002. **321**(2): p. 297-305.

151. Miyashita, O. and F. Tama, *Hybrid methods for macromolecular modeling by molecular mechanics simulations with experimental data.* Integrative Structural Biology with Hybrid Methods, 2018: p. 199-217.

152. Kmiecik, S., et al., *Modeling of protein structural flexibility and large-scale dynamics: Coarse-grained simulations and elastic network models.* International journal of molecular sciences, 2018. **19**(11): p. 3496.

153. Metropolis, N., et al., *Equation of state calculations by fast computing machines.* The journal of chemical physics, 1953. **21**(6): p. 1087-1092.

154. Duane, S., et al., *Hybrid monte carlo.* Physics letters B, 1987. **195**(2): p. 216-222.

155. Neal, R.M., *MCMC using Hamiltonian dynamics*, in *Handbook of Markov Chain Monte Carlo*, A.G. Steve Brooks, Galin Jones, and Xiao-Li Meng, Editor. 2011, Chapman & Hall / CRC Press. p. 113–162.

156. Hoffman, M.D. and A. Gelman, *The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.* J. Mach. Learn. Res., 2014. **15**(1): p. 1593-1623.

157. Durand, P., G. Trinquier, and Y.H. Sanejouand, *A new approach for determining low-frequency normal modes in macromolecules.* Biopolymers: Original Research on Biomolecules, 1994. **34**(6): p. 759-771.

158. Bauer, J.A., J. Pavlović, and V. Bauerová-Hlinková, *Normal mode analysis as a routine part of a structural investigation.* Molecules, 2019. **24**(18): p. 3293.

159. Hoffmann, A. and S. Grudinin, *NOLB: Nonlinear rigid block normal-mode analysis method.* Journal of chemical theory and computation, 2017. **13**(5): p. 2123-2134.

160. Pearson, K., *LIII. On lines and planes of closest fit to systems of points in space.* The London, Edinburgh, and Dublin philosophical magazine and journal of science, 1901. **2**(11): p. 559-572.

161. Wold, S., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometrics and intelligent laboratory systems, 1987. **2**(1-3): p. 37-52.

162. Papaleo, E., et al., *Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: the myoglobin case.* Journal of molecular graphics and modelling, 2009. **27**(8): p. 889-899.

163. Ferguson, A.L., et al., *Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach.* Chemical Physics Letters, 2011. **509**(1-3): p. 1-11.

164. Trozzi, F., X. Wang, and P. Tao, *UMAP as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: a comparison study.* The Journal of Physical Chemistry B, 2021. **125**(19): p. 5022-5034.

165. Tenenbaum, J.B., V.d. Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* science, 2000. **290**(5500): p. 2319-2323.

166. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(11).

167. McInnes, L., J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* ArXiv, 2018. **abs/1802.03426**.

168. Trabuco, L.G., et al., *Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography.* Methods, 2009. **49**(2): p. 174-180.

169. Wu, X., et al., *Targeted conformational search with map-restrained self-guided Langevin dynamics: application to flexible fitting into electron microscopic density maps.* Journal of structural biology, 2013. **183**(3): p. 429-440.

170. Igaev, M., et al., *Automated cryo-EM structure refinement using correlation-driven molecular dynamics.* Elife, 2019. **8**: p. e43542.

171. Bonomi, M., R. Pellarin, and M. Vendruscolo, *Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy.* Biophys J, 2018. **114**(7): p. 1604-1613.

172. Habeck, M., *Bayesian Modeling of Biomolecular Assemblies with Cryo-EM Maps.* Frontiers in Molecular Biosciences, 2017. **4**.

173. Rieping, W., M. Habeck, and M. Nilges, *Inferential structure determination.* Science, 2005. **309**(5732): p. 303-306.

174. Habeck, M., M. Nilges, and W. Rieping, *Bayesian inference applied to macromolecular structure determination.* Physical Review E, 2005. **72**(3): p. 031912.

175. Tama, F., O. Miyashita, and C.L. Brooks III, *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**(4): p. 985-999.

176. Navaza, J., et al., *On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation.* Acta Crystallographica Section D: Biological Crystallography, 2002. **58**(10): p. 1820-1825.

177. Suhre, K., J. Navaza, and Y.-H. Sanejouand, *NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps.* Acta crystallographica section D: biological crystallography, 2006. **62**(9): p. 1098-1100.

178. Tama, F., O. Miyashita, and C.L. Brooks Iii, *Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM.* Journal of structural biology, 2004. **147**(3): p. 315-326.

179. Schröder, G.F., A.T. Brunger, and M. Levitt, *Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution.* Structure, 2007. **15**(12): p. 1630-1641.

180. Wang, Z. and G.F. Schröder, *Real-space refinement with DireX: From global fitting to side-chain improvements.* Biopolymers, 2012. **97**(9): p. 687-697.

181. De Groot, B., et al., *Prediction of protein conformational freedom from distance constraints.* Proteins: Structure, Function, and Bioinformatics, 1997. **29**(2): p. 240-251.

