COMPUTATIONAL METHODS TO IMPROVE CRYO-EM STRUCTURAL ANALYSIS

By
Satinder Kaur

Department of Anatomy and Cell Biology
McGill University, Montreal

August 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Contents

List of Figures and Tables

Abbreviations

2D	Two Dimensional
3D	Three Dimensional
3D-EM	Three Dimensional - Electron Microscopy
Å	Angstrom
ArnA	4-amino-4-deoxy-L-arabinose
B-factor	Debye–Waller factor
BIM	Beam Induced Motion
CCD	Charge-Coupled device
CMOS	Complementary Metal Oxide Semiconductor
CNB	National Center for Biotechnology
CNN	Convolutional Neural Network
CPU	Central Processing Unit
Cryo-EM	Cryo-Electron Microscopy
CT	Contrast Transfer
CTF	Contrast Transfer Function
DDD	Direct Detection Device
DPR	Differential Phase Residual
DQE	Detective Quantum Efficiency
ECT	Electron Cryo-Tomography
EM	Electron Microscopy
EMD	EMDB map accession id (EMD-XXXX)

EMDB	Electron Microscopy Data Bank
EMPIAR	Electron Microscopy Public Image Archive
eV	Electronvolt
FEI	Field Electron and Ion Company
FSC	Fourier Shell Correlation
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IR-ECD	Insulin Receptor-Ectodomain
Isomap	Isometric mapping
K	Kelvin
KeV	Kilo-Electronvolt
LaB6	Lanthanum hexaboride
LLE	Locally Linear Embedding
LTSA	Local Tangent Space Alignment
MDS	Multidimensional Scaling
ML2D	Maximum Likelihood Two-Dimensional (classification approach)
ML3D	Maximum Likelihood Three-Dimensional (classification approach)
MRA	Multi-Reference Alignment
MRC lab	Medical Research Council Laboratory
MSA	Multiple Sequence Alignment
NAS	Network-Attached Storage
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance

PACRG	PArkin Co-Regulated Gene
PCA	Principal Component Analysis
PDB	Protein Data Bank
PSF	Point Spread Function
RAID	Redundant Arrays of Disks
RANSAC	Random Sample Consensus
RCT	Random Canonical Tilt
Relion	REgularized LIkelihood OptimizatioN
S (eg:30S)	Svedbergs
SANs	Storage Area Networks
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SNR	Signal-to-Noise Ratio
SPA	Single-Particle Analysis
SPT	Spiral Phase Transformation
SSNR	Spectral Signal-to-Noise Ratio
STA	Subtomogram Averaging
TB	TeraByte
TEM	Transmission Electron Microscopy
t-RNA	Transfer Ribonucleic Acid
t-SNE	T-distributed Stochastic Neighbor Embedding
VPP	Volta Phase Plate

Abstract (English)

Cryo-EM has yielded important discoveries about fine cellular structures, viruses, and protein complexes at high resolution over recent years. This popular technique is progressing stronger than ever, owing to various experimental and computational methods advancements. Handling multiple conformations and providing finer structural interpretability of a dynamic macromolecular has strengthened the potential of this technique. However, analysis of heterogeneous data remains a major challenge for cryo-EM, even after using tailored image processing techniques. Here, heterogeneity is concerned with varying resolution distribution on a cryo-EM map or several existing conformations of a macromolecule resulting from its flexibility. These dynamic developments can arise due to various reasons such as when interacting with biomolecules and ligands or in spontaneous fluctuation due to biological functions. Failing to analyse these movements can interrupt the interpretation of a given macromolecule structure and function. Thus, sophisticated methods are required to record these biologically significant macromolecular movements. The focus of this thesis is to demonstrate improvements in interpreting and analysing the cryo-EM maps with computational methods to account for map changes locally for building efficient atomic models and to visualize the conformational dynamics for a flexible or dynamic macromolecule.

Biologists may use atomic models, even without expert knowledge, to explore the function of a macromolecule depending on mechanistic hypotheses concluded from its structure. Building an accurate atomic model, therefore, is an important part of the structural analysis procedure and is based on the quality of the EM density map. However, various cryo-EM density maps from recent periods contain local regions whose resolution differs from the global map resolution reported. This heterogeneous distribution of resolution can arise due to flexibility as well as subunit

occupancy in diverse parts of the structure. Analysing such a map can cause poorer quality reconstruction. Recent image-processing algorithms have suffered limitations in describing these local resolution differences among distinct parts of the structure.

Domain flexibility of many large macromolecules also contributes to generating various conformations. In cryo-EM, these dynamic structures require sorting into their least variable forms by performing 3D classifications, which usually capture these conformations as snapshots in the form of 3D classes. The resultant 3D classes can prefer lower-energy states within the thermodynamic equilibrium, such that analysing the whole cryo-EM data can show the trajectory of movements corresponding to various conformations of a macromolecule. Yet, generating a suitable method for recording conformation mapping from representative 3D classes remains an ongoing process. The work done in this thesis discusses the details of single-particle image analysis methods designed for better utilizing the heterogeneous data.

Abstract (French)

Cryo-EM a permis d'importantes découvertes sur les structures cellulaires fines, les virus et les complexes protéiques à haute résolution au cours des dernières années. Cette technique populaire progresse plus fort que jamais grâce aux diverses avancées des méthodes expérimentales et informatiques. La manipulation de multiples conformations et la fourniture d'une interprétabilité structurelle plus fine d'une macromoléculaire dynamique ont renforcé le potentiel de cette technique. Cependant, l'analyse de données hétérogènes reste toujours un défi majeur pour la cryo-EM, même après avoir utilisé des techniques de traitement d'image sur mesure. Ici, l'hétérogénéité concerne la distribution de résolution variable sur une carte cryo-EM ou plusieurs conformations existantes d'une macromolécule résultant de sa flexibilité. Ces développements dynamiques peuvent survenir pour diverses raisons telles que lors de l'interaction avec des biomolécules et des ligands ou lors de fluctuations spontanées dues à des fonctions biologiques. Ne pas analyser ces mouvements peut interrompre l'interprétation de la structure et de la fonction d'une macromolécule donnée. Ainsi, des méthodes sophistiquées sont nécessaires pour enregistrer ces mouvements macromoléculaires biologiquement significatifs. L'objectif de cette thèse est de démontrer des améliorations dans l'interprétation et l'analyse des cartes cryo-EM avec des méthodes de calcul pour tenir compte des changements de carte localement pour construire des modèles atomiques efficaces et pour visualiser la dynamique conformationnelle d'une macromolécule flexible ou dynamique.

Les biologistes peuvent utiliser des modèles atomiques, même sans connaissances spécialisées, pour explorer la fonction d'une macromolécule en fonction d'hypothèses mécanistes tirées de sa structure. La construction d'un modèle atomique précis est donc une partie importante de la procédure d'analyse structurelle et est basée sur la qualité de la carte de densité EM. Cependant,

diverses cartes de densité cryo-EM de périodes récentes contiennent des régions locales dont la résolution diffère de la résolution de la carte globale rapportée. Cette distribution hétérogène de la résolution peut survenir en raison de la flexibilité ainsi que de l'occupation des sous-unités dans diverses parties de la structure. L'analyse d'une telle carte peut entraîner une reconstruction de moins bonne qualité. Les algorithmes récents de traitement d'images ont souffert de limitations dans la description de ces différences de résolution locales entre les différentes parties de la structure.

La flexibilité du domaine de nombreuses grandes macromolécules contribue également à générer diverses conformations. En cryo-EM, ces structures dynamiques nécessitent un tri dans leurs formes les moins variables en effectuant des classifications 3D, qui capturent généralement ces conformations sous forme d'instantanés sous la forme de classes 3D. Les classes 3D résultantes peuvent préférer des états à faible énergie dans l'équilibre thermodynamique, par conséquent, l'analyse de l'ensemble des données cryo-EM peut montrer la trajectoire des mouvements correspondant à diverses conformations d'une macromolécule. Cependant, la génération d'une méthode appropriée pour enregistrer la cartographie de conformation à partir de classes 3D représentatives est toujours un processus en cours. Le travail effectué dans cette thèse traite des détails des méthodes d'analyse d'images à une seule particule conçues pour mieux utiliser les données hétérogènes.

Acknowledgments

I sincerely thank **Dr. Javier Vargas** (primary supervisor 2019 to 2020, present co-supervisor) for giving me this wonderful opportunity to join such an esteemed university. I am and will be always grateful for his guidance and enormous patience with me. I want to thank **Dr. Mike Strauss** (current supervisor) for his unconditional support and virtuous counseling. I really appreciate his amazing passion for research and for providing constant motivation to his students, which has considerably influenced me in strenuous times.

I want to thank all the past and present members from Dr. Vargas' lab and Dr. Strauss' lab for their assistance and friendship. Also, I want to mention **Dr. Joaquin Ortega's** help in offering his suggestions to progress my research work. Special thanks to **Dr. Chantal Autexier and Shalom Chaim Spira** for her help with the organization and careful review of the thesis format.

I am thankful for my mentor **Dr. Susanne Bechstedt** and two Doctoral Advisory Committee members **Dr. Khanh Huy Bui** and **Dr. Jean-Francois Trempe** along with former member **Dr. Alba Guarné** for their advice throughout my PhD degree. Finally, I want to thank McGill's **Department of Anatomy and Cell Biology** for its prompt support and guidance in my PhD.

This thesis is dedicated to my personal role model **Dr. Veena Puri** and my parents **Jatinder Singh and Manjit Kaur**. Thank you for giving me this freedom of education.

Contributions to original knowledge

The work from this thesis was aimed to contribute to the improving of cryo-EM image analysis methods with a special focus on highly heterogeneous samples. Chapter 2 presents a novel method to improve cryo-EM maps locally, which are affected by a non-homogeneous distribution of SNRs. This method calculates local B-factors and local map occupancies for cryo-EM maps, dampening the Fourier amplitudes of less-well-ordered domains while boosting the high-resolution regions. The approaches described in this publication help to improve the map connectivity as well as produce a better coverage of atomic models.

Studies in chapter 3 discuss the proposed Bayesian inference-based method to discover various conformations in cryo-EM data. It has a two-fold approach. The first part helps to deduce a large number of conformations in extensive heterogenous datasets by performing hierarchical automatic 2D and 3D classifications that eliminate the need for human expertise as well as the so-called "attractor problem," a phenomenon whereby conformations with low population numbers cannot be captured by typical Bayesian approaches. The second part of the method presents these resultant conformations in the form of a trajectory on a free-energy landscape.

This manuscript-based thesis consists of two original research articles showcased in chapter 2 and 3, as follows:

1. *Kaur, S., Gomez-Blanco, J., Khalifa, A.A., Adinarayanan, S., Sanchez-Garcia, R., Wrapp, D., McLellan, J.S., Bui, K.H. and Vargas, J., 2021. Local computational methods to improve the interpretability and analysis of cryo-EM maps. Nature communications, 12(1), pp.1-12*[1]

2. *Gomez-Blanco, J., Kaur, S., Strauss, M. and Vargas, J., 2022. Hierarchical autoclassification of cryo-EM samples and macromolecular energy landscape determination. Computer Methods and Programs in Biomedicine, p.106673*[2]

First-authored manuscripts and author contributions

Chapter 2: My contribution consisted of developing the method, processing preliminary sample data, devising the theory, figures preparations, and writing the first draft of the method, which provided the baseline for the present publication method section, and providing feedback and comments on the manuscript's drafts.

Contributions of other authors: Dr. Javier Vargas conceived the idea. Dr. Javier Vargas and Dr. Josue Gomez-Blanco devised the theory, participated in developing and implementing the algorithm, performed experiments, and wrote the manuscript. Data analysis and interpretation were also orchestrated by Dr. Ruben Sanchez-Garcia and Swathi Adinarayanan. Ahmad A. Z. Khalifa, Dr. Daniel Wrapp, Dr. Jason S. McLellan and Dr. Khanh Huy Bui analysed data, wrote part of the manuscript and provided comments and feedback. All authors reviewed the manuscript, supervised the experiments and discussed the results.

Chapter 3: My contribution, as a first co-author (along with visiting scholar Dr. Josue Gomez-Blanco), involved designing the algorithm for the free-energy landscape generation method in the manuscript, implementing the algorithm, processing data, providing results and interpretations, and reviewing the manuscript.

Contributions of other authors: Dr. Javier Vargas formulated the idea, Dr. Josue Gomez-Blanco and Dr. Javier Vargas devised the theory, performed the experiments, and wrote the manuscript. Dr. Josue Gomez-Blanco and Dr. Javier Vargas developed and implemented the algorithm. All authors reviewed the manuscript, supervised experiments and discussed results.

Advances toward PhD objectives, (e.g., awards and presentations)

- **Achievements:**
    - 2021 Microscopy & Microanalysis (M&M) Student Scholar Award
    - Exclusive interview with Microscopy Society of America (MSA) for July 2021 spotlight issue
    - Differential fee waiver by Training & Awards Committee of the Centre de Recherche en Biologie Structurale (CRBS), Sept 2019 – May 2020
    - Graduate Research Enhancement and Travel Award by Graduate Studies in Cell Biology & Anatomy McGill University, Montreal, QC
- **Presentations:**
    - Microscopy & Microanalysis (M&M) 2021 platform presentation
    - Invited speaker to Microscopy society of America's 5th pre-meeting Congress (PMCx60) on 'Advanced Data Analysis.'
    - 2021 Canadian Microscopical Society Symposium
    - Centre de Recherche en Biologie Structurale (CRBS), 2021
    - 1st Symposium of Structural Biology, 2019, Université de Montréal (UdeM)
    - 1st Annual CSB Symposium, 2019, McGill Centre for Structural Biology
    - Annual Retreat 2019, Department of Anatomy and Cell Biology, McGill University
    - XXVIII. Symposium of GÉPROM –2019, Université de Montréal
    - ACB Undergraduate Research & Lab Recruitment Day, 2019, McGill University

CHAPTER 1

INTRODUCTION

The structural analysis of macromolecules to achieve functional information inside the cell is the core of structural biology. Using morphological determining techniques, scientists can analyse the various aspects of a macromolecule such as its biological behaviour, drug development, action mechanisms, or design of new structural complexes, among others. Various structural analysis techniques include NMR, X-ray crystallography, and cryo-EM, among others. However, no single technique suffices to elucidate all aspects of any given macromolecule. Rather, these techniques should complement each other to describe the macromolecule's complete structural and functional analysis. Among these techniques, cryo-EM has gained popularity over the last few years, as shown in Figure 1(a), as the resolution of the structures solved by this technique has been highly increasing. Cryo-EM uses electrons to pass through the flash-frozen solutions of macromolecules for creating its 3D structure. This technique employs advanced image processing methods to reconstruct biological macromolecules, and represents the topic of this thesis.

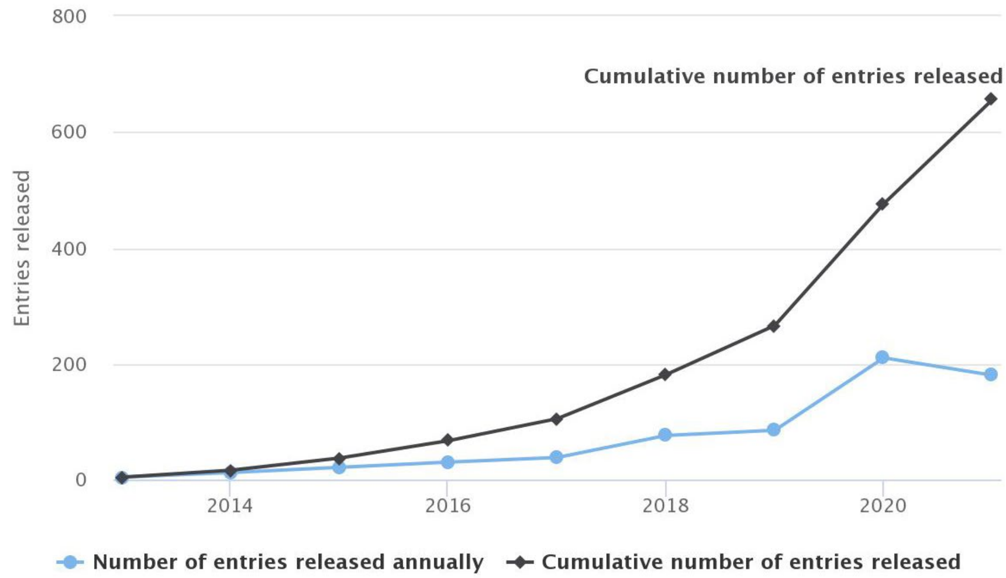
1.1 History of EM

Transmission electron microscopy (TEM) typically uses electrons accelerated to 200-300 keV to magnify biological samples to elucidate their structure and composition. Ernst Ruska initiated the discipline of electron microscopy by developing a mini microscope in 1925[3–5]. In 1926, Hans Busch invented the first electromagnetic lens which used electric and magnetic fields to shape the paths followed by electrons, similar to how glass lenses are used to bend and focus visible light[3,5]. Ernst Ruska and Max Knoll, from the University of Berlin, created the first transmission electron microscope in 1931, for which Ruska was awarded the Nobel Prize for Physics in 1986[6] (Figure 1(b)). It is considered one of the most influential scientific inventions for allowing us to visualize the nanoworld. In the first image of a eukaryotic cell[7], TEM was able to show many of the organelles for the first time[8,9]. Other breakthroughs included the imaging of neurons communicating through neurotransmitters[10,11], sliding Earnys-filament theory on muscle contraction[10,12,13], and revealing viruses to be defined particles for the first time[14,15].

1.2 The rise of cryo-EM

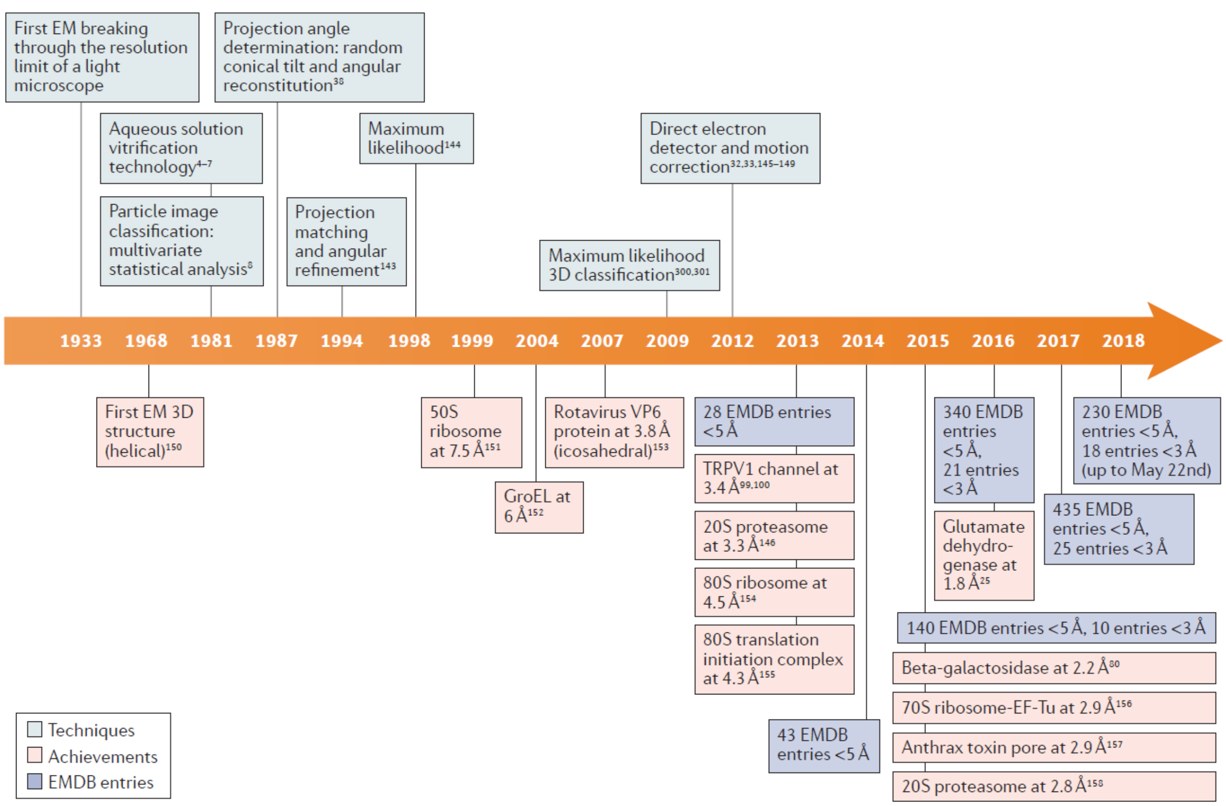
The ability to detect features in an image is determined by image contrast; in particular, imaging in a TEM is typically achieved by phase contrast, which increases with the difference in the atomic number of the atoms that make up the sample. But in biology, the phase contrast is usually low due to the similarly low atomic number of elements comprising the biological sample and the aqueous solution that carries the samples in their native state: carbon (Z=6) and oxygen (Z=8). The contrast of biological elements could be increased by extending the exposure time, yet the high energy electrons used for imaging will lead to the burning of the sample[16]. R.M. Glaeser proposed the solution to this issue by averaging multiple images[16].

(a)



800
600
400
200
0
Entries released
Cumulative number of entries released
2014
2016
2018
2020
Number of entries released annually
Cumulative number of entries released

(b)



First EM breaking through the resolution limit of a light microscope
Projection angle determination: random conical tilt and angular reconstitution[38]
Aqueous solution vitrification technology[4–7]
Maximum likelihood[144]
Direct electron detector and motion correction[32,33,145–149]
Particle image classification: multivariate statistical analysis[8]
Projection matching and angular refinement[143]
Maximum likelihood 3D classification[300,301]
1933
1968
1981
1987
1994
1998
1999
2004
2007
2009
2012
2013
2014
2015
2016
2017
2018
First EM 3D structure (helical)[150]
50S ribosome at 7.5 Å[151]
GroEL at 6 Å[152]
Rotavirus VP6 protein at 3.8 Å (icosahedral)[153]
28 EMDB entries <5 Å
TRPV1 channel at 3.4 Å[99,100]
20S proteasome at 3.3 Å[146]
80S ribosome at 4.5 Å[154]
80S translation initiation complex at 4.3 Å[155]
43 EMDB entries <5 Å
340 EMDB entries <5 Å, 21 entries <3 Å
Glutamate dehydro-genase at 1.8 Å[25]
230 EMDB entries <5 Å, 18 entries <3 Å (up to May 22nd)
435 EMDB entries <5 Å, 25 entries <3 Å
140 EMDB entries <5 Å, 10 entries <3 Å
Beta-galactosidase at 2.2 Å[80]
70S ribosome-EF-Tu at 2.9 Å[156]
Anthrax toxin pore at 2.9 Å[157]
20S proteasome at 2.8 Å[158]
Techniques
Achievements
EMDB entries

__Figure 1.1: Historical timeline.__ (a) Chart shows the EMPIAR entries released per year and cumulatively. (b) Timeline of cryo-EM with the selected key events that lead to the development of single-particle electron microscopy. Achievements of the techniques are shown on the upper part, while the lower part shows the structures solved. On the lower part, EMDB entries are shown as a single-particle method. (Reproduced with permission from Renaud JP et al, 2018 from[17]).

Electron crystallography revealed that freezing crystals reduces the effects of radiation damage[18]. This discovery led to the usage of cryo-samples (-140°C) in EM[19] and x-ray crystallography[20]. In the 1970s, Richard Henderson, a molecular biologist and biophysicist from MRC Laboratory of Molecular Biology in Cambridge, UK, along with his colleague Nigel Unwin, employed EM combining weaker rays and mathematical analysis to produce the first 3D model of helices arranged within the bacterial membrane of bacteriorhodopsin[21]. Along the same timescale, Joachim Frank, a biophysicist currently at Columbia University in NYC, developed with his colleagues image-processing software and produced a 3D structure from blurry 2D images. In the following years of the 1980s, the method of freezing samples in their native state, called vitreous ice samples, was invented by Jacques Dubochet[22] from European Molecular Biology Laboratory in Heidelberg. And finally in 1990, building upon all these advancements, Henderson was able to create the first atomic-resolution images of a protein using cryo-EM[23], thereby fathering the field of cryo-EM. Accordingly, in 2017, Dubochet, Frank and Henderson were jointly awarded the Nobel Prize in chemistry.

1.3 TEM setup

TEM allows the imaging of biological objects at atomic resolution. TEM generates an electron beam that transmits through the frozen specimen, gets scattered, and is later focused via the electromagnetic lenses onto an imaging device (e.g. a CCD camera, a fluorescent screen, a direct detector, or a photographic film). Figure 1.2 displays the schematic of TEM.

It uses a cathode to generate electrons and a series of anodes that serve to extract electrons from the source tip, focus them, and accelerate them to the desired voltage (typically between 200 and 300 keV. Condenser lenses adjust the physical size of the beam to project it on the biological object. The objective lens focuses electrons passed through the sample and magnifies the initial image, which is later projected on the fluorescent screen using the projection lenses.

In EM, the difference in electron densities of organic molecules in a cell represents a feature called “contrast”. Amplitude contrast and phase contrast are considered to be the two sources of contrast in TEM. Amplitude contrast occurs when the specimen absorbs the transmitted waves to a high degree, or scatters the electron to a large angle, and can result in changing the wave amplitude in the final image. Alternatively, electron interaction with the specimen can cause electron refraction at small angles, which changes the phase of the electron wave. Phase-contrast comes about when this scattered wave constructively and destructively interferes with the unscattered wave to produce an image.

The above-mentioned parts of the microscope column are responsible for producing the electron beam, focusing it on the specimen and projecting an image on the camera. This whole process must take place in a vacuum to minimize the collision frequency between electrons and gas atoms. The vacuum system, though not outlined here in detail, plays a significant role in maintaining the frozen sample in a near-native environment, as discussed in section 1.4.2.

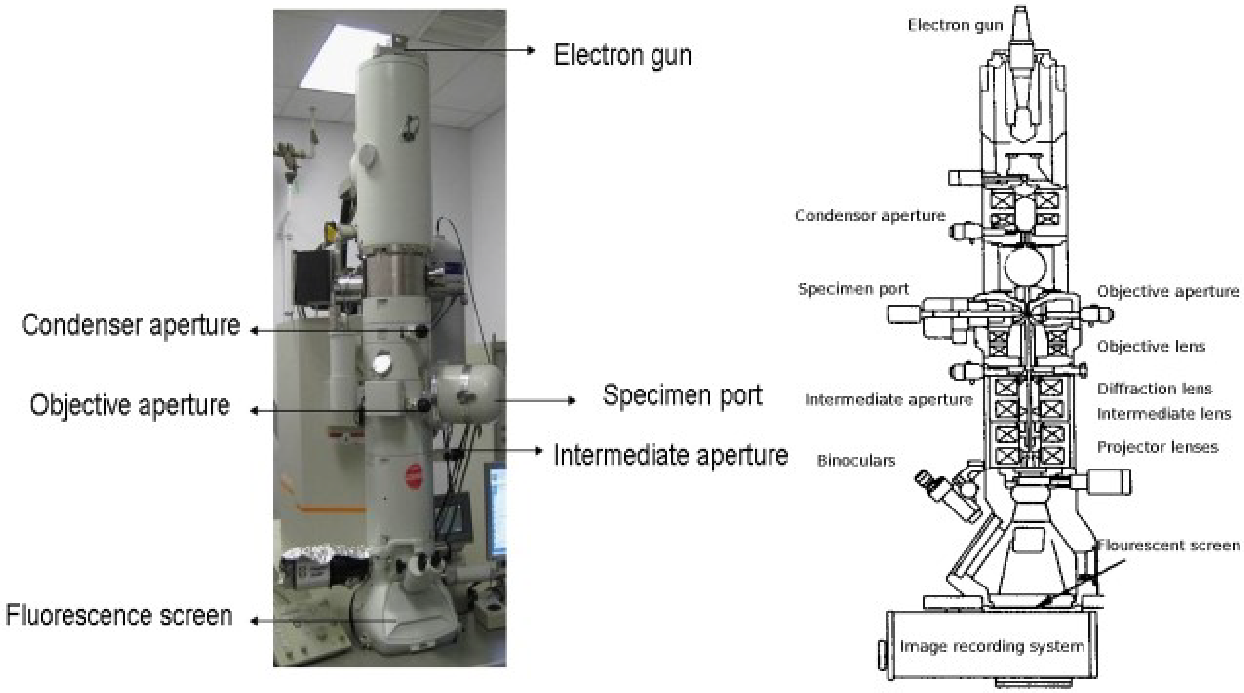


Figure 1.2: Principal components of TEM.
(Reproduced with permission from "Transmission electron microscope (TEM)," by Online biological notes for students, 2021 (http://www.biosciencenotes.com/transmission-electron-microscope-tem/). Copyright 2021 by https://wordpress.org from[24]*).*

For cryo-EM, the biological samples usually contain light atoms that interact weakly with electrons, where we assume that phase-contrast is the main source of contrast for cryo-EM.

A famous technique mitigating the effects of poor phase contrast, shown on imaged micrographs, is negative staining. The latter technique is well suited for collecting the initial observations about the sample such as shape or size. in the early stage. It involves applying heavy metal salts to the sample, resulting in the contrast between the stain (dark) and the specimen (light) where stain is excluded, for initial observations. However, usage of negative stain can cause loss of high-resolution information of the biological sample as well as introduce artifacts such as sample flattening.

1.4 TEM workflow

1.4.1 Sample preparation for molecular TEM

Sample preparation for biological macromolecules is a very complex and crucial task, as the specimen has to survive damage from both electron radiation and vacuum evaporation inside the electron microscope. Accordingly, the sample must be thin (to permit electron transmission) and stabilized to be introduced into the evacuated microscope column. To preserve the sample, different sample preparation techniques are used according to sample and study type, thereby maintaining its native structure. This section describes such sample preparation techniques.

First, sample treatment is performed to preserve intactness as well as obviate subsequent reduction of image contrast. In some cases, this includes sample stabilization by increasing the intramolecular interactions using crosslinking methods such as GraFix[25], removing crowding agents such as glycerol or sugar that can cause strong image background, and stabilizing transmembrane domains using amphipols, weak detergents, or nanodiscs in the case of membrane proteins.

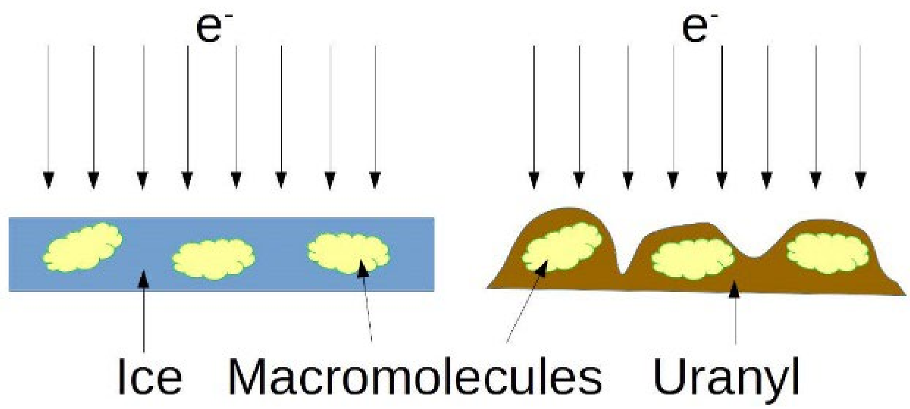
Sample preparation must be carefully orchestrated in order to avoid multiple paths of potential failure (e.g. water evaporation in the high-vacuum conditions, radiation damage). Using negative staining or vitreous ice freezing can resolve these limitations.

1. **Negative stain:** First, the sample is placed on a hydrophilic thin continuous carbon support film. After water wash and blotting, the sample is enveloped with a solution of heavy metal salt such as uranyl acetate[26]. This envelope of heavy metal atoms results in amplitude contrast and high SNR. Next on the carbon grid, excess of the solution is removed. After blotting, the stain dries to produce an electron-dense thin layer in which the particles are embedded[27]. Finally, the contrast is achieved due to the high-density difference between the macromolecule complex and the uranyl salt.

To improve stain quality, various buffer components, such as glycerol, detergent or specimen can be diluted with a specific buffer before their application to the grid. It is important to note that, on the micrograph, the macromolecule complex appears white on a black background, as illustrated in figure 1.3. Negative staining only shows the overall shape of the macromolecule and does not recover high-resolution information, owing to various factors. The main limiting factors are the grain size of the stain, the presence of uneven staining artifacts, dehydration, preferred orientation presence due to usage of continuous carbon film that limits sample views, and flattening of the samples, which can cause considerable structural distortions[28,29]. Despite the aforementioned, this technique is favoured because of its ease of preparation, high-contrast property as well as its ability to furnish information about the sample size, shape and arrangement.

2. **Vitreous ice:** This technique involves freezing samples quickly to reach a non-crystalline ice stage[30,31] as shown in figure 1.3. It helps to preserve the hydration of the sample without meddling with its shape. This technique also helps to achieve high-resolution reconstructions, enabled by the low contrast generated from the similar density between the ice and macromolecular complex. On the micrograph, particles look black against a white background. Combining this technique with DDDs has revolutionized the cryo-EM field in the last few years.

(i)



(ii)

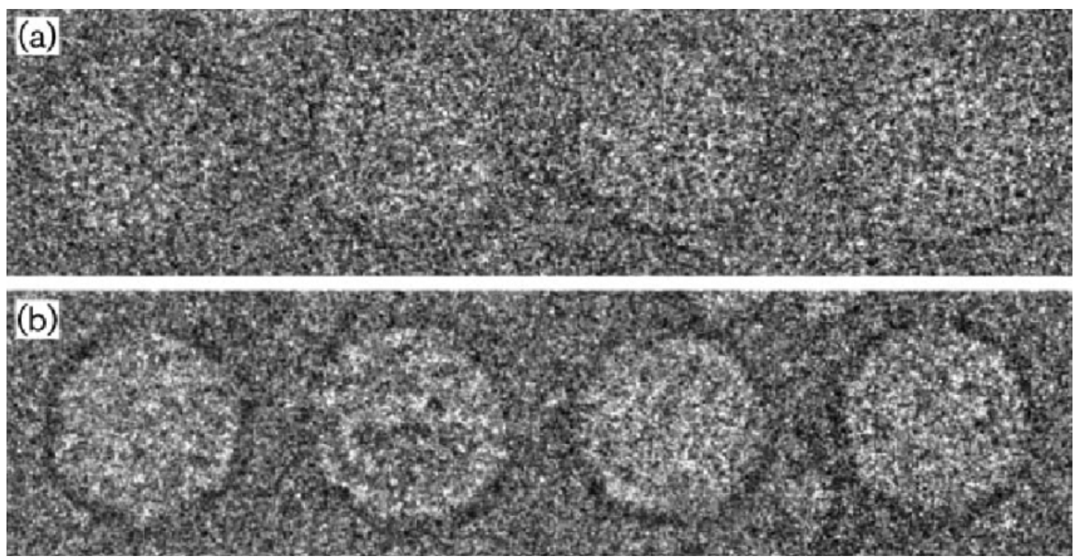


Figure 1.3: Vitreous ice technique. *(i) (Top left) vitreous samples and Negative stain (top right), (ii) with EM images of HcRNAV109 particles (a) embedded in vitreous ice and (b) Negative stain. (Reproduced with permission from Saibil HR et al, 2000 from[32] and Miller JL et al, 2011 from[33]).*

1.4.2 Maintaining a near-native state

Generally, biological macromolecules in an organism exist in a partially hydrated (embedded in lipid membrane) or fully hydrated state and perform their functions in an aqueous environment as well. Therefore, it is very crucial to analyze these macromolecules in their native state to understand their true structure and biological function. Cryo-EM enables structural analysis of macromolecules in their near-native environment by performing their rapid freezing and then placing them into the electron microscope column, while kept under high-vacuum and in a low-temperature state. Both these conditions minimise the effect of radiation damage on specimen[34] and keep the ice in an amorphous state. Based on the temperature-dependent modifications, ice can exist in various forms, namely cubic (− 123 °C to − 148 °C, 115 to 150 K) and hexagonal crystal (above − 103 °C, 170 K)[35]. However, during the rapid freezing treatment, ice skips the crystalline phase to form an amorphous phase of ice, which is generally lower in temperature than the crystalline transition phase, called vitrification. Among various ultrarapid procedures formatted for general vitrification procedure, the most widely used is the plunge-freezing method, where after removing the excess solution, an EM grid is rapidly plunged into the liquid ethane (cooled using liquid nitrogen)[36]. Ethane is one of the commonly used cryogens for vitrification of sample because of its adequate boiling point (184 K), freezing point (90 K), high heat capacity (68.5 J/ mol K at 94 K) and high thermal conductivity. These all conditions are necessary to prevent the formation of a vapor layer between the sample and the cryogen. As a primary coolant, liquid nitrogen helps to keep ethane liquefied and cool during the procedure of vitrifying the sample. Liquid nitrogen is the commonly used condition for handling cryo-EM samples (-193 °C, 80 K)[37], as the lower temperature protects from secondary chemical reactions and retards the displacement of molecular factors caused by ionizing radiation, generally known as the cage effect [38,39], thereby increasing the SNR. Cryo-EM has two main sub-disciplines to determine the structure of a 3D macromolecule, viz. ECT and SPA.

1.4.3 Imaging with electrons

1.4.3.1 ECT: Several images of a sample from the same region are acquired by tilting it by

various angles (tilt degrees range to 60°–70°) with respect to an incident electron beam. This series of images, or tilt series, is later combined computationally to form a 3D map, called a 3D tomogram. The diagrammatical information can be seen in Figure 1.4[40].

In recent years, this technique has allowed us to visualize 3D structural details for biological macromolecules *in vivo* at 2–6 nm resolution[41,42]. During sample preparation, chemical fixation and staining can alter the macromolecular organization of the cell. However, rapid freezing to vitrify the sample (as detailed in sections 1.4.1 (2) and 1.4.2) yields pristinely preserved samples. ECT has progressively gained importance to decipher a large range of specimens, from isolated protein complexes to large eukaryotic cells as well as the molecular architectures of viruses, bacteria and cellular components *in situ*[43–45]. Moreover, it provides the opportunity of understanding the spatial relationship of macromolecules within a cellular tomogram. For advanced analysis, similar sub-tomograms or sub-volumes can be selected for further alignment and averaging, revealing additional structural information[46]. This whole procedure is referred to as ECT STA.

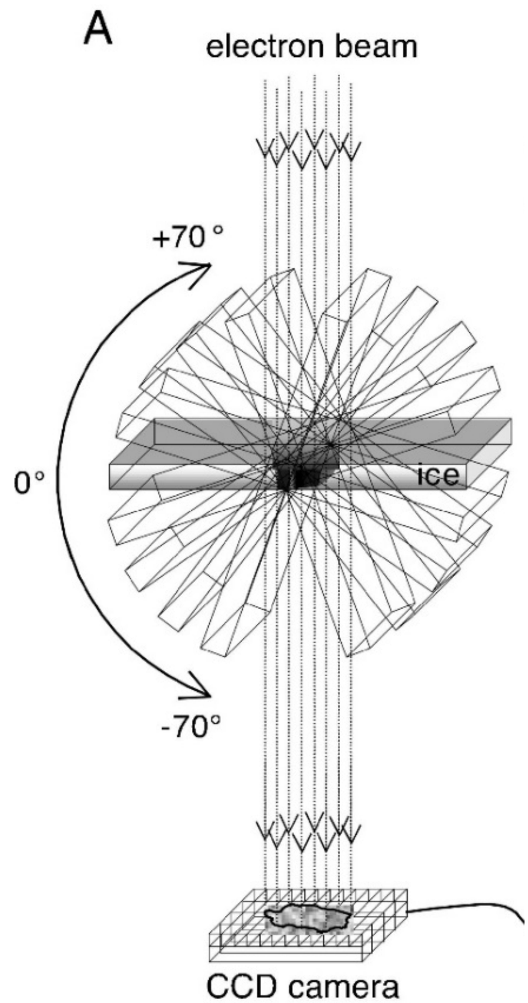


Figure 1.4: Principle of electron tomography[40]*. Electron tomography works by collecting a sample's projection caused by an electron beam passing through the sample, within the defined degrees of rotation (between 60°–70°) around the centre of the specimen. (Reproduced with permission from Steven A et al, 2005 from*[40]*).*

Beyond analyzing secondary structure elements, STA has yielded high-resolution density maps for nuclear pore complexes[47,48], chemotaxis signaling arrays[49], coat protein complex I[50], polysomes[51], ribosomes[52], retrovirus assembly[53–57] and bacteria surface layers[58]. However, when compared to cryo-EM SPA, STA generates lower resolution due to several reasons[59–61] such as sample thickness increase while acquiring tilting series, distortion of densities in tomogram, or

3D-image artifacts known as missing wedge, the latter which affects subtomogram alignment accuracy and classification procedure. Another challenge is the usage of the thick sample in cryo-ET, which further increases during the tilt series procedure. Thus, the defocus gradient from sample thickness as well as the sample tilt has to be considered while going ahead[62]. Thicker samples also suffer from poorer image quality due to inelastic and plural scattering. To avoid damage from the cumulative irradiation, the electron dose is limited per tilted image, resulting in low SNR when compared to SPA. Nonetheless, STA has an advantage over SPA in the case of 3D classification, where each particle exists as a unique 3D reconstruction, allowing analysis of various conformations adopted by a biological object in the form of direct 3D variance.

1.4.3.2 SPA (basic overview):

The main goal of SPA is to determine the high-resolution maps from a macromolecule of interest using EM. SPA is based on two assumptions to achieve this goal:

1. All the particles in the sample are identical copies of macromolecules in the same conformation but showing different orientations. Without the presence of such a condition, the sample is said to be heterogeneous, which is a challenge for SPA. This hypothesis is called the identity condition. It is considered a weak condition as the slight presence of heterogeneity doesn't affect the sample much. On the other hand, a highly heterogeneous area can cause a blurring effect in the final reconstruction.
2. Micrographs are electron-density projections of the object and have the same magnification. This is called the scale condition.

When a suspension of biological molecules or complexes is prepared for imaging, it adopts an arbitrary position on the grid, or in the ice in cryo-EM. Thus, multiple copies and different orientations of the same macromolecule are imaged on a single micrograph. The orientation of the molecules in the images is unknown and represents the key set of parameters to be determined in SPA. Therefore, to be able to reconstruct a 3D map, Euler angles (planar rotation angles around x, y, and z axes, whose values depend on the choice and order of the rotation axes) and shifts are employed to describe these unknown orientations of the particle images in 3D Euclidean space, a space in which postulates of geometry can be applied with linear and finite-dimensions[63]. Using an initial volume, which can be estimated from previous knowledge of the sample or by image processing methods[64], the orientation parameters are estimated by comparing the individual particle images with projections of the reference volume. After finding the Euler angles, particle images are projected back to form the final 3D reconstruction of the specimen using the reconstruction algorithm[65]. SPA's particle images suffer from the low electron dose, which leads to a low SNR. This problem is resolved by aligning and averaging many similar particle images (typically around ~10k-100k particles). The standard reconstruction workflow follows particle selection, particle alignment, particle classification, 3D reconstruction, and model refinement[66,67]. Detailed information is given in section 1.4.4.

However, TEM application to biological samples suffers from resolution limitations[68] due to several reasons such as high voltage beam damage to specimen, small electron doses used for imaging, BIM, and heterogeneity of the sample. BIM is a phenomenon whereby the energy deposited on the sample by the beam causes the sample to move, often because of the relaxation of tension within the vitreous ice. This motion leads to the blurriness of images and thus limits the final reconstruction resolution[69]

Some of these limitations have been overcome by advances in image acquisition and image processing enabling superior resolution, 0.6 Å[70] for micro-ED and 1.15 Å[71] for single particle analysis. Electron microscopes have undergone various improvements over the decades in case of electron sources, vacuum, sample holders, power supplies, optical lens systems, and detectors. In terms of electron sources, tungsten filaments and LaB6 sources have been upgraded to field emission electron sources to generate a more coherent electron beam for preserving the phase of the structural information of a macromolecule. A detailed overview of each image processing step in the SPA workflow with advances is presented below.

1.4.4 SPA image processing workflow

In the first step of this workflow, images in the form of movies or micrographs are acquired. With the high sensitivity and fast frame rate of modern recording devices, cryo-EM data can be recorded as movies. Individual frames of a movie are aligned in order to correct BIM and to acquire a single, de-blurred micrograph. In the next step, the CTF is estimated to correct possible aberrations caused by EM that would otherwise modify the image. Subsequently, the particle picking step includes selecting and extracting particles in the micrograph. After the CTF correction, particles of similar orientation are grouped and averaged into 2D classes by performing 2D classification. The classification step checks the quality of data by removing poor particle images, e.g., those containing a high amount of noise. Averaging the particles within 2D classes increases the SNR and helps to determine the initial volume for the next step. An initial volume provides the first and coarse estimation of the macromolecular structure. Following that, initial volume is used to reclassify the particles and to find the different conformations of the same volume. In the refinement step, maps are enhanced by the assignment of better angular orientations.

Figure 1.5 describes the image processing workflow. A detailed overview of each step is given in section 1.4.4.

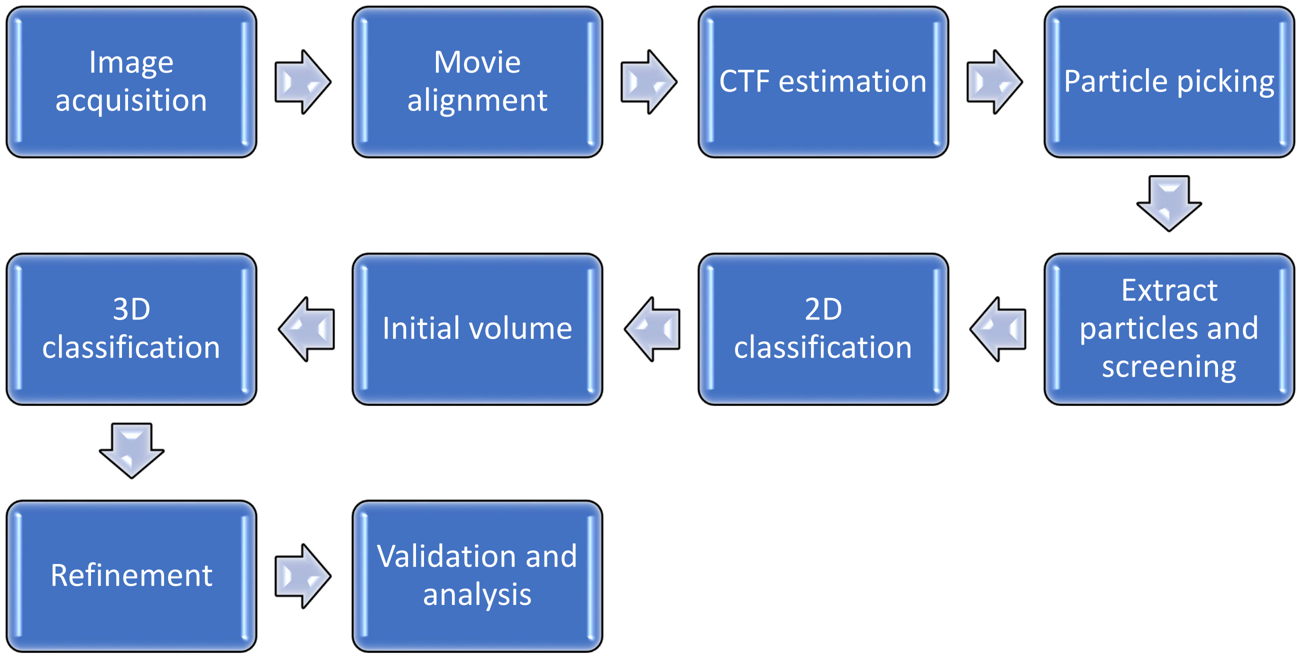


Figure 1.5: Main steps in SPA workflow.

1.4.4.1 Movies and micrographs

High contrast between the ice (background) and the complex (signal)[72] is a necessary component for observation of structural details. The small difference in atomic number between the vitreous ice and the specimen can limit such contrast, thereby compromising structural information about the macromolecule. Increasing the electron dose[73,74] is one solution but will lead to radiation damage of the specimen, since high-energy electrons break bonds when they interact with the specimen.

Hence, in single-exposure images, electron dose is kept below ~20 $e^{-}/Å^{2}$ to achieve high resolution. Instead of taking a single image (called a micrograph), DDDs work in movie mode by taking multiple frames in the form of a motion picture at the same electron dose. For electron-counting DDD cameras, the electron dose rate is typically kept below ~10 e^{-}/pixel/sec, depending on the camera. DDDs make use of a superior DQE, which measures the combined effect of signal and noise performance of the camera. They do this by detecting individual electron impact events on the sensor, which are then collected into dose-fractionated image stacks, called movies, with the help of high-speed CMOS technology. These movies can be aligned to computationally correct specimen movements. The aligned frames are averaged later to use for structure determination[75]. The result is an image with reduced blurriness, since the alignment routine can adjust for BIM and specimen drift.

This movie mode of DDDs helps to optimize the SNR of images affected by radiation damage. Early frames have been subjected to a lower total electron dose, and thereby possess a higher resolution signal, but may be affected by fast specimen movement. In later frames, specimen movement decelerates, but a higher accumulated dose results in reduced high-resolution information. Finally, a relative weight is applied while averaging these frames to optimize the signal in the final average[76]. This allows for higher total dose exposures, since the weighting during dose compensation will take care of the high-resolution noise in later frames, but keep the low-frequency signal. The next step includes the processing of movies. One technique is to average each movie and reduce it to micrograph images.

The alignment can be executed at two levels, micrograph (global)[72,77–79] or particle (local)[80,81].

1. **Global:** The goal is to align the frames of a movie and present a corrected average. The first alignment approach called MotionCor[82] uses this approach by estimating the relative shift between two frames of a micrograph using correlation. In the next step, frame alignment is performed with the known displacements.
2. **Local:** The BIM is local as different particles show different movements. Some methods in this category impose prior knowledge assuming that particles which are close in the

micrograph will move similarly. This approach is used by Optical Flow[77], alignframes_lmbfgs, and alignparts_lmbfgs[81], Unblur[83](without local tracking), and Summovie[72] or MotionCor2[84](performs local tracking). Another way to perform local alignment is to divide the frames and grids, and then track each grid independently without *a priori* knowledge. If the particle position is known, each particle can be tracked. Such a function can be accessed by using software packages, including *Relion*[85].

The descriptive information of all these approaches can be found in a publication of *Ripstein JA et al, 2016*[86]. Figure 1.6 shows the advancement in technology to record data and generate unprecedented quality reconstructions.

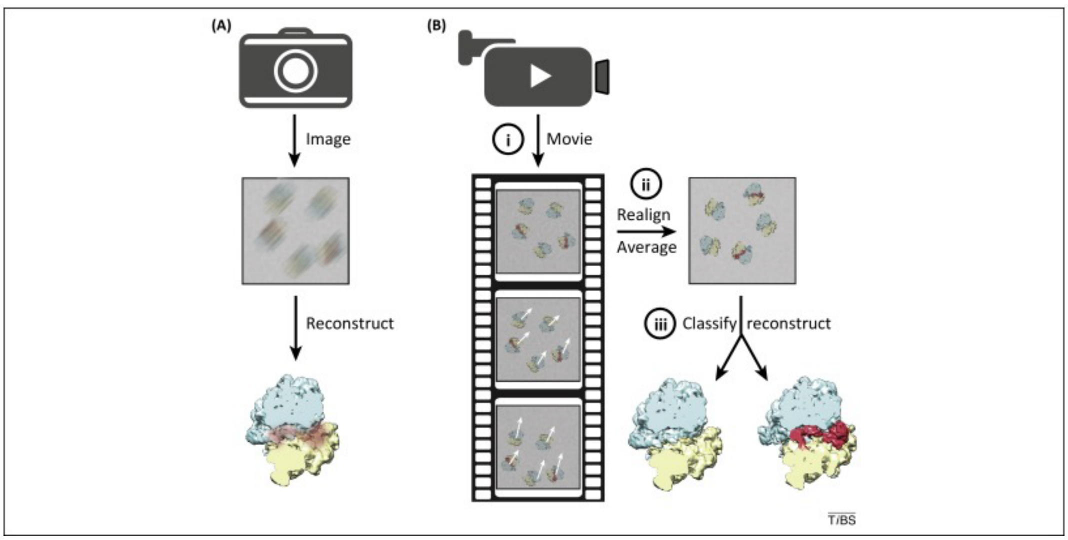


Figure 1.6: Images and movies. *(a) Before, photographic films were used to record the noisy images with blurriness caused by BIM and a mixture of structurally different particles combined into a single reconstruction. (b) Due to recent advances, better reconstructions have been achieved. (i) a movie of significant-good quality is recorded as a set of frames using DDDs; (ii) movie frames are aligned and averaged to compensate for sample movement; (iii) dynamic classification methods generate different conformations from a sample containing a mixture of states. (Reproduced with permission from Bai XC et al, 2015 from*[66]*).*

1.4.4.2 CTF correction

The TEM imaging system produces a systematic artifact in the image contrast. The CTF is evident in the Fourier transform of a projection image as an oscillating function that depends mainly on defocus and the spherical aberration coefficient of the objective lens[87]. Regulation of the electron micrograph signal can be described by an oscillating non-linear CTF in the form of an equation given as:

$$\mathrm{CTF}\ (f) = E(f)\sin\left(\pi\mathrm{Cs}\ \lambda^3 f^4/2 - \pi\lambda d f^2\right),$$

where spherical aberration of the electromagnetic lens is represented by Cs; λ shows the electron wavelength; f shows the spatial frequency, d is the applied defocus (negative for underfocus), and $E(f)$ is an envelope function showing degradation of high-resolution information. [Equation taken from *Erickson HP et al, 1971*[75], *Wade RH, 1992*[76].]

By estimating the CTF parameters, the effect of the CTF can be reduced computationally during the reconstruction process. CTF estimation measures the defocus of each image because the Cs term and electron wavelength are constant for each cryo-EM dataset. As the spatial frequency increases, the function begins to oscillate more rapidly; thus, estimating the CTF parameters is essential for retrieving high-resolution information. Different ways to correct CTF include application of CTF estimation to the image, phase-flipping[89] and Wiener Filtering[90]. The CTF for the standard TEM imaging system is shown with the starting point near zero (low spatial frequencies region), followed by oscillation between positive maxima (positive CT) and negative maxima (negative CT) with zero-crossings that represent the frequencies with zero information.

Analysing the power spectrum of the cryo-EM image can show the oscillations in CTF, as depicted in Figure 1.7, where the pattern of the concentric ring reflects the changing of high and low CT as a function of spatial frequency (for high-quality images). The origin of power spectrum or image's centre represents the low spatial frequencies, while the edges show high-spatial frequencies, also known as the Shannon-Nyquist limit (2 x pixel size).

If the micrograph is of good quality, it contains a CTF pattern of several concentric rings, as shown in Figure 1.7 (left). For poor quality micrographs, CTF pattern has specific asymmetric rings or elliptical rather than circular rings (astigmatism) as visualized in Figure 1.7 (middle); or rings fading in a particular direction, indicated by a directional fall-off in the pattern of Thon rings (drift) as seen in Figure 1.7 (right). These bad-quality micrographs can be excluded from the further processing steps after the CTF pattern analysis. Therefore, this step is helpful for the screening of the micrographs.

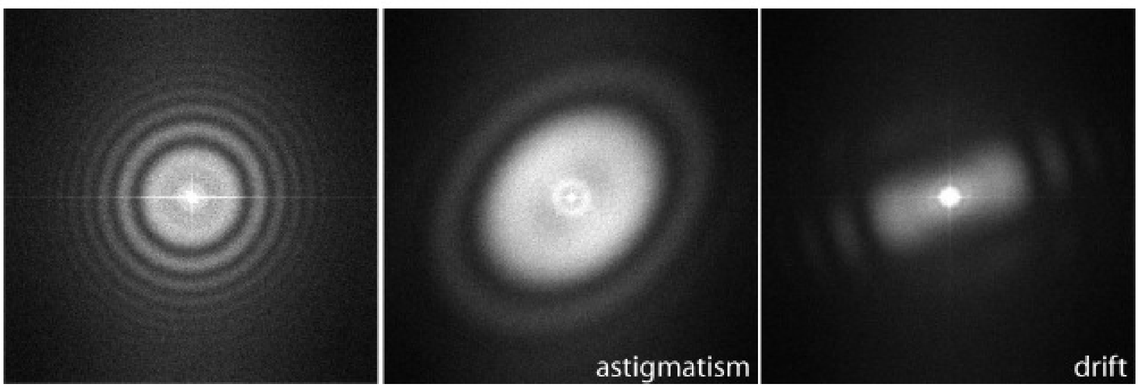


***Figure 1.7: CTFs**. CTF of good (left), astigmatic (centre) and drifted (right) micrographs, respectively. (Reproduced with permission from Scheres, SHW et al, 2008 from[91]).*

1.4.4.3 Particle picking

Particle picking involves identifying particles from the background noise in the micrograph. In this step, particles are selected and cropped from the micrograph to be further processed for structural analysis. The low SNR of cryo-EM micrographs, particularly for smaller molecules or complexes, renders this step difficult, but not impossible, to automate. There are many methods developed for this purpose and can be classified as:

1. **Manual picking:** As the name suggests, particle picking is performed via the user's bare eyes. This type of selection can be time-consuming and can lead to bias based on the user's subjectivity**.**
2. **Automatic picking:** The automatic particle picking methods are based on deep learning solutions involving CNNs, such as DeepPicker[92] or DeepEM[93]. In other methods, for example, APPLE Picker[94], a template is automatically chosen by the algorithm. Approaches such as *Gautomatch*[95] use GPUs to accelerate the process of particle picking. Other methods such as *gEMpicker*[96] and *DoGPicker*[97] operate by subtracting two Gaussian blurred versions of the same image resulting in a Difference of Gaussian (DoG) map. In the following step, particles are sorted based on size and then extracted. By contradistinction, CrYOLO[98] employs a deep-learning object detection framework colloquially styled as "you only look once"[99], and can detect particles (after training with at least 200–2500 particles per dataset) with an impressive speed of five micrographs per second. Methods implementing automatic particle picking remove the limitation of bias based on manual picking or the user-selected template.

1.4.4.4 2D classification

The 2D classification step consists of the clustering of the particle images with similar projection directions so as to average them after proper alignment, which – in turn – allows calculating 2D class averages with improved SNR. These 2D class averages can help assess the quality of the particle set by the identification of secondary-structure elements, e.g., alpha-helices can be seen in the high-resolution class images. Additionally, the analysis of different 2D class averages can provide an early idea of heterogeneity in a dataset.

2D class averages can be determined by iterative approaches such as MRA[100–102,85,103]. These methods first generate a set of 2D classes by evenly and randomly assigning particle images to them. From these initial 2D classes, 'reference' images or 2D averages are calculated by averaging corresponding particle images. In the next iteration, particle images at different 2D orientations and shifts are compared with the 'reference' images, calculating for each case a similarity metric that could be the correlation[100,101] or an empirical likelihood of the matching[85,102,104]. Next, 'reference' images of 2D classes are updated by weighted averaging particle images according to associated likelihoods in maximum-likelihood based approaches[85,102,104] or by averaging particles at the orientation and class that maximized the correlation[100,101]. This process is typically iterated a few dozen times, generating a final set of 2D class averages. Note that ML2D based approaches as RELION[85] 2D classification assume that each particle image participates in every 2D class and at every 2D orientation but contributing with different weights. On the other hand, correlation-based approaches demand that each particle contribute only in one 2D class with one 2D orientation. Thanks to GPU paralellization[105], RELION has become a popular method for fast processing speed. However, the outcome of RELION (or any ML2D based approach) can suffer from the attractor problem[106] wherein particle likelihoods are biased toward classes containing more particles. Thus, these methods may tend to classify particles according to the SNRs of the different classes rather than by the actual conformation of the particles.

Currently, ISAC is one of the well-used 2D classification approaches, as well[107]. ISAC ensures the homogeneity of each class by performing repeated stability tests to validate each class' member. It can be a favoured approach for analysing heterogeneous data because each class size is restricted by using modified K-means which reduces the attractor problem faced in RELION. ISAC may not require human intervention for selecting good classes, as it can automatically remove unstable or non-reproducible classes. However, due to being highly time-consuming, ISAC's utility is limited to difficult cases.

The MSA approach, which flows from PCA, involves reference-free alignment[108], and represents a powerful tool to sort out structural heterogeneities in cryo-EM samples. MSA converts the particle images into a linear combination of its main eigenvectors and performs the classification based on similar orientations of particle images[109–113]. As such, particle images in the dataset are compared using distances and correlations, each with its own metric system. One of the commonly used metrics is Euclidian, which is based on the assumption that the smaller the distance between the two particle images, the higher the correlation between them, and when their distance is zero, their correlation is at its maximum. This metric is associated with PCA[108].

Figure 1.8 shows class averages of 256 classes, where some poor-quality particles are eliminated to obtain a clean dataset suitable for the downstream pre-processing steps.

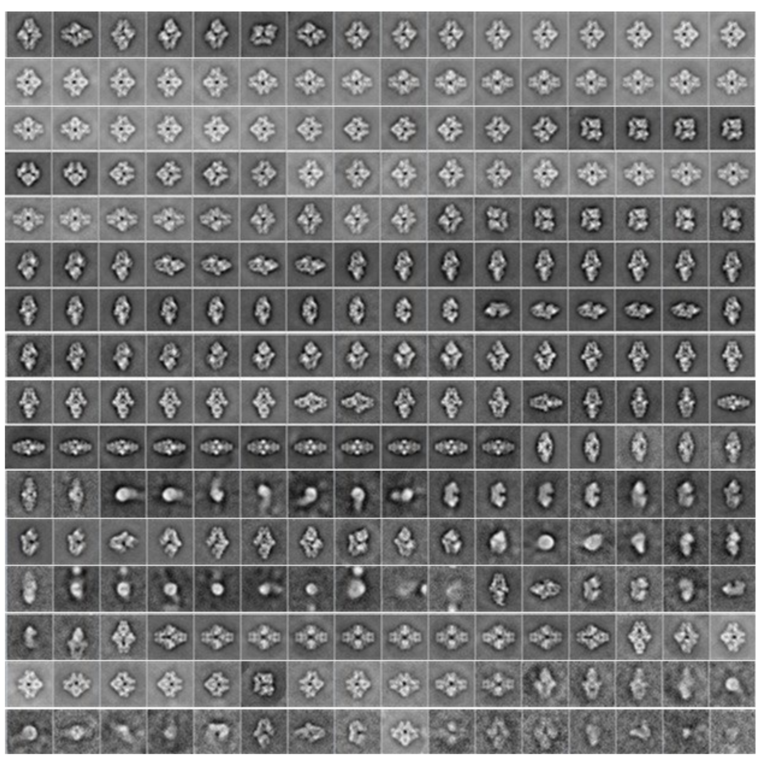


Figure 1.8: 256 class averages after classification of 90,503 β-galactosidase particles. *(Reproduced with permission from Sorzano COS et al, 2010, Sorzano COS et al, 2015 from*[101,114]*).*

1.4.4.5 3D reconstruction

The Fourier slice theorem establishes a relationship between a 3D map and a 2D projection, as explained below**.** The first step in 3D reconstruction is to select the projection of the macromolecule, called particles, and combine them to form the final structure. Using high numbers of particles provides better coverage of the projection sphere as well as increases the SNR of the final reconstruction. This statement is based on the conventional resolution determination method known as Crowther criterion, presented in following equation:

$$N_\theta = \pi D/d,$$

where N_θ is the minimum number of views required to obtain full Fourier domain sampling for 3D reconstruction of the object; D is the size of the object for imaging; and d is the isotropic spatial resolution (equation from Jacobson C et al, 2018)[115]. The final goal is to reconstruct the 3D structure of the macromolecule with all suitable particles.

Among various approaches[116], the central slice theorem[117,118] establishes a remarkable relationship between the 2D projection images (particles) and the 3D reconstruction, and is the basis for the second assumption in section 1.4.3.2**.** The Fourier central slice theorem states that the Fourier transform of each 2D projection is equal to a central slice through the 3D Fourier transform of the 3D object perpendicular to the direction of projection.

Figure 1.9 shows a graphical view of the theorem, whereby a 3D object or molecule is defined as function $f(x)$. The projection of this function along a certain direction d is shown as $P_{\alpha,\beta}[f]$, with

$$P_{\alpha,\beta}[f](r,\psi) = \int_{-\infty}^{\infty} f(r,\psi,s)ds,$$

(Equation from Barnett A et al, 2017 from[119]*).*

where (r, ψ, s) represents a cylindrical coordinate system in $\mathbb{R}^3$, and (r, ψ) donates the polar coordinates in the projection plane orthogonal to d, and s is the component along with d. The orientation vector is represented by $\mathrm{d} = (1, \alpha, \beta)$.

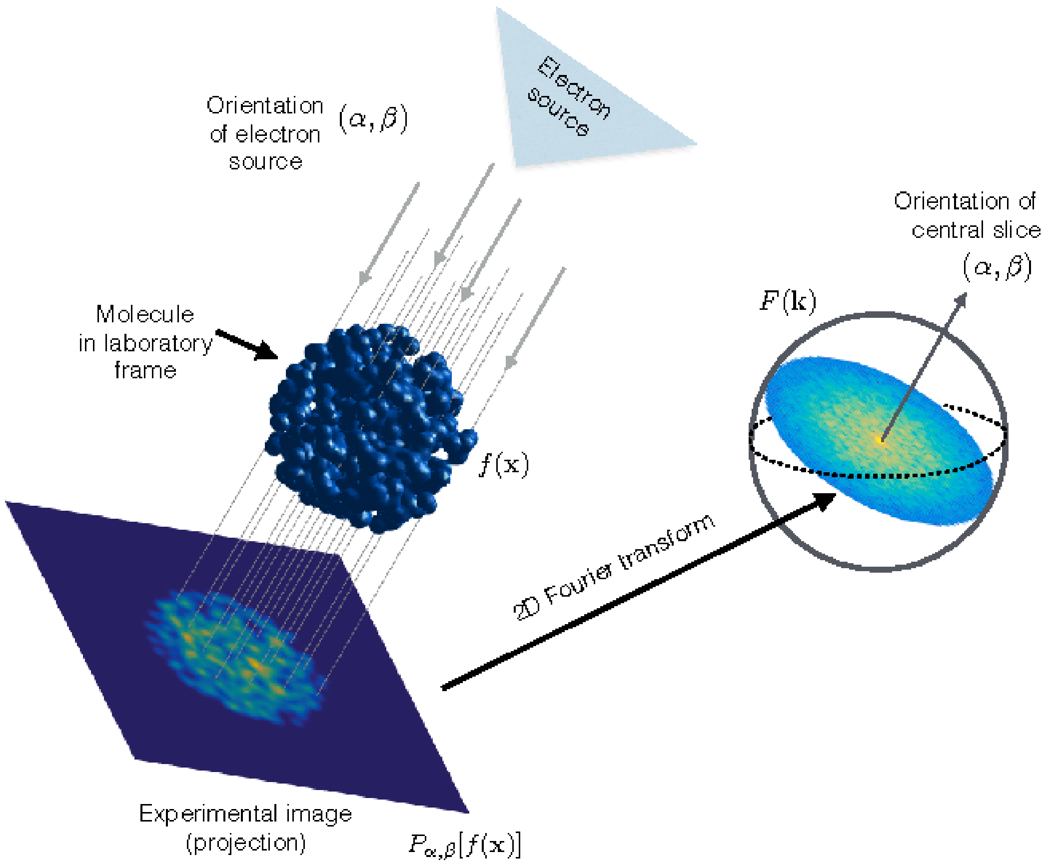


Figure 1.9: Principle of Central slice theorem *(Reproduced with permission from Barnett A et al, 2017 from[119]).*

Although the central slice theorem helps to reconstruct the 3D structure of the macromolecule from the projections (particles), the orientation of these projections remains unknown. To compute the orientation of the particles, an initial volume is required in SPA.

1.4.4.6 Initial volume

The initial volume is a coarse and low-resolution first estimation of the 3D structure of the macromolecule, and is employed to obtain the orientation of the single particles. This initial and low-resolution structure is typically generated by harnessing a few classes with sufficient SNR. The determination of a “good enough” initial volume is critical in SPA because choosing a poor initial volume can lead to bias in the final map[120] as would occur, for example, by manually selecting the best initial map from several obtained low resolution initial maps based solely on the user's experience[121–127].

A proper initial model can be obtained in a variety of different ways. It can be an already-available similar macromolecular model, which should be low-pass filtered to typically 40-90 Å to avoid bias issues such as the whimsically termed the “Einstein-in-noise” problem. In this issue, recorded images of particles are in reality pure noise from which even a portrait of Einstein can be extracted. This was a result of using data with low or zero contrast, small or invisible particles as well as automatic particle picking[128,129].

However, this initial model may not be available in all cases. So, another option includes using one untitled and one tilted orientation to obtain a set of tilt pairs. The tilt pairs undergo alignment and classification to provide Euler angles, which can be further harnessed for an angular assignment without the need for a reference volume. This approach is called RCT[130–132]. Another method performs *ab initio* reconstruction based on common line methods[117,133–140]. This framework usually considers the Fourier transform of different projections that will intersect in a common line. This common line can help in determining the angular assignment of the classes.

Other methods for initial volume generation include RANSAC[141] and Significant[142]. RANSAC works by assigning random orientations to different class averages and then computing scores for each volume by comparison between the class averages. Volumes with higher scores are selected for further processing. As an example, Figure 1.10 shows an initial map generated from RANSAC of the β-galactosidase sample. By contradistinction, the Significant approach employs statistical methods to calculate weights based on the cumulative density function, thereby deciphering different image similarity measures.

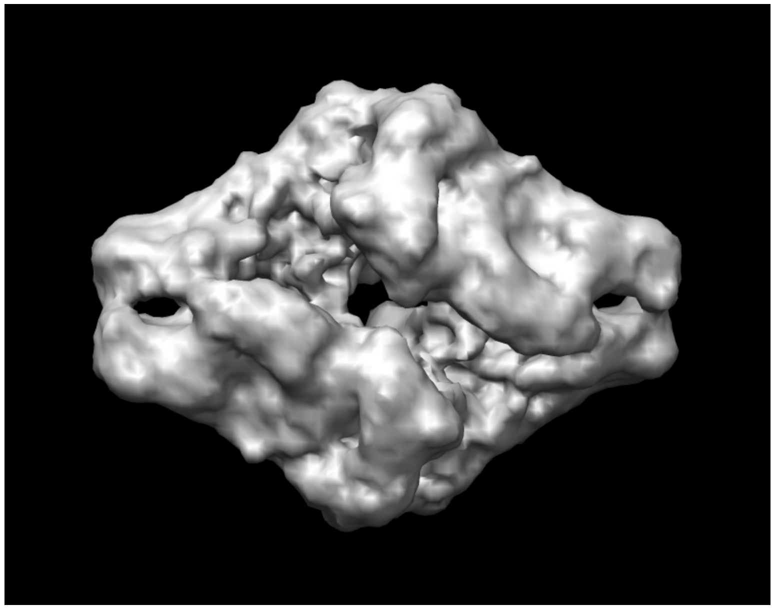


Figure 1.10: Initial volumes for β-galactosidase generated by RANSAC method. *(Reproduced with permission from Sorzano COS et al, 2015, Vargas J et al, 2014 from*[114,141]*).*

1.4.4.7 3D classification

3D classification deals with the major problem of heterogeneity in cryo-EM[143]. Heterogeneity can be present as compositional heterogeneity, in which the macromolecular complexes assume different biochemical states and adopt different structural arrangements. This may include the occupancy of binding partners in complexes. Another form is conformational heterogeneity, where some macromolecules are not rigid and have a certain degree of flexibility, leading to slight but non-negligible differences in their structure. Presence of either of the above-mentioned kinds of heterogeneity shatters the assumption of SPA workflow, viz. that all particles are identical copies of a macromolecule but in different orientations. Hence, to continue the further processing, a set of particles belonging to one particular conformation of the macromolecule should be grouped together accordingly, i.e. 3D-classified. The most commonly used method for dealing with heterogeneity is ML3D, a maximum likelihood approach integrated into RELION[85,144–146]. Another type of method includes defining a phase space that contains all possible conformations, using vibrational modes and dynamics of the macromolecule[147–149]. PCA[150], Bayesian marginalization algorithms[151], and analysis of covariance matrix[152] are among the alternative approaches employed to this end.

1.4.4.8 Refinement

After obtaining a homogenous set of particles (or projections) of the same macromolecular complex, refinement is executed to obtain a high-resolution map. Usually, single-particle EM packages use a 3D-projection matching procedure for structure refinement, in a manner of a more or less elaborate version. It involves modifying the orientation parameters of projections for achieving a better match with reprojections computed from the current approximation of the structure[153].

This step is equivalent to solving a linear equation A**x** = **b**, where A is the projector or can be called P_θ, **x** is the structure to be determined (also represented by V), and **b** are the particle images. This equation can be solved from a given homogenous set of N particle projections of the same structure V captured from different orientations. The goal is to minimize the distance between the image and the projection of the structure, $P_\theta V$, along the direction θ, in order to determine Euler angles and shifts.

Maximum likelihood[103,144,154,155], Maximum *a posteriori* (RELION approach[85]) and the traditional projection matching[156–158] are a few examples that integrate the above-mentioned equation to perform the refinement step. Maximum likelihood is based on the assumption that a given particle image can present in all possible orientations but with different probability weights. The goal is to decrease the possible number of directions by collapsing the probability. It can be explained as:

$$(\theta^*_i, V^*) = \text{argmin}_{\theta_i, V} \| I_i - P_\theta V \|^2_w,$$

where I_i are the particle images with i = 1, ...,N [such that N represents homogenous set of particles]; P_θ is the projector along the direction θ; V are the projections of the structure; and 'w' is the given weight. (*Equation taken from Sigworth FJ et al, 2010*[104]).

On the other hand, *a posteriori* approaches work by penalizing some of the possible orientations discussed above by introducing prior information. This step of refinement is distinguished by splitting the maximum likelihood into two tasks, namely, a) assigning angles to the particles according to the initial map; and b) reconstructing a new map by using the assigned angles. The experimental images are then compared to measure compatibility.

Progress of refinement is analysed by various indicators, especially the FSC, which reflects the level of SNR as a function of spatial frequency as well as the map's resolution[159]. An FSC curve is obtained by comparing the Fourier transforms of two maps over shells of the same spatial frequency. The comparison between maps is orchestrated by computing respective correlation coefficients. This FSC curve should decrease with spatial frequency until a resolution limit is reached, as indicated by a cut-off threshold[160]. The 'resolution' value in single-particle EM is the spatial frequency at which the SNR or FSC curve crosses a certain threshold value. For example, resolution is the spatial frequency where FSC is equal to 0.5 or spatial frequency where SNR is 1.0, generally a level where the power of signal is equal to the power of noise. Another typically used threshold is FSC = 0.143[161], based on relating EM results to those in X-ray crystallography[162].

A common issue observed in structure refinement is the so-called 'overfitting' of data, which occurs in the EM map due to alignment of noise instead of signal. Caused by lack of careful judgment, it leads to obfuscation and confusion between signal and noise[163]. Therefore, by chance artifacts are created which are further increased by alignment of the noise components in the data, resulting in inflated FSC values and an artificially high resolution. To avoid the issue of exaggerated resolution estimation using the FSC, one has to ensure independence of noise in the half-dataset maps used to calculate the FSC[164]. This approach is called the "gold standard" refinement procedure[85,165]. Another way to avoid overfitting during iterative structure refinement is the elimination of high-resolution data in the alignment step[166].

1.4.4.9 Validation and analysis

With the completion of the reconstruction procedure, the next step involves validation and analysis of the final structure achieved. The validation step is performed due to:

a) low SNR of particle images.

b) user decisions in the SPA workflow that lead to bias in the final reconstruction.

There are many quantitative methods for the validation of the final 3D structure obtained. A final 3D reconstruction can be compared with results of other techniques such as NMR, X-ray crystallography, or already existing similar structures from PDB[167] or EMDB[168].

Tilt-pair validation analysis[169–172] can be used to quantify the accuracy in the orientation estimation of particle image in SPA, yet has not been broadly adopted by cryo-EM practitioners because it increases the amount of necessary data to collect and process. There are also proposed methods for gauging particle alignability, viz. the precision in the orientation estimation of single particles that do not require tilt-pair acquisition[173,174]. The latter approaches can provide metrics to detect incorrect reconstructions under the assumption that the distribution of likely orientations of a particle image with respect to the obtained 3D map should be clustered as opposed to random.

An issue called overfitting, introduced before in section 1.4.4.8, can be detected at this stage. To identify such overfitting, one possibility is to take the 3D map with a small subset of particles used in the reconstruction. These particles are realigned to the map without masking or resolution limits, thereby obtaining a new 3D reconstruction. Another solution is to computationally replace the random subset of particles with noisy images. Overfitting may be detected by the analysis of obtained resolution in either case. [175].

After the 3D reconstruction is validated, its spatial reliability must be verified via map resolution estimation. Such resolution can be calculated by several approaches, including FSC[176], DPR[157,177], and SSNR[178,179]. FSC, considered as the standard measure, discovers the cross-correlation between two 'half maps' at different spatial frequencies, where each map contains a random half-subset of the data, explained in section 1.4.4.8.

Locally varying resolution may result from sample heterogeneity and image processing errors that cannot be estimated by the standard FSC. Blocres[180] rendered the first attempt to calculate the local resolution of a 3D map by measuring the FSC between two maps within a moving window. However, major problems arise concerning the adjustment of window size and the need for two half maps. Alternative local resolution-determining methods include ResMap[181], MonoRes[182] and DeepRes[183]. It must be noted that the highest nominal resolution of a cryo-EM map is not necessarily the best one, because lower frequency values often matter more for map connectivity and interpretability. Therefore, rather than treating a firm number, the reported EM map resolution should be considered as a broad guideline[184].

A decades-old method, cryo-EM has proven to be a significant source for structural analysis owing to advances in technology for detecting ricocheting electrons and image-analysis software. These improvements catalyzed the process of resolution-revolution, yielding the sharpest protein structures ever. Some of the major advances in cryo-EM technology over the years are mentioned below.

1.5 Major advances in cryo-EM:

From the microscope point of view, the following are major improvements:

- The cryo-EM field has grown progressively in the image-recording media, initially from photographic film to CCD cameras to currently direct electron detectors. DDDs have several advantages over indirect electron detectors[185–188] as shown in Figure 1.11. First, DDDs do not use a scintillator which can introduce artifacts to limit the final resolution. Second, the thin layer sensors let the electrons pass through with lesser backscattering, resulting in a small PSF and hence higher resolution. Lastly, DDDs can capture images at high frame rates, which have multipurpose usage of motion correction, damage compensation, and various image processing techniques[189,190]. Note that scintillator-based CMOS detectors can also be used as well at high frames rates.

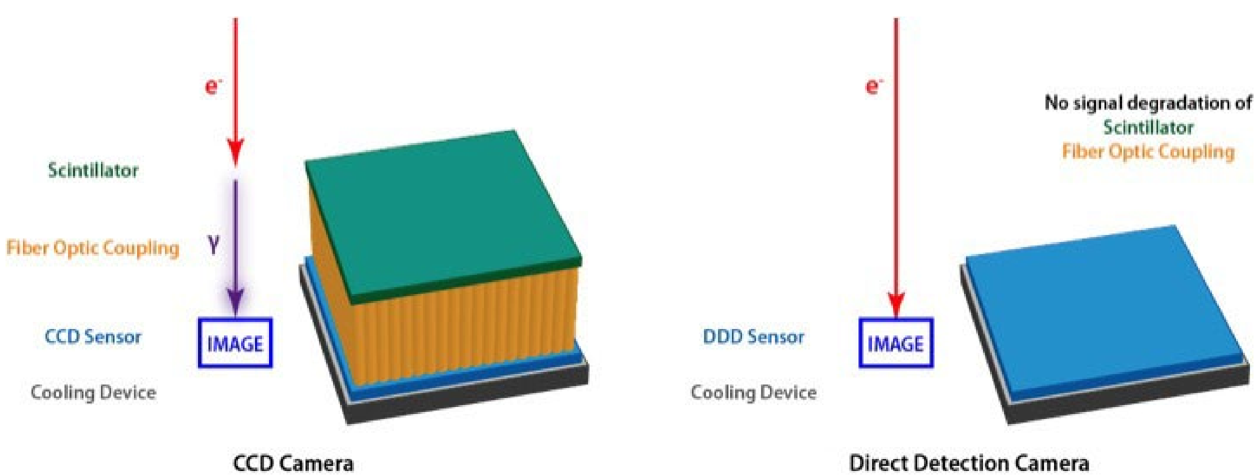


Figure 1.11: Working principle of DDD*. On the left half of the figure, the CCD (used in conventional digital cameras) has a scintillator that converts the electrons into photons for*

detection by the imaging sensor, whereas the right half illustrates how DDDs directly harness the electrons. (Reproduced with permission from Bammes BE et al, 2013 from[191]).

- Introducing multi specimen autoloaders in cryo-EM enabled high-throughput, giving the space to load up to 12 grids[192].
- Upgrading the optics by achieving parallel illumination over a wide range of conditions through the addition of a third condenser lens (ThermoFisher Scientific Titan). As the electron beam tilts, it can facilitate spatial coherency.
- High coherence advancement in electron guns.
    - Schottky-type electron gun: Recognized for its high-stability of electron current. Using a lower temperature and stronger electric field conditions than a thermionic emitter causes a decrease in the potential barrier to emit electrons, resulting in superb coherence by the Schottky field emitter[193].
    - Cold field-emission electron gun: Here, the tungsten tip is kept at room temperature and in a strong electric field, with electrons emitted by tunneling the potential barrier (∼4.5 eV). The emitted electron beam has a high coherence[193].
- In the microscope column, specimen stage stability has been improved as now the column can completely contain the specimen cartridge in concert with removal of the side entry holder.
- ***Data collection and automation:*** Automation enables high-throughput structure determination, with modern microscopes able to collect hundreds of high-resolution micrographs per hour. Various automated software examples include SerialEM [194], Leginon[195], Gatan Latitude[196], JEOL JADAS[197], ThermoFisher Scientific EPU/Tomography4[198], and UCSF-Image4[199]. Those produce larger datasets, up to 10 TB[37]

of movie data per sample. This large amount of data has traditionally been processed using large computer clusters of many CPU cores and stored on special high speed storage devices such as on SANs, NAS or local RAID. Increasingly, GPUs are being used to accelerate highly parallel numerically intensive computations, which eliminate the need for centrally maintained data centres, and also save energy. Modern gaming computers with consumer-grade graphics cards can yield an equivalent in performance to 100–1000 CPUs, and fit in a single chassis[37]. GPUs are now generally used in motion correction for movie frames[84,200] and often for large portions of the image processing workflow[201] (e.g. Relion[85]).

1.6 Thesis challenges

Cryo-EM is a structure determination technique that is useful for a wide range of macromolecules. The main goal of this technique lies in achieving a high-resolution reconstruction while discovering the conformational changes of dynamic macromolecules[202,203]. These goals are currently possible due to the above-mentioned major improvements in section 1.5. However, the presence of heterogeneity can hinder these goals. Such heterogeneity results from relevant conformational and compositional changes of macromolecules during their functional cycles. This structural variability demands advanced methodologies. The aim of the research in this thesis explores two significant computational techniques for improving the interpretability and analysis of cryo-EM data.

1.6.1 B-factor and map occupancy

The structure determination by cryo-EM depends on the quality of reconstructed EM density maps to generate an accurate atomic model. Hence, the reliability of resulting atomic models greatly depends on methodological advancements that can enhance the map quality. Sharpening methods in the post-processing step of the cryo-EM workflow helps to achieve such improvements, which are commonly used in X-ray crystallography[204,205].

In cryo-EM, B-factor correction is the favourable method that involves structure factor modification based on the Guinier plot[162]. This method overcomes the contrast loss in high-resolution maps by boosting the amplitudes of structure factors in a resolution range, shown by 'B-factor,' the slope of the amplitude falloff that will be boosted.

Most of the modern computational methods used by single-particle cryo-EM for map sharpening and atomic modelling are based on Wilson statistics[162,206], which describe the power spectrum of proteins at a high frequency. Wilson statistics are based on the assumption that all of the proteins show a similar power spectrum at high frequency regardless of their shape-specific atomic positions. This phenomenon helps to correct the Fourier coefficient magnitudes for the reconstructed map to match with the theoretical prediction. In this case, an exponentially growing filter, whose parameter is estimationally derived from a Guinier plot, is applied to the reconstructed map for increasing medium and high frequencies such that the sharpened map has a flat power spectrum which is also consistent with Wilson statistics[207]. Boosting the medium and high-frequencies or called as B-factor sharpening results in increasing the contrast of the reconstructed map and thus, helps to model the atomic structure.

Historically, the B-factor, temperature value or Debye–Waller factor in X-ray crystallography quantifies the degree to which the electron density is spread out locally, thus, the uncertainty in the position for each atom and the positions where errors may exist in the model building[208].

As introduced above and in analogy to X-ray crystallography[209], the B-factor of a cryo-EM 3D reconstruction represents the signal fall-off inside a defined resolution range and is experimentally calculated from the slope of the Guinier plot (logarithm of the structure factor amplitudes of a reconstruction versus the square of the spatial frequency).

Different sharpening algorithms have been introduced over the years, which can be categorized into two classes: global and local. All these algorithms apply the above-mentioned basic amplitude correction method. As such, global sharpening methods measure a single B-factor value for the whole cryo-EM map. RELION post-processing[85] and AutoSharpen in the Phenix Package[210] both fall into this category, working directly on the Guinier plot. Additionally, the AutoSharpen method endeavours to find the B-factor values that increase both the connectivity and the isosurface area of the cryo-EM map. However, by assuming single B-factor-values, these global sharpening algorithms neglect the important point that a cryo-EM map can also harbour different local resolutions. For such cases, local sharpening methods come to the rescue. LocScale is one of the examples that works by comparing the radial average of the structure factors inside a moving window in both the experimental map and the map calculated from the corresponding atomic model. Subsequently, it locally rescales the map amplitudes in Fourier space according to the atomic model. However, in this case, starting the atomic model is an important requirement that can reduce its usage. Another local sharpening, named LocalDeblur, requires a local resolution estimate as an input, which acts as a frequency cutoff to generate a lowpass filter cryo-EM map. LocalDeblur determines the map's local density values by convolution between the lowpass filter map and the actual cryo-EM map. An obvious shortcoming of this method is the requisite input estimation of a local resolution, which can only be obtained with expert experience and can easily go wrong. Therefore, a local sharpening algorithm that can overcome this shortcoming is necessary in a cryo-EM workflow. Another essential map post-processing step is determining the local map occupancy values, which describe the presence of an atom at its mean position with a range of 0.0 to 1.0, thereby enabling construction of an atomic model. Regrettably, though, cryo-EM lacks any method to calculate local map occupancies.

Chapter 2 surveys the methods to solve the above-said issues by using 3D SPT[1,211]. Dr. Javier Vargas, who is one of the authors of the manuscript from chapter 2, has been working with SPT to extract the modulating phase from interferometry[212–214], which we have used (in chapter 2[1]) to obtain the modulation or amplitude maps in case of cryo-EM map with non-homogenous distribution of resolution. Broadly speaking, the SPT helps to factorize a band-pass signal (1D, 2D, 3D or ND) into two components: the amplitude term and the phase term as: $\boldsymbol{V_{\omega}}(\mathbf{r}) = \boldsymbol{m_{\omega}}(\mathbf{r}) \cdot \mathbf{cos}\,(\boldsymbol{\varphi_{\omega}}(\mathbf{r}))$, where $\boldsymbol{V_{\omega}}(\mathbf{r})$ is the input map, $\boldsymbol{m_{\omega}}(\mathbf{r})$ is the amplitude map and $\mathbf{cos}\,(\boldsymbol{\varphi_{\omega}}(\mathbf{r}))$ corresponds to the cosine of the phase map. Note that the amplitude map refers to the amount of the signal or energy, while the phase map refers to its shape.

SPT has been used before in cryo-EM, as well, such as for the particle picking step in image analysis[215]. In this case, SPT is employed to facilitate morphology descriptors, which in turn identify the correctly picked particles from the incorrect ones based on their shape. SPT is also used in the case of CTF estimation[216], where CTF is considered as a fringe pattern, and its 2D phase has the details about the microscopic aberrations in the form of a sine function or so-called CTF signal. When applied to the CTF, SPT converts the CTF signal (or sine of the phase map) to the cosine, which represents the CTF quadrature signal. With these two maps (CTF and CTF quadrature signal), it is possible to obtain the phase map. The ultimately recovered phase map, called the absolute phase, provides information about microscope aberrations.

SPT is also used by several approaches[182,217] for local resolution determination, which works by calculating the local amplitude map of the cryo-EM map at each spatial frequency. For example, in the *MonoRes* approach[182], a Riesz transform is implemented[218–220] which is similar to the SPT. Here, at a certain given spatial frequency, the amplitude map is extracted from the input map, to compare it with amplitude distribution of noise at that certain resolution. A test is performed to check if the observed amplitude signal is higher than the noise signal at that specific resolution and location.

Essentially, the Riesz transform is a generalization of a Hilbert transform for a 1D signal[220]. In relation to the Fourier transform, a Hilbert transform can be explained as below. Mathematically, Fourier transform can be expressed as the 1D function f(t) which decomposes as a combination of waves with different frequencies, ω. Here, if a Hilbert transform is applied, it changes +π/2 to the negative frequencies and -π/2 to the positive frequencies, when applied to the Fourier transform. For example, for the sinusoidal function $f(t) = \cos(\omega \cdot t)$, if Hilbert transform is applied (represents shift of ±π/2 for negative and positive frequencies, respectively), the final output will be $H[f(t)] = \sin(\omega \cdot t)$. Therefore, in summary, here the Hilbert transform will convert the sines into negative cosines and the cosines into sines[182].

SPT is a generalization of the Hilbert transform for a 2D signal, and in chapter 2, methods representing 3D generalization of the 2D SPT will be discussed[1].

1.6.2 Observations of molecular dynamics

In cryo-EM, particle classification tools were the first studies to calculate free energies, which used Brownian machines, the ribosomes[221]. While in case of back-translocation tRNA studies[222], free-energy landscape is derived using the particles classified for each sub-state (n_i, where n represents the particle population of state *i*) and the Boltzmann law has been used to calculate the free-energy difference with respect to a reference state (ΔG; with particle population n_o). The Boltzmann factor can be shown as:

$$n_i/n_o=\exp(-\beta\Delta G), \text{ where } \beta=1/(K_BT),$$

where K_B is the Boltzmann constant and T is the temperature. The equation is taken from *Giraldo-Barreto J et al, 2021*[223].

In other words, the free-energy landscape in cryo-EM reflects the various conformations of the molecular system and their corresponding energy levels, called Gibbs-free energy. Nowadays, the free-energy landscape is commonly used to describe the aging of the protein folding mechanism[224,225], such as protein unfolding (denaturation) and protein folding to its native state mechanism[226]. In a free-energy landscape, many thermodynamical configurations present a number of local minima separated by barriers. The progression of the system's trajectory can be visualized in the form of local energy minima and saddle points (transition states), which are decided by the particle population per state[227], as shown in figure 1.12.

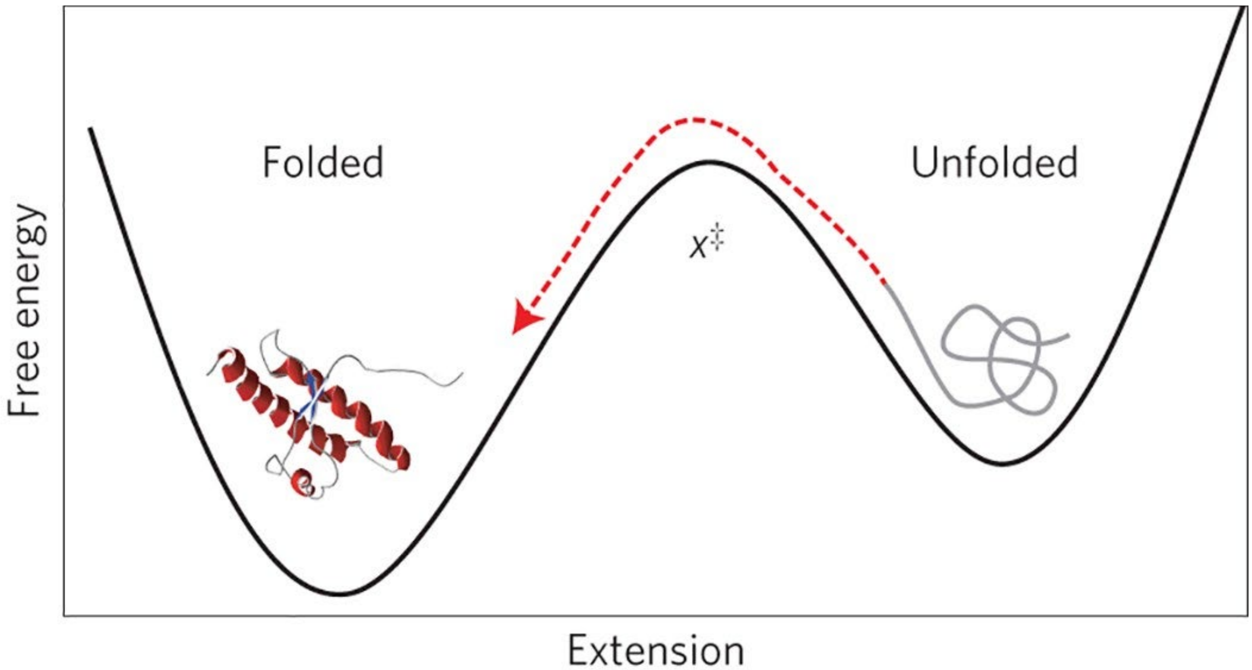


Figure 1.12: Trajectory of the two-state folding is shown in the form of an energy profile. *It contains two wells separated by a barrier, where the wells represent the unfolded and folded states of the system, while a transitional conformation is shown by the barrier (red). (Reproduced with permission from Neupane K et al. 2016 from[227]).*

1.6.3 Analysing a large conformational data for dynamic macromolecule

When interacting with biomolecules and ligands or in spontaneous fluctuation due to biological functions, macromolecules usually display different conformational motions. The 3D classification step of SPA workflow (recognized in section 1.4.4.7) helps to discover these various conformations of macromolecules. However, existing 3D classification methods[91,144,155,228] heavily depend on user expertise and experience, being susceptible to the so-called "attractor problem" that affects the output number of 3D classes. Additionally, the analysis of these conformations is essential to understand the molecular mechanism of macromolecule function, which can be explained in the form of energy landscape[229,230]. The challenge is to obtain a free-energy landscape that can provide quantitative information about the macromolecule internal energy at all possible conformations and can evaluate the likelihood of potential conformational changes as a function of the available thermal energy. Therefore, the development of an automatic 3D classification method with least user expertise demand is required, which can produce a large set of 3D classes, which can be further analysed in the form of the trajectory on energy landscape. This method is integrated on Scipion[231].

1.6.4 Scipion

Scipion[231] is open-source software that provides a platform for various 3DEM data analysis tools such as SIMPLE[232], EMAN[233], SPIDER[234], and RELION[85], among others. It was developed at Biocomputing Unit at the CNB in Spain. Using this integrated software has eliminated various limitations, including lack of standardization in the format of the output/input files for a specific program, and the burden of keeping tabs of each package's workflows to get an appropriate result. Scipion helps to simplify the interaction between different programs. Scipion also traces the workflow performed to reconstruct the 3D structure of a macromolecule and saves it in the form of a 'log' file. This function helps the user to find the possible error in a sorted way as well as examine the result properly. Figure 1.13 shows the various steps of cryo-EM 3D reconstruction workflow with their specific tasks in Scipion.

Scipion contains two different kinds of programs, as follows:

- Protocols: Written in python, protocols are used to manage and execute the algorithms of different packages as well as manage the format of various inputs and outputs.
- Libraries: written in C++, algorithms of libraries perform major functions with protocols in reconstruction procedure.

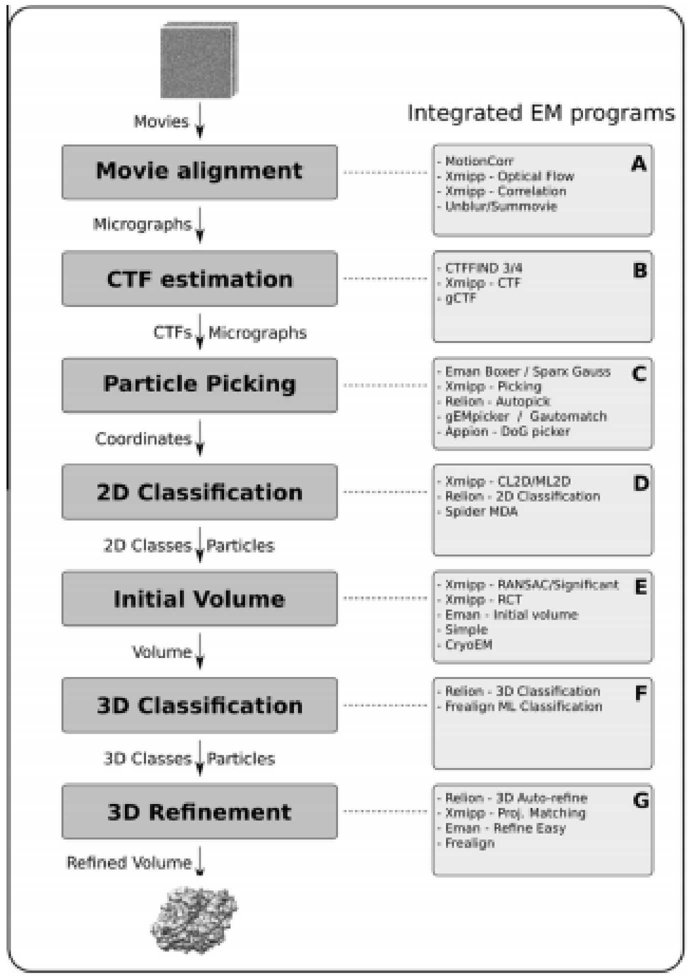


Figure 1.13: Workflow executed to determine the 3D structure of a macromolecule in Scipion. *(Reproduced with permission from JM, de la R-T et al, 2016 from[231]).*

1.7 Thesis objectives

To summarize, the objectives of this thesis are:

- *Presenting a novel method to measure the local B-factors and electron density occupancy maps while enhancing the high-resolution feature of cryo-EM maps.*

  This method is the best fit for cryo-EM maps with non-homogeneous distribution of SNR. Application of global sharpening approaches on such maps can cause over-sharpening (noise and broken densities) as well as under-sharpening (reduced contrast at high resolutions).