182. Costa, M.G., et al., *A new strategy for atomic flexible fitting in cryo-EM maps by molecular dynamics with excited normal modes (MDeNM-EMfit).* Journal of chemical information and modeling, 2020. **60**(5): p. 2419-2423.

183. Harastani, M., et al., *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**: p. 663121.

184. Cossio, P. and G. Hummer, *Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies.* Journal of structural biology, 2013. **184**(3): p. 427-437.

185. Vuillemot, R. and S. Jonić. *Combined Bayesian and Normal Mode Flexible Fitting with Hamiltonian Monte Carlo Sampling for Cryo Electron Microscopy*. in *2021 29th European Signal Processing Conference (EUSIPCO)*. 2021. IEEE.

186. Kobayashi, C., et al., *GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms*. 2017, Wiley Online Library.

187. Oh, B.-H., et al., *Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand.* Journal of Biological Chemistry, 1993. **268**(15): p. 11348-11355.

188. Müller, C., et al., *Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding.* Structure, 1996. **4**(2): p. 147-156.

189. Müller, C.W. and G.E. Schulz, *Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 Å resolution: A model for a catalytic transition state.* Journal of molecular biology, 1992. **224**(1): p. 159-177.

190. Haridas, M., B. Anderson, and E. Baker, *Structure of human diferric lactoferrin refined at 2.2 Å resolution.* Acta Crystallographica Section D: Biological Crystallography, 1995. **51**(5): p. 629-646.

191. Norris, G., B. Anderson, and E. Baker, *Molecular replacement solution of the structure of apolactoferrin, a protein displaying large-scale conformational change.* Acta Crystallographica Section B: Structural Science, 1991. **47**(6): p. 998-1004.

192. Jørgensen, R., et al., *Two crystal structures demonstrate large conformational changes in the eukaryotic ribosomal translocase.* Nature Structural & Molecular Biology, 2003. **10**(5): p. 379-385.

193. Tama, F., et al., *Building-block approach for determining low-frequency normal modes of macromolecules.* Proteins: Structure, Function, and Bioinformatics, 2000. **41**(1): p. 1-7.

194. Suhre, K. and Y.H. Sanejouand, *ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W610-4.

195. Goddard, T.D., et al., *UCSF ChimeraX: Meeting modern challenges in visualization and analysis.* Protein Science, 2018. **27**(1): p. 14-25.

196. Davis, I.W., et al., *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids.* Nucleic acids research, 2007. **35**(suppl_2): p. W375-W383.

197. Vilas, J.L., et al., *MonoRes: automatic and accurate estimation of local resolution for electron microscopy maps.* Structure, 2018. **26**(2): p. 337-344. e4.

198. Harastani, M., et al., *ContinuousFlex: Software package for analyzing continuous conformational variability of macromolecules in cryo electron microscopy and tomography data.* Journal of Structural Biology, 2022: p. 107906.

199. Cock, P.J.A., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.

200. Takada, S., *Gō model revisited.* Biophysics and physicobiology, 2019. **16**: p. 248-255.

201. Noel, J.K., et al. *SMOG 2: A Versatile Software Package for Generating Structure-Based Models*. PLoS computational biology, 2016. **12**, e1004794 DOI: 10.1371/journal.pcbi.1004794.

202. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**(null): p. 2825–2830.

203. Tipping, M.E. and C.M. Bishop, *Mixtures of Probabilistic Principal Component Analyzers.* Neural Computation, 1999. **11**(2): p. 443-482.

204. Sorzano, C.O.S., et al., *XMIPP: a new generation of an open-source image processing package for electron microscopy.* Journal of Structural Biology, 2004. **148**(2): p. 194-204.

205. Peng, L.-M., et al., *Robust parameterization of elastic and absorptive electron atomic scattering factors.* Acta Crystallographica Section A: Foundations of Crystallography, 1996. **52**(2): p. 257-276.

206. Frank, J. and R.K. Agrawal, *A ratchet-like inter-subunit reorganization of the ribosome during translocation.* Nature, 2000. **406**(6793): p. 318-322.

207. Spiegel, P.C., D.N. Ermolenko, and H.F. Noller, *Elongation factor G stabilizes the hybrid-state conformation of the 70S ribosome.* Rna, 2007. **13**(9): p. 1473-1482.

208. Ling, C. and D.N. Ermolenko, *Structural insights into ribosome translocation.* WIREs RNA, 2016. **7**(5): p. 620-636.

209. Grigorieff, N., *Frealign: an exploratory tool for single-particle cryo-EM.* Methods in enzymology, 2016. **579**: p. 191-226.

210. Murakami, R., et al., *The Interaction between the Ribosomal Stalk Proteins and Translation Initiation Factor 5B Promotes Translation Initiation.* Mol Cell Biol, 2018. **38**(16).

211. Ratje, A.H., et al., *Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites.* Nature, 2010. **468**(7324): p. 713-716.