- *To identify the presence of various conformations, present in cryo-EM data.*

  This method allows uncovering the various conformations of dynamic macromolecules through cryo-EM 2D and 3D classification algorithms, which is importantly made by considering the factor of automation (or no user's input) in the form of number of final classes or iterations.

- *Presenting the large conformational data output in the form of a trajectory, from the aforementioned automatic 2D and 3D classification method.*

  This technique uncovers the conformational trajectory for a dynamic macromolecule. It is shown in the form of free-energy landscape which employs machine learning tools and Boltzmann's distribution law that connects the energy of the macromolecule's dynamics with the temperature and population of particle images.

CHAPTER 2

LOCAL COMPUTATIONAL METHODS TO IMPROVE THE INTERPRETABILITY AND ANALYSIS OF CRYO-EM MAPS

2.1 Abstract

Cryo-EM maps usually show heterogeneous distributions of B-factors and electron density occupancies and are typically B-factor sharpened to improve their contrast and interpretability at high-resolutions. However, 'over-sharpening' due to the application of a single global B-factor can distort processed maps causing connected densities to appear broken and disconnected. This issue limits the interpretability of cryo-EM maps, i.e. ab initio modelling. In this work, we propose 1) approaches to enhance high-resolution features of cryo-EM maps, while preventing map distortions and 2) methods to obtain local Bfactors and electron density occupancy maps. These algorithms have as common link the use of the spiral phase transformation and are called LocSpiral, LocBSharpen, LocBFactor and LocOccupancy. Our results, which include improved maps of recent SARS-CoV-2 structures, show that our methods can improve the interpretability and analysis of obtained reconstructions.

2.2 Introduction

Cryo-EM has become a mainstream technique for structure determination of macromolecular complexes at close-to-atomic resolution and ultimately for building an atomic model[1,2]. With its unique ability to reconstruct multiple conformations and compositions of the macromolecular complexes, cryo-EM allows the understanding of the structural and assembly dynamics of macromolecular complexes in their native conditions[3–5]. However, the presence of heterogeneity in cryo-EM maps leads to high variability in resolution within different regions of the same map.

This directs to challenges and errors in the process of building an atomic model from a cryo-EM reconstruction. Additionally, current reconstructions from cryo-EM do not provide essential information to build accurate ab initio atomic models as atomic Debye–Waller factors (B-factors) or atomic occupancies, while their counterparts from X-ray crystallography do by analysing the attenuation of scattered intensity represented at Bragg peaks. Cryo-EM structures exhibit loss of contrast at high-resolution coming from many different sources, including molecular motions, heterogeneity and/or signal damping by the transfer function of the electron microscope (CTF). Interpretation of high-resolution features in cryo-EM maps is essential to understanding the biological functions of macromolecules. Thus, approaches to compensate for this contrast loss and improve map visibility at high-resolution are crucial. This process is usually referred to as 'sharpening' and is typically performed by imposing a uniform B-factor to the cryo-EM map that boosts the map signal amplitudes within a defined resolution range. When the map is sharpened with increasing positive B-factors, the clarity and map details initially improve, but eventually, the map becomes worse as the connectivity is lost, and the map densities appear broken and noisy. In the global sharpening approach[6–8], the B-factor is automatically computed by determining the line that best fits the decay of the spherically averaged noise-weighted amplitude structure factors, within a resolution range given by [15–10 Å, Rmax], with Rmax the maximum resolution in the map given by the Fourier Shell Correlation (FSC). More recently, the AutoSharpen method within Phenix[9] calculates a single B-factor that maximises both map connectivity and details of the resulting sharpened map. AutoSharpen automatically chooses the B-factor that leads to the highest level of detail in the map, while maintaining connectivity. This combination is optimised by maximising the surface area of the contours in the sharpened map. The approaches presented above are global, so the same signal amplitude scaling is applied to map regions that may exhibit very different signal to noise ratios (SNRs) at medium/high-resolutions. Thus, cryo-EM maps showing inhomogeneous SNRs (and resolutions) can result in sharpened maps that show both oversharpened and under-sharpened regions. The former may be strongly affected by noise and broken densities, while the latter may present reduced contrast at high-resolutions. Both cases make it difficult or even impossible to interpret the biological relevance of these regions or even the whole map[10]. Thus, local sharpening methods have been proposed to overcome these limitations[11,12]. LocScale approach11 compares radial averages of structure factor amplitudes inside moving windows between the experimental and the atomic density maps. After, the method modifies locally the map amplitudes of the experimental map in Fourier space to rescale them accordingly to those of the atomic map. This approach requires as input a complete atomic model (without major gaps) fitted to the cryo-EM map to be sharpened, which is not always available. In addition, the size of the moving window should be provided and depending on the quality of the map to be sharpened, this process may lead to overfitting. More recently, the LocalDeblur method[12] proposed an approach for map local sharpening using as input an estimation of the local resolution. The method assumes that the map local density values have been obtained by the convolution between a local isotropic low-pass filter and the actual map. This local low-pass filter is assumed Gaussian-shaped so that the frequency cutoff is given by the local resolution estimation. In X-ray crystallography, the B-factor (also called temperature value or Debye–Waller factor) describes the degree to which the electron density is spread out, indicating the true static or dynamic mobility of an atom and/or the positions where errors may exist in the model building. The B-factor is given by $B_i = 8\pi^2 u_i^2$, where u_i^2 is the mean square displacement for atom i. These atomic B-factors can be experimentally measured in X-ray crystallography, introduced as an amendment factor of the structure factor calculations since the scattering effect of X-ray is reduced on the oscillating atoms compared to the atoms at rest[13]. B-factors can be further refined by model building packages, i.e. Phenix[14] or Refmac[15] to improve the quality and accuracy of atomics models. Although B-factors are essential to 'sharpen' cryo-EM maps at high-resolution, they also provide key information to analyse cryo-EM reconstructions. Effective B-factors are used to model the combined effects of issues such as molecular drifting due to charging effects, macromolecular flexibility or possible errors in the reconstruction workflow that lead to a signal fall-off[6,16,17]. However, cryo-EM maps are usually analysed with a single B-factor, even though maps may largely differ in different regions. Thus, methods to determine local B-factors are much needed to accurately analyse cryo-EM maps and improve the quality of fitted atomic models. Another local parameter usually provided by X-ray crystallography in contrast with cryoEM are atomic occupancies (or Q-values). The occupancy estimates the presence of an atom at its mean position and it ranges between 0.0 to 1.0. Note that these parameters can be also refined by model building packages if the electron density map is of sufficient resolution. To our knowledge, currently, there is not any available method to estimate local occupancies from cryo-EM maps, even though this information (in addition to local B-factors) is essential to building accurate atomic models. For example, in ref. [18] authors found that 31% of all models examined in this analysis possess unrealistic occupancies or/and B-factor values, such as all being set to zero or other unlikely values. They also reported that 40% of models analysed show cross-correlations between cryo-EM maps and respective models below 0.5, and they indicated as a possible hypothesis an incomplete optimisation of the model parameters (coordinates, occupancies and Bfactors). In this work, we propose semi-automated methods to enhance high-resolution map features to improve their visibility and interpretability. More importantly, these approaches do not require input parameters as fitted atomic models or local resolution maps, which reduces the possibility of overfitting. In particular, our proposed local map enhancement approach (LocSpiral) is robust to maps affected by inhomogeneous local resolutions/SNRs, thus the method strongly improves the interpretability of these maps. Secondly, we also propose approaches to determine local B-factors and density occupancy maps to improve the analysis of cryo-EM reconstructions. The link between the different proposed approaches is the use of the spiral phase transform to estimate a modulation or amplitude map of the cryo-EM reconstruction at different resolutions.

2.3 Results

In this section, we first provide a brief and comprehensive description of the approaches developed in this work. A deeper and more technical explanation of these methods is given in the 'Methods' section at the end of the manuscript. Then, we present results obtained by our approaches in a variety of situations. We tested our proposed methods with five different samples ranging from near-atomic single-particle reconstructions (∼1.54 Å) to maps with more modest resolutions (∼6.5 Å). In all cases, we compared our results with the ones provided by the Relion postprocessing approach[7,19].

2.3.1 Overview of the proposed methods

The input parameters of the different methods (LocSpiral, LocBSharpen, LocBFactor and LocOccupancy) is the unfiltered map to process, a resolution range given by [Rmin, Rmax] and, in some cases, a tight solvent mask. The different algorithms start by filtering the input map to a given resolution 1/ω within the resolution range. Then, the 3D spiral phase transform is calculated to factorise in real space the filtered map into amplitude and a phase map as

$$V_{\omega}(\mathbf{r}) = m_{\omega}(\mathbf{r})\cos\left(\varphi_{\omega}(\mathbf{r})\right) \tag{1}$$

The amplitude map $m_\omega(\mathbf{r})$ is related to the 'strength' of the local map signal at resolution $1/\omega$, while the phase map refers to its shape and it is limited to the $[-1, +1]$ range. The different methods proposed here are based on the analysis of the amplitude maps. In some cases, the approaches compute new amplitude maps ($\breve{m}_\omega(\mathbf{r})$), which are used to determine a sharpened map (LocSpiral, LocBSharpen) as

$$\breve{V}(\mathbf{r}) = \sum_\omega \breve{V}_\omega(\mathbf{r}) = \sum_{\boldsymbol{\omega}} \boldsymbol{C}_{\mathbf{ref},\boldsymbol{\omega}}(\mathbf{r})\breve{\boldsymbol{m}}_{\boldsymbol{\omega}}(\mathbf{r})\mathbf{cos}\,(\boldsymbol{\varphi}_{\boldsymbol{\omega}}(\mathbf{r})) \tag{2}$$

with $C_{ref,\omega}(\mathbf{r})$ a SNR weighting parameter (please see 'Methods' section). In other cases, the amplitude maps are further analysed to provide local B-factor maps (LocBFactor) or a local occupancy map estimation (LocOccupancy).

In LocSpiral at every resolution inside the resolution range, the amplitude map is compared locally with a noise threshold value computed from the 90–95% quantile of the empirical noise/background distribution at this resolution. This empirical distribution is generated collecting the amplitude values at resolution $1/\omega$ for all voxels outside the tight solvent mask. The computed noise threshold is then used to obtain a new normalised and filtered amplitude map, $\breve{m}_\omega(\mathbf{r})$, which is used to reconstruct the sharpened map as shown in Eq. (2).

In LocBSharpen the amplitude map at a resolution $1/\omega_0\left(m_{\omega_0}(\mathbf{r})\right)$ is stored. The resolution $1/\omega_0$ is provided by the user and is typically 15–10 Å. In the process of building the sharpened map, the new amplitude map $\breve{m}_\omega(\mathbf{r})$ at any resolution equal or higher than $1/\omega_0$ is equal to $m_{\omega_0}(\mathbf{r})$, while for the rest of resolutions inside the resolution range, $\breve{m}_\omega(\mathbf{r})$ is equal to $m_\omega(\mathbf{r})$

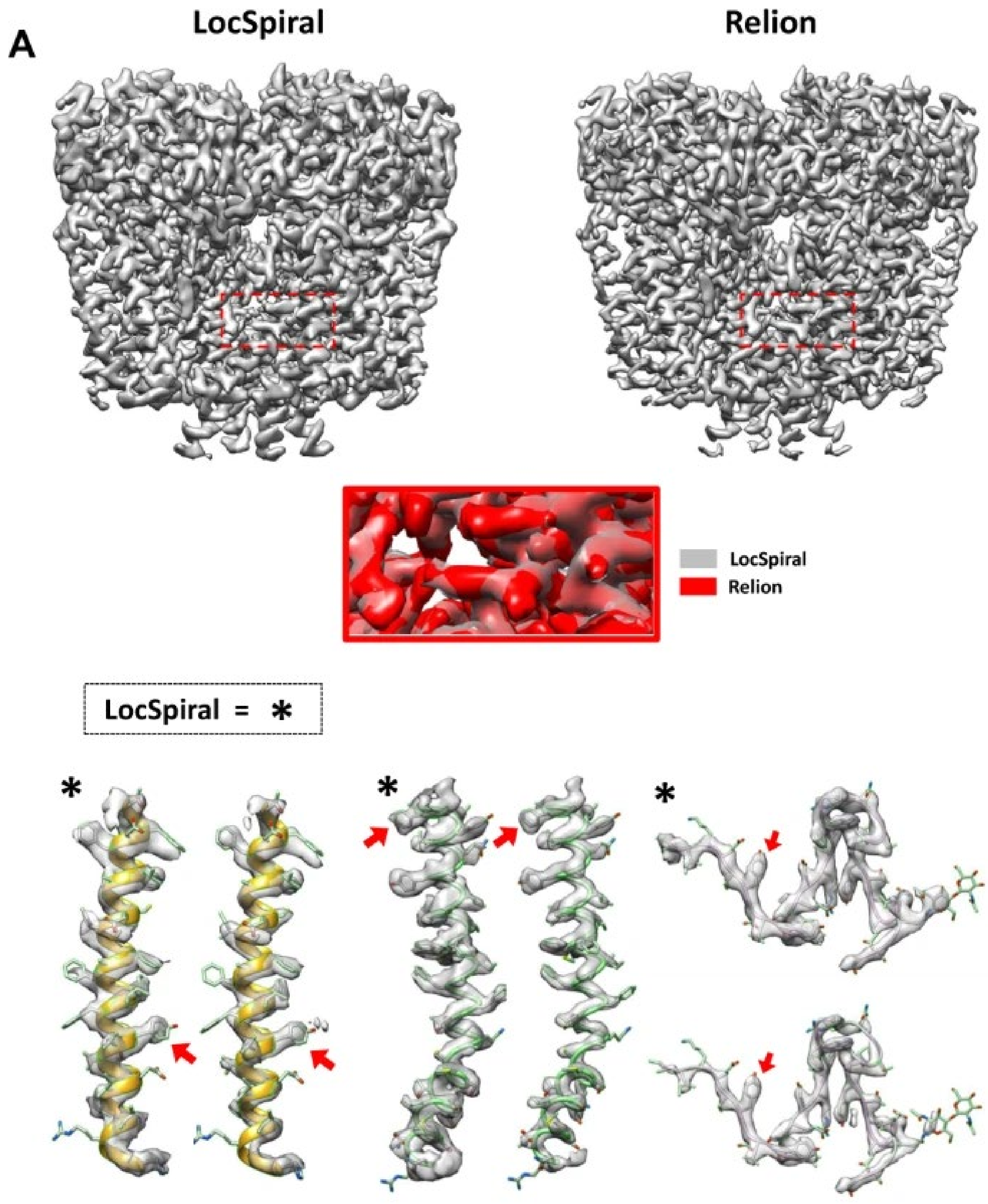
In LocBFactor the amplitude maps at different resolutions inside the defined resolution range are used to estimate map local B-factors. A typical resolution range is of $[15, R_{max}]$ Å, being R_{max} the global map resolution. To compute the local B-factors, the method obtains at every voxel $\mathbf{r}$ the linear fitting between $\log\left(C_{ref,\omega}(\mathbf{r})m_\omega(\mathbf{r})\right)$ and ω^2 within the resolution rage. The method provides as output the B-map (local B-factor map) and the A-map (local values of the logarithm of structure factor amplitudes at 15 Å).

In LocOccupancy, the local occupancy map is estimated comparing the amplitude map with a macromolecule density threshold for every resolution inside the defined resolution range. The macromolecule density threshold at a given resolution indicates the density value at which we are confident that the electron density occupancy is of 100% at this resolution. This threshold is obtained from the empirical macromolecule amplitude probability distribution (m_ω^M) at frequency ω. This amplitude probability distribution is calculated from density values at voxels that are included inside the solvent mask. From this distribution, the macromolecule density threshold may be calculated from the macromolecule amplitude value corresponding to the 25% quantile, given by $m_\omega^M(q = 25\%)$. Then, for every voxel and resolution within the resolution range, the amplitude map $m_\omega(\mathbf{r})$ is compared with $m_\omega^M(q = 25\%)$, providing a value between 0 and 1. Finally, the average value over all resolutions is computed and provided as an estimation of the map occupancy within the resolution range.

2.3.2 Polycystin-2 (PC2) TRP channel

First, we analysed a single-particle reconstruction of the polycystin-2 (PC2) TRP channel (EMDataBank: EMD-10418)[20]. In this case, we focussed on showing the capacity of LocSpiral approach, though, for the sake of consistency, we also show results of obtained B-factor and occupancy maps. The original publication reports a resolution of 2.96 Å with a final B-factor to be used for sharpening of −84.56 Å^2 (slope of Guinier plot fitting equal to −21.14 Å^2).

In Figure 2.1A, we show maps with high threshold values obtained by LocSpiral and by the postprocessing method of Relion 3[7,19]. The map densities are similar in the inner core of the protein as can be seen from the solid red rectangle in the figure, where we show a zoomed view of LocSpiral and Relion maps of the region indicated in the red rectangles over the maps. However, the map densities are quite different in the outer regions, where the Relion map shows thin and broken densities. In addition, we show comparisons of fitted densities with the corresponding atomic model (PDB ID: 6t9n) of two α-helices and one loop. The asterisks label results obtained by LocSpiral. The residues marked with a red arrow were used to adjust the threshold values between maps. These comparisons show that the map obtained by LocSpiral shows fewer fragmented and broken densities and better coverage of the atomic model, helping in the interpretation of the maps and in the process of building accurate atomic models. In Supplementary Figure 2.1, we show additional figures comparing LocSpiral and Relion postprocessing maps.


A
LocSpiral
Relion
LocSpiral
Relion
LocSpiral = *
*
*
*

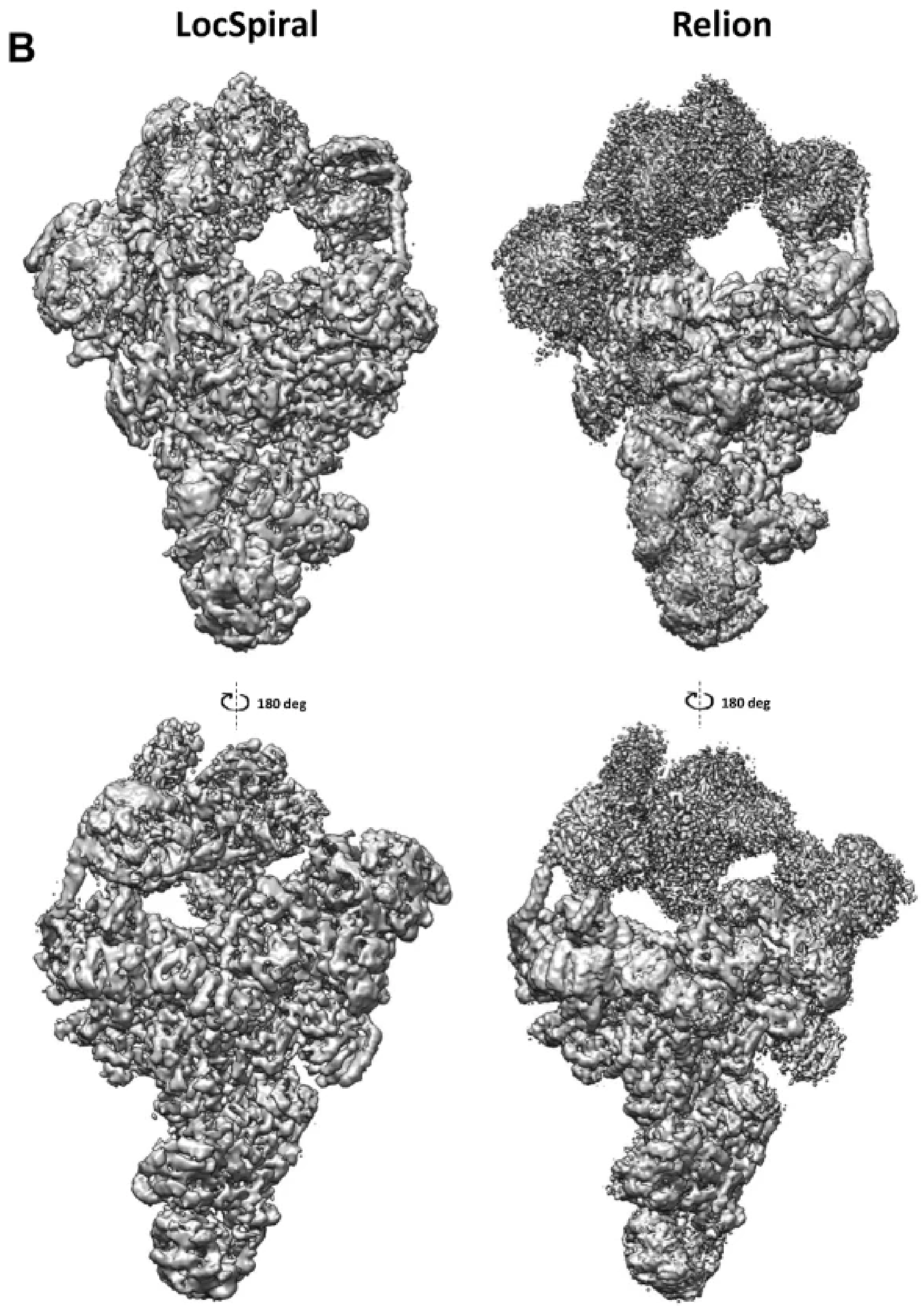


Figure 2.1: Capacity of LocSpiral to improve the interpretability of cryo-EM maps. A *Top: sharpened maps of the TRP channel obtained by LocSpiral (left) and Relion postprocessing (right) methods. The threshold values are adjusted to provide similar densities in the core inner part of the protein. The red square in the figure shows a zoomed view of the protein inner core where both maps (LocSpiral and Relion) are superimposed. Relion map appears in red colour, while LocSpiral is in grey. Bottom: Fitted map densities (LocSpiral and Relion) with the corresponding*

atomic model (PDB ID: 6t9n) of two α-helices and one loop. The asterisks mark results obtained by LocSpiral approach. The residues marked with a red arrow were used to adjust the threshold values between maps. ***B*** *Spliceosome maps at different orientations and similar threshold values obtained by LocSpiral and the postprocessing method of Relion.*

We also compared the performance of LocSpiral with other methods, including LocalDeblur, our proposed local *B*-factor correction method (LocBSharpen) and the global *B*-factor correction approach as implemented in Relion. The results are shown in Supplementary Note 1, Supplementary Table 2.1 and Supplementary Figure 2.2 where we also provide results obtained by LocBFactor and LocOccupancy methods.

2.3.3 Pre-catalytic spliceosome

Next, we processed the Saccharomyces cerevisiae pre-catalytic B complex spliceosomal single particles deposited in EMPIAR (EMPIAR 10180)[4,21]. This dataset exhibits a high degree of conformational heterogeneity, thus, it represents a perfect use case to test our proposed approaches. We used the approach described in ref.[22] to obtain a reconstruction at 4.28 Å resolution after Relion postprocessing[7,19]. In the 'Methods' section, we provide a detailed description of the image processing workflow used to obtain this reconstruction. The unfiltered map provided by Relion autorefine was used as input to LocSpiral, LocBFactor and LocOccupancy.

We first show results obtained by LocSpiral method for this highly heterogeneous case. In Figure 2.1B, we show maps at different orientations and similar threshold values obtained by LocSpiral and by the postprocessing method of Relion 3[7,19]. As before, the LocSpiral map shows fewer fragmented and broken densities, especially in the flexible part of the spliceosome reconstruction, and enhanced details in the central core portion improving the visibility of the reconstruction.

We then concentrate on showing the capacity of LocBFactor method. In Figure 2.2A, we show a central slice along the *Z* axis of this map with several points marked with coloured squares. These points show parts of the map that correspond to clear spliceosome densities (green and red), flexible and low-resolution spliceosomal regions (yellow and blue) and background (magenta). Figure 2.2B shows the corresponding Guinier plots at these locations. Solid lines represent measured values of the logarithm of SNR-weighted structure factor amplitudes, while dashed lines show fitted curves. This figure also provides the obtained *B*-factors for the different curves. The Guinier plots and *B*-factors are determined within a resolution range of 15 Å to the FSC resolution, given by 4.28 Å. As can be seen from Figure 2.2B, the red and green curves, which correspond to clear spliceosomal densities, present high amplitude values at 15 Å, while the yellow, blue and magenta curves show low amplitudes at 15 Å and a flat profile within the resolution range. In Figure 2.2B, we also show in the black curve, the Guinier plot of the noise/background amplitudes obtained from the 90–95% quantile of the empirical noise/background distribution for reference. The discontinuous black line indicates the linear fit of this noise Guinier plot. Comparing the yellow, blue, magenta, and black curves, it is clear that these plots are below our noise level and that the shape of these curves is similar to that of the noise curve. Thus, these *B*-factors describe mainly noise *B*-factors that show how the noise signal fall off inside the used resolution range and they should be filtered out from our *B*-factor map. Moreover, Figure 2.2C shows the spliceosome map coloured according to the occupancy map obtained by LocOccupancy using a resolution range of [30, 10] Å. From Figure 2.2C, we see that the flexible and moving parts of the spliceosome, like the ones indicated with the yellow and blue points in Figure 2.2A, show low occupancies (close to zero) within the used resolution range. Figure 2.2D renders the spliceosome map coloured with the obtained *B*-factor map to be used for sharpening (slope of the local Guinier plot multiplied by 4). In Figure 2.2D the noise *B*-factors (*B*-factors obtained from amplitudes below the noise level for the used resolution range) are filtered out and appear with black colour. Note that Guinier plots at regions with amplitudes below the noise level are dominated by the noise signal and describe the noise signal fall-off inside the used resolution range. The noise signal presents typically a flat spectrum, thus, artefactual close to zero *B*-factors, which are not in agreement with the concept of *B*-factor as a measure of position uncertainty or disorder. Figure 2.2E shows the corresponding local resolution map as obtained by Resmap[23] of the spliceosome reconstruction. As can be seen from this figure, the local resolution values of the flexible parts (helicase and SF3b domains) are lower than the others and within a range of [10, 15] Å. Consequently, the obtained amplitudes for these flexible parts within the resolution range of [15, 4.28] Å are dominated by the noise/background signal. The average inside a solvent mask of the signal *B*-factors (*B*-factors obtained from amplitude values above the noise level for the used resolution range) is -567.62 Å^2, while the value reported by Relion postprocessing is -158.08 Å^2. Note that Relion postprocessing does not filter out regions dominated by noise/background when computing the global *B*-factor. As mentioned before, regions dominated by the noise signal within the used resolution range present artefactual low *B*-factors. Consequently, this global *B*-factor may be overestimated. A more detailed description of this point is given in Supplementary Note 3: *B*-factor analysis of low and high-resolution maps. In Figure 2.2F, we show the local values of the logarithm of the structure factor's amplitudes at 15 Å (A map). As expected, this map shows low amplitudes at highly flexible and moving regions. We have recalculated *B*-factors using a new resolution range of [20, 10] Å. The results are shown in Figure 2.2G-I. As can be seen from these figures, now the flexible parts show unfiltered low signal *B*-factors and low amplitudes at 20 Å. However, it is important to note that at this resolution range, the *B*-factors are dominated by the molecular shape and solvent contrast and not by resolution limiting factors such as errors in the reconstruction procedure (as the presence of heterogeneity), radiation damage or imaging imperfections, for example[6].

Consequently, it is not recommended to use a resolution range of [20, 10] Å as obtained *B*-factors may not be used to evaluate map quality.

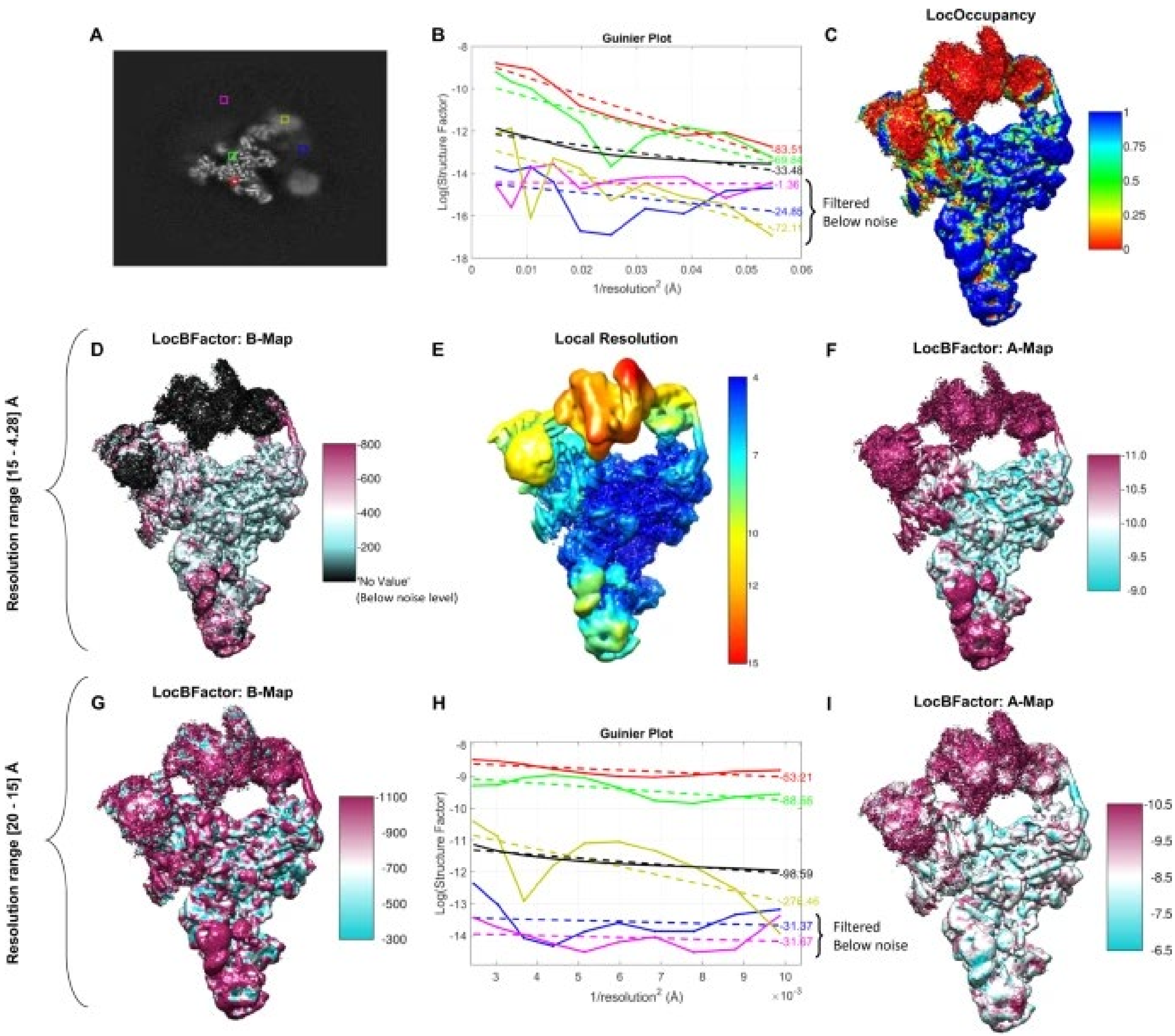


Figure 2.2: Results obtained by LocBFactor and LocOccupancy for the Saccharomyces cerevisiae pre-catalytic B complex spliceosome sample. *A Central slice along Z axis of the obtained unsharpened map using EMPIAR 10180 single particles. Coloured squares mark parts of the map corresponding to clear spliceosome densities (green and red), flexible and low-resolution spliceosomal regions (yellow and blue) and background (magenta). B Guinier plots at map points indicated in the coloured squares using a resolution range of [15–4.28] Å. Solid lines represent SNR-weighted values of the logarithm of structure factor amplitudes, while discontinued lines*

show the fitted lines. C Spliceosome map coloured with the obtained occupancy map by LocOccupancy. The occupancy ranges from 0 (red colour) to 1 (blue colour), indicating no macromolecular density and full occupancy, respectively. D Spliceosome map coloured with the B-factor map to be used for sharpening (slope of the local Guinier plot multiplied by 4) obtained by LocBFactor using a resolution range of [15–4.28] Å. The local B-factor values in the figure range approximately between −800 and −200 Å2. In this figure, noise B-factors (B-factors obtained from amplitudes below the noise level for the used resolution range) are filtered out and appear with black colour. E Local resolution map obtained by Resmap approach. The local resolution ranges between 4 (blue colour) and 15 Å (red colour). F Spliceosome map coloured with the obtained A map (local values of the logarithm of structure factor amplitudes at 15 Å). The values range between −11.0 (magenta colour) and −9.0 (cyan colour) approximately. G Spliceosome map coloured with the B-factor map (B-map) obtained by LocBFactor using a resolution range of [20–15] Å. The Bfactor values range between −1100 (magenta colour) and −300 (cyan colour) approximately. H Guinier plots at map points indicated in the coloured squares using a resolution range of [20–15] Å. Solid lines represent SNR-weighted values of the logarithm of structure factor amplitudes, while discontinued lines show the fitted lines. I Spliceosome map coloured with the obtained A map (local values of the logarithm of structure factor amplitudes at 20 Å). The values range between −10.5 (magenta colour) and −6.5 (cyan colour) approximately

2.3.4 Apoferritin

We have also applied these techniques to recently reported high-resolution cryo-EM reconstructions of mouse apoferritin: EMD-9865 and EMD-21024. The reported global resolution of these reconstructions is 1.54 and 1.75 Å for EMD-9865 and EMD-21024, respectively.

In Figure 2.3A, B, we show the results obtained by LocBFactor (B and A maps) and LocOccupancy methods (occupancy maps). The resolution range used to estimate the B and A maps was between 15 Å to the reported global resolution for both cases. The occupancy maps were calculated for these high-resolution maps between 5 Å to the global resolution. As can be seen from Figure 2.3B, EMD-9865 shows lower *B*-factors and higher local amplitudes than EMD-21024, indicating a better-quality reconstruction, however, the low values of both B maps indicate the high quality of these reconstructions. In both cases, the highest *B*-factors are in the outer regions of the protein. Moreover, local occupancies show similar maps for both cases, showing occupancies as low as approximately 0.5 at the outer part and indicating the presence of flexibility in these outer residues. Note that the obtained average and standard deviation of B-factors inside a solvent mask is of -56 and 7.20 Å^2 (EMD-9865) and -78 and 8.93 Å^2 (EMD-21024), respectively, which reflects the high quality of these reconstructions.

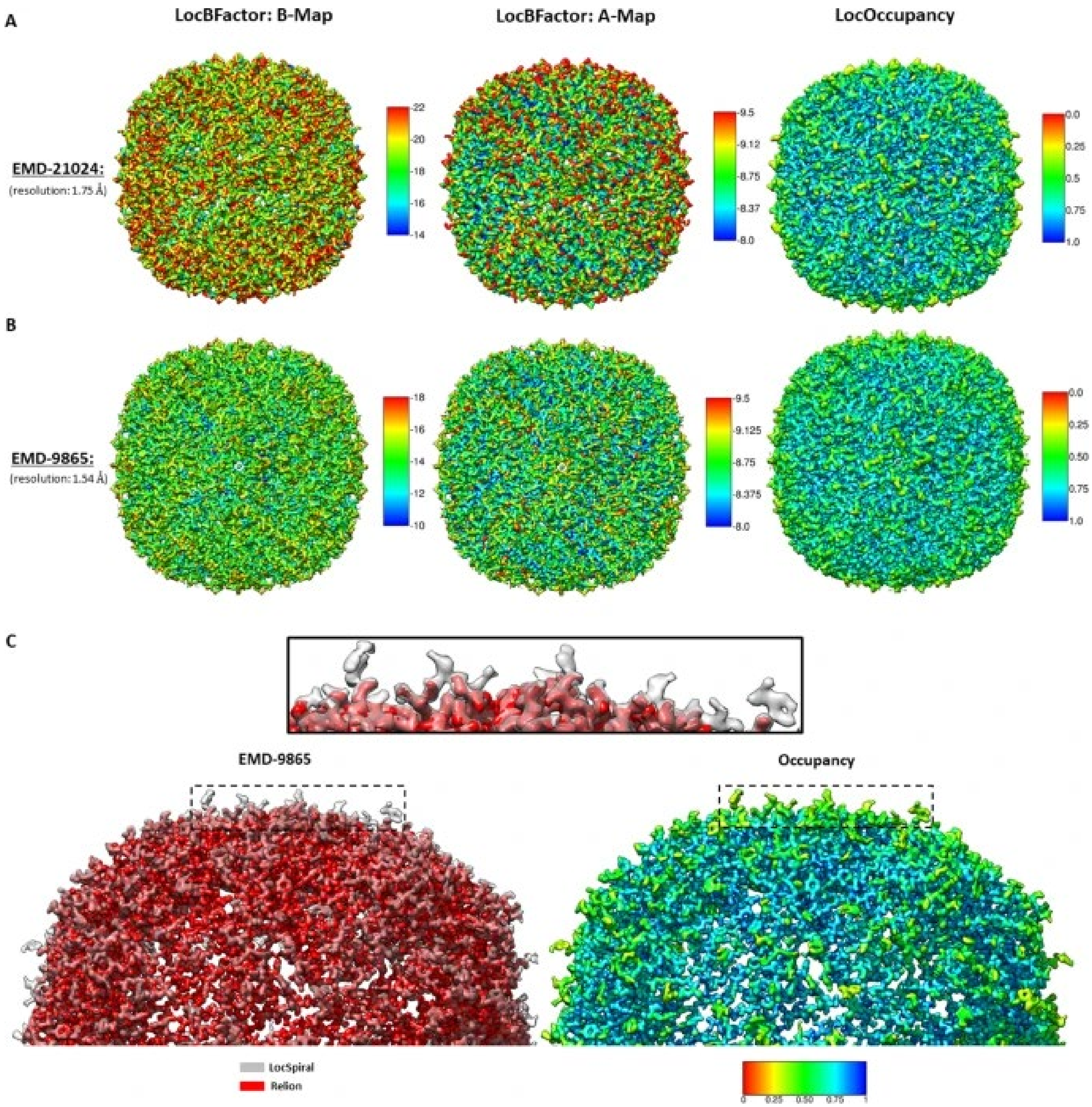


Figure 2.3: Results obtained by LocBFactor, LocOccupancy and LocSpiral for apoferritin sample. *Obtained B-maps (local B-factor map corresponding to the slope of the local Guinier plots), A-maps (local values of the logarithm of structure factor amplitudes at 15 Å) and occupancy*

maps by LocBFactor and LocOccupancy for EMD-21024 (A) and EMD-9865 (B). The B-factor ranges between [−22, −14] Å2 in A and [−18, −10] Å2 in B. The A-map ranges between [−9.5, −8.0] in A and B. The occupancy ranges between [0, 1] in A and B. In C, we show on the left side, superimposed sharpened maps obtained by LocSpiral (grey colour) and Relion (red colour) for EMD-9865. The black rectangle shows a zoomed view of the region indicated with the dashed rectangles. On the right, we show the respective occupancy map obtained by LocOccupancy at the same orientation that these sharpened maps. In this figure, 0 (red colour) indicates no density occupancy and 1 (blue colour) full occupancy.

In Figure 2.3C, we show sharpened maps obtained from EMD-9865 by LocSpiral and by the postprocessing method of Relion 3. The LocSpiral map is shown in grey colour, while the Relion map is rendered in red. The solid black rectangle shows a zoomed view of the outer region of the protein, which is indicated with the dashed black rectangles in the figure. Supplementary Figure 3 shows that the extra densities that appear in the LocSpiral map correspond to missing residues in EMD-9865. Additional, at the right of Figure 2.3C, we show the respective occupancy map obtained by LocOccupancy at the same orientation that these sharpened maps. As can be seen from Figure 2.3C, the LocSpiral map shows fewer fragmented and broken densities, especially in the parts of the map that shows low occupancies. We compute also EMRINGER and MolProbity scores[26] between these maps (EMD-9865 and LocSpiral) and the atomic model (PDB 6v21) after refining the structure against corresponding maps by Phenix real_space_refine approach[27] using 5 refining iterations. The results obtained are shown in Supplementary Table S1.

2.3.5 Immature prokaryote ribosomes

We processed immature ribosomal maps of the bacterial large subunit[3]. These maps were obtained after depletion of bL17 ribosomal protein and are publicly available from the Electron Microscopy Data Bank (EMDB) (EMD-8440, EMD-8441, EMD-8445, EMD-8450, EMD-8434)[28]. In this case, we focussed on showing the capacity of LocOccupancy to interpret and analyse reconstructions showing a high degree of compositional heterogeneity.

Figure 2.4 shows the obtained results. The first row shows the different maps to be processed as deposited in the EMDB. Next, we show the obtained occupancy maps by LocOccupancy, where the mature 50S ribosome (EMD-8434) is coloured according to corresponding occupancy maps. The resolution range used was [30, 10] Å. These figures clearly show regions that are lacking in the different immature maps with respect to the mature map. Thus, occupancy maps were used to create binary masks to segment the mature 50S ribosome map, extracting after the densities that are missing in the respective immature maps. These densities are shown in the third column of Figure 2.4 with different colours (yellow, red, indigo and green). The obtained occupancy maps also allow us to define a 'maturity level' index. This index is calculated by comparing the number of voxels activated in the solvent mask of the mature 50S reconstruction with the ones in the occupancy masks (see methods section for a more detailed description). As can be seen from Figure 2.4, the larger the unfolded regions in the immature maps are, the smaller the maturity level is. This maturity level index allows us to quantitatively sort the different immature maps in a spectrum according to their maturity.

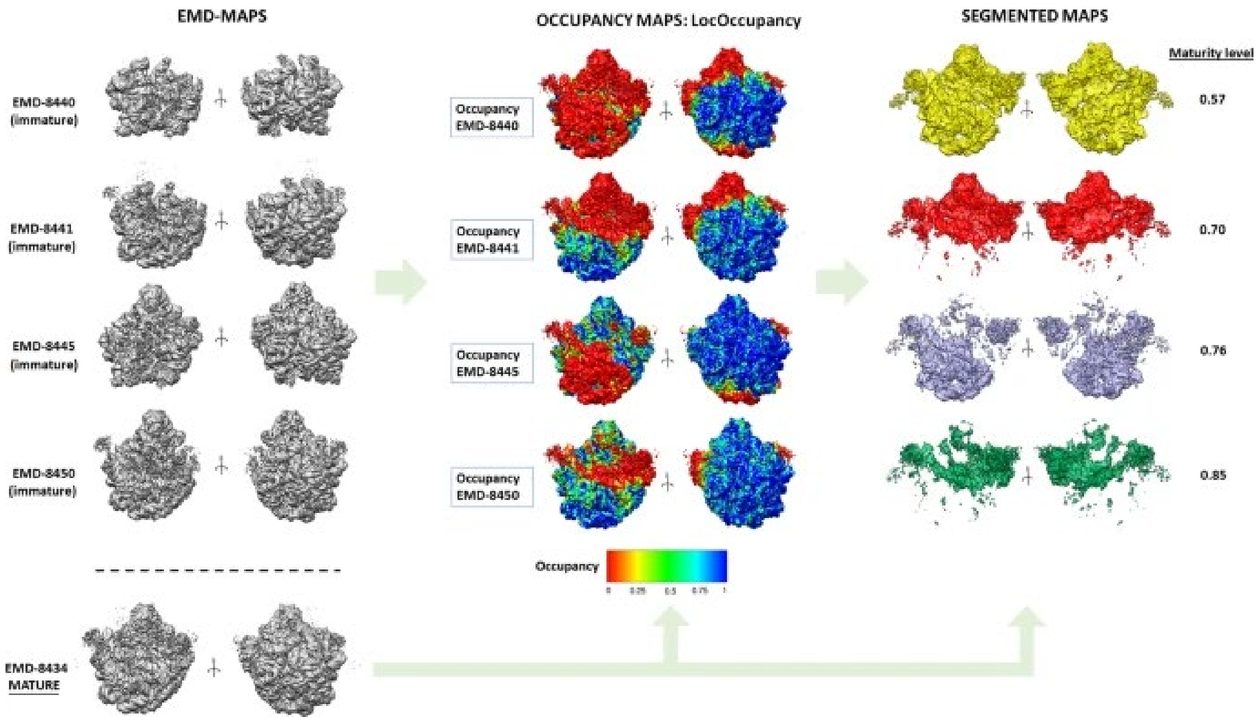


Figure 2.4: Results obtained by LocOccupancy for immature 50S ribosomes. *First column: immature maps at different orientations as deposited in EMDB. Second column: obtained occupancy maps by LocOccupancy, where the mature 50S ribosome (EMDB-8434) is coloured with corresponding occupancy maps obtained from the immature ribosomes. The occupancy ranges between [0, 1]. Third column: Segmented maps showing the densities that are missing in the different immature maps when compared to the mature 50S reconstruction and obtained maturity levels. The different colours (yellow, red, purple, green) label the different corresponding segmented regions for each case*

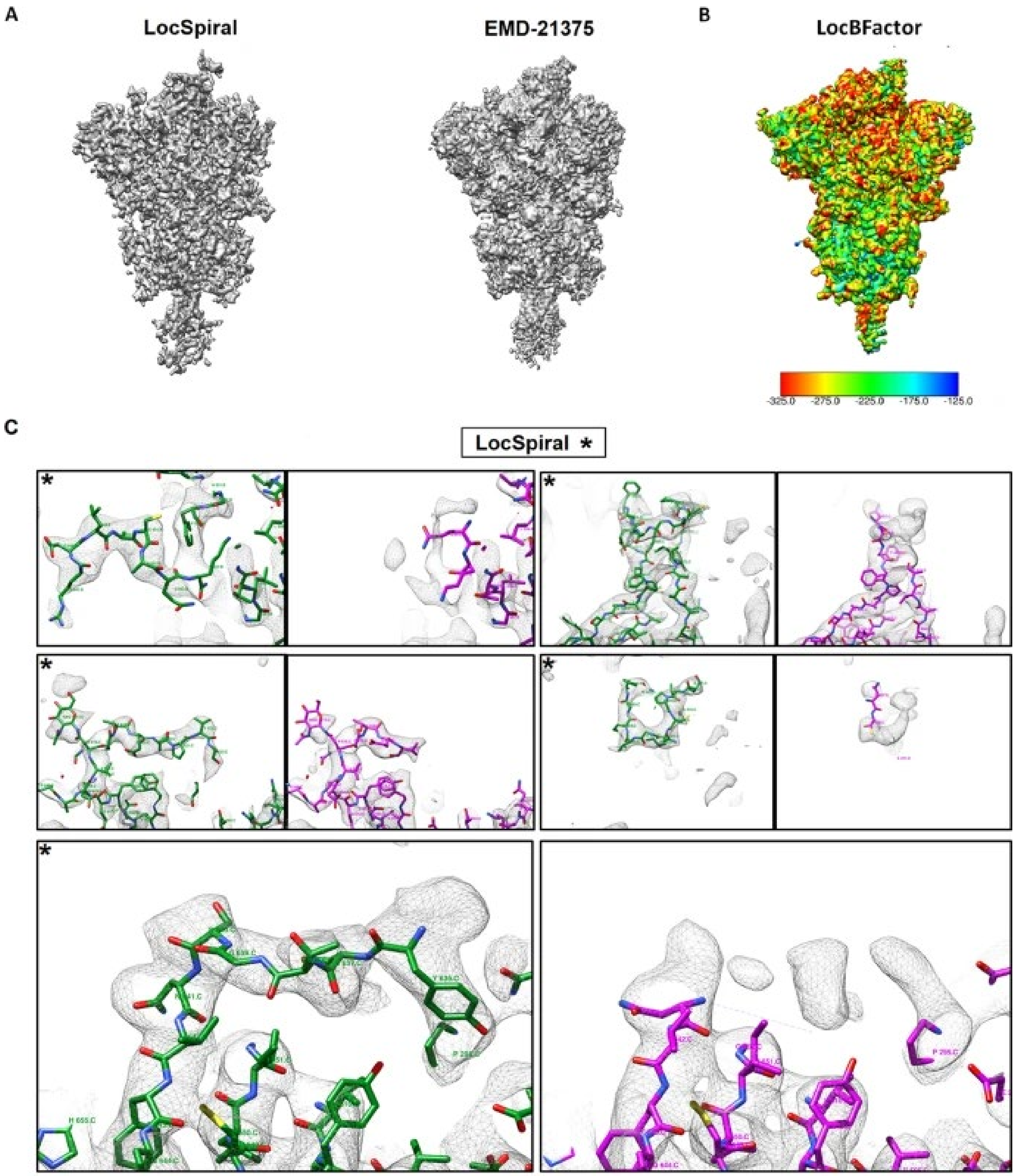
In the Supplementary Note 2 and Supplementary Figure 2.4, we further show the advantages of LocSpiral and LocBFactor approaches in these highly heterogeneous datasets compared to the global sharpening approach.

2.3.6 SARS-CoV-2

We have processed recent cryo-EM maps of the SARS-CoV-2 spike (S) glycoprotein[27,28]. These maps include cryo-EM reconstructions of the SARS-CoV-2 spike in the prefusion conformation with a single receptor-binding domain (RBD) up (EMD-21375) and after imposing C3 symmetry in the refinement to improve visualisation of the symmetric S2 subunit (EMD-21374). We also processed additional cryo-EM reconstructions from the Veesler lab of the SARS-CoV-2 spike glycoprotein with three RBDs down (EMD-21452) and the SARS-CoV-2 spike ectodomain structure (EMD-21457) with a single RBD up. The reported global resolution of these maps is 3.46, 3.17, 2.8 and 3.2 Å, respectively. Interesting deposited atomic models (PDBs PDB 6vsb, PDB 6vxx and PDB 6vyb) incompletely cover the reconstructed cryo-EM maps, showing the existence of disordered or over-sharpened regions after *B*-factor correction that could not be modelled. Supplementary Figure 2.5 displays corresponding maps and fitted atomic models showing a large amount of protein that is not currently modelled.

In Figure 2.5A, we show EMD-21375 map and the obtained LocSpiral reconstruction. In this figure, we use a relatively low threshold to visualise the outer parts of the protein. This figure shows that our obtained reconstruction presents less fragmented and broken densities and better map connectivity than the one deposited in EMD, suggesting that our approach improves the analysis and visualisation of the outer regions and potentially aides in the modelling of additional map motifs. In Supplementary Figure 2.6A, we show similar results for EMD-21374, EMD-21452 and EMD-21457 maps. Interesting, the LocSpiral EMD-21374 map shows some additional fragmented densities at the top of the spike, however, we believe that these additional densities are in fact artefacts that come as a result of artificially imposing C3 symmetry on particles that are asymmetric. In Figure 2.5B, we show the local *B*-factor map to be used for sharpening (slope of the local Guinier plot multiplied by 4) obtained by LocBFactor for EMD-21375 and in Supplementary Figure 2.6B, we compare obtained local *B*-factor maps from EMD-21375, EMD-21374, EMD-21452 and EMD-21454 maps using a similar colourmap. Supplementary Figure

2.6B shows that EMD-21452 and EMD-21454 present lower *B*-factors than EMD-21374 and EMD-21374, and then a better localizability of secondary structure and residues.



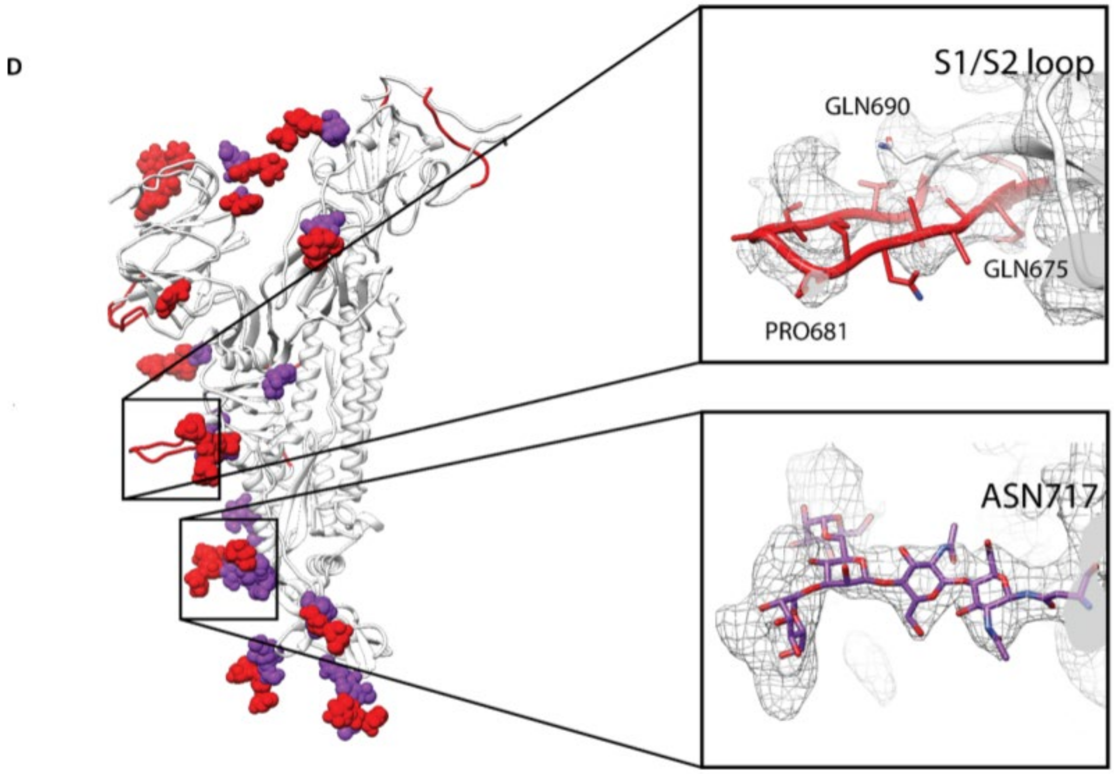


Figure 2.5: Results obtained by LocSpiral, LocBFactor and improved atomic model for EMD-21375 SARS-CoV-2 sample. *A Map obtained by LocSpiral approach (left) compared with the map as deposited in EMDB for EMD-21375. B B-factor maps to be used for sharpening (slope of the local Guinier plot multiplied by 4) obtained by LocBFactor approach for EMD-21375. The B-factor ranges between [−325.0, −125.0] Å2. C Visual examples of map regions corresponding to EMD-21375 that could be further modelled after processing the corresponding unfiltered and unsharpened map with LocSpiral approach. On the left and marked with asterisks, we show the LocSpiral maps with the improved atomic models in green, and on the right the deposited EMD-21375 map with the PDB 6vsb in magenta. D In white, PDB 6vsb with traced parts of the glycan proteins marked with purple spheres. In red, additional parts that could be traced using LocSpiral map. Inside the black squares, zoomed views of two glycan proteins that could be further modelled.*

Then, we used the LocSpiral EMD-21375 reconstruction to improve the deposited atomic model (PDB 6vsb). As result, we could model additional loops and motifs: K444.C-F490.C; E96.C-S98.C; NAG1322.C; P812.C-K814.C, and some additional amino acids, which are now visible in the improved map: P621.C-G639.C; S673.C-V687.C; A829.B-A825.B. We were also able to visualise map densities corresponding to numerous additional *N*-linked glycans that could not be resolved in the original reconstruction. Examples of some regions that could be further modelled are shown in Figure 2.5C, D. In Figure 2.5C, we show the obtained LocSpiral map with the improved atomic model in green at the left and marked with an asterisk. At the right, it is rendered the deposited EMD map with the PDB 6vsb in magenta. Figure 2.5D shows in white the PDB 6vsb with the traced parts of the glycan proteins marked with purple spheres and in red the additional parts that could be traced using LocSpiral map. In addition, in this figure, we provide also zoomed views of two glycan proteins that could be further modelled with our improved map. Corresponding EMRINGER and MolProbity scores, calculated between LocSpiral map and the improved atomic model, and between EMD-21375 and the deposited model (PDB 6vsb), are shown in Supplementary Table S1. In both cases, the atomic structures were refined against corresponding maps by Phenix real_space_refine approach[25] using 5 refining iterations.

2.4 Discussion

In this paper, we have introduced methods to improve the analysis and interpretability of cryo-EM maps. These methods include map enhancement approaches (LocSpiral and LocBSharpen), and approaches to calculate local *B*-factors (LocBFactor) and density occupancy maps (LocOccupancy). We have shown in our experiments that LocSpiral approach improves map connectivity showing fewer fragmented and broken densities and better coverage of the atomic model. In fact, our LocSpiral approach has been applied on several published publications[31-35], enabling molecular modelling on maps with flexibility and light anisotropic resolution.

We envision that our proposed methods to estimate local *B*-factors and occupancy maps could be used to improve de novo model building. First, these maps can be employed to guide the manual tracing. These maps can be informative to estimate the range of structures that could be compatible with the given electron microscopy density. Second, for very high-resolution cryo-EM maps, these values can be used as an approximation of the atomic *B*-factors and occupancies to be further refined as part of the automatic model refinement process by automatic model building packages as Phenix[14] or Refmac[15]. *B*-factor maps provide complementary information to local resolution maps, though, these results are usually correlated. The latter usually determines the resolution at a given point by comparing the map to noise or background amplitudes[36], while the former determines the rate of signal amplitude fall off within a resolution range. Then, we can find map regions with similar local resolution (map amplitude similar to noise/background amplitude at this resolution and coordinates), while different *B*-factor as the signal damping could be different within the used resolution range (highly or slowly sloped).

We have seen that we must be careful when processing maps affected by high flexibility and heterogeneity or when analysing maps with moderate global resolution (close to 10–15 Å) as the obtained *B*-factors could be overestimated if the selected resolution range is above the local resolution at these regions. Note that obtained *B*-factors at these low-resolution regions describe mainly noise *B*-factors that show how the noise signal fall off inside the used resolution range and they should be filtered out from our *B*-factor map. However, these problematic cases can be easily detected as the amplitude values in corresponding Guinier plot will be below the noise level (obtained from the 90–95% quantile of the empirical noise/background distribution). Thus, these regions can be automatically filtered out and not taken into consideration. In our analysis of *B*-factors for low and high-resolution maps shown in the Supplementary Material, we show that existing methods to determine the map global *B*-factor, as Relion postprocessing, do not filter problematic low-resolution regions so the estimated *B*-factor may be overestimated.

In principle, it might be possible to differentiate between compositional and moderate conformational flexibility from the obtained occupancy maps for samples accurately 3D classified.

In the former case, the occupancy map is expected to show close to zero values at missing regions, as the density values of these parts should be low and close to the noise level. Oppositely, in the latter case, the occupancy is likely to show higher values as the density values of moving parts, while slightly blurred because of the movement, should be similar to the ones at other static regions of the macromolecule. However, we should be extremely careful about these analyses as 3D classification approaches are not perfect, thus, macromolecules showing different compositions could provide 3D maps with significant density values in regions that should be empty. Additionally, samples showing large conformational changes could present low-density values at moving regions when compared to density values at static parts, providing close-to zero occupancy values.

The methods proposed here are semi-automated and essentially only require the unfiltered map to enhance or analyse, a resolution range and, in some cases, a binary solvent mask as inputs. They do not require additional information as atomic models or local resolution maps. The common link between all these approaches is the use of the spiral phase transform, which is used to factorise cryo-EM maps into amplitude and phase terms in real space for different resolutions. The spiral phase transform has been extensively used in optics for phase extraction in interferometry[35-39] or