212. Natchiar, S.K., et al., *Visualization of chemical modifications in the human 80S ribosome structure.* Nature, 2017. **551**(7681): p. 472-477.

213. Deshaies, R.J., *Proteotoxic crisis, the ubiquitin-proteasome system, and cancer therapy.* BMC biology, 2014. **12**(1): p. 1-14.

214. Kobakhidze, G., et al., *The AAA+ ATPase p97 as a novel parasite and tuberculosis drug target.* Trends in Parasitology, 2022.

215. Valimehr, S., et al., *Molecular Mechanisms Driving and Regulating the AAA+ ATPase VCP/p97, an Important Therapeutic Target for Treating Cancer, Neurological and Infectious Diseases.* Biomolecules, 2023. **13**(5): p. 737.

216. Rouiller, I., et al., *A major conformational change in p97 AAA ATPase upon ATP binding.* Molecular cell, 2000. **6**(6): p. 1485-1490.

217. Rouiller, I., et al., *Conformational changes of the multifunction p97 AAA ATPase during its ATPase cycle.* Nature structural biology, 2002. **9**(12): p. 950-957.

218. Mountassif, D., et al., *Cryo-EM of the pathogenic VCP variant R155P reveals long-range conformational changes in the D2 ATPase ring.* Biochemical and biophysical research communications, 2015. **468**(4): p. 636-641.

219. Cooney, I., et al., *Structure of the Cdc48 segregase in the act of unfolding an authentic substrate.* Science, 2019. **365**(6452): p. 502-505.

220. Pan, M., et al., *Mechanistic insight into substrate processing and allosteric inhibition of human p97.* Nature structural & molecular biology, 2021. **28**(7): p. 614-625.

221. Twomey, E.C., et al., *Substrate processing by the Cdc48 ATPase complex is initiated by ubiquitin unfolding.* Science, 2019. **365**(6452): p. eaax1033.

222. Blythe, E.E., et al., *Multisystem proteinopathy mutations in VCP/p97 increase NPLOC4· UFD1L binding and substrate processing.* Structure, 2019. **27**(12): p. 1820-1829. e4.

223. Schütz, A.K., E. Rennella, and L.E. Kay, *Exploiting conformational plasticity in the AAA+ protein VCP/p97 to modify function.* Proceedings of the National Academy of Sciences, 2017. **114**(33): p. E6822-E6829.

224. Valimehr, S., *Structure and dynamic studies of AAA+ ATPase p97 molecular machine*. 2021, The University of Melbourne.

225. Turoňová, B., et al., *In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges.* Science, 2020. **370**(6513): p. 203-208.

226. Chen, Y., et al., *Fast and accurate reference-free alignment of subtomograms.* Journal of Structural Biology, 2013. **182**(3): p. 235-245.

227. Hamming, I., et al., *Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis.* The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 2004. **203**(2): p. 631-637.

228. Hoffmann, M., et al., *SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor.* cell, 2020. **181**(2): p. 271-280. e8.

229. Benton, D.J., et al., *Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion.* Nature, 2020. **588**(7837): p. 327-330.

230. Narayan, A., B. Berger, and H. Cho, *Density-Preserving Data Visualization Unveils Dynamic Patterns of Single-Cell Transcriptomic Variability.* bioRxiv, 2020. **10**(2020.05): p. 12.077776.

231. Heumann, J.M., A. Hoenger, and D.N. Mastronarde, *Clustering and variance maps for cryo-electron tomography using wedge-masked differences.* Journal of structural biology, 2011. **175**(3): p. 288-299.

232. Jonić, S. and C.Ó.S. Sorzano, *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**(1): p. 161-173.

233. Badaczewska-Dawid, A.E., A. Kolinski, and S. Kmiecik, *Computational reconstruction of atomistic protein structures from coarse-grained models.* Computational and structural biotechnology journal, 2020. **18**: p. 162-176.

234. Harauz, G. and M. van Heel, *Exact filters for general geometry three dimensional reconstruction.* Optik., 1986. **73**(4): p. 146-156.

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

236. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.

237. Carlson, H.A., *Protein flexibility is an important component of structure-based drug discovery.* Current Pharmaceutical Design, 2002. **8**(17): p. 1571-1578.

238. Lees, J.A., J.M. Dias, and S. Han, *Applications of Cryo-EM in small molecule and biologics drug design.* Biochemical Society Transactions, 2021. **49**(6): p. 2627-2638.

239. Zhu, K.-F., et al., *Applications and prospects of cryo-EM in drug discovery.* Military Medical Research, 2023. **10**(1): p. 10.

240. Luque, I. and E. Freire, *Structural stability of binding sites: consequences for binding affinity and allosteric effects.* Proteins: Structure, Function, and Bioinformatics, 2000. **41**(S4): p. 63-71.

241. Sánchez, R.M., R. Mester, and M. Kudryashev. *Fast alignment of limited angle tomograms by projected cross correlation*. in *2019 27th European Signal Processing Conference (EUSIPCO)*. 2019. IEEE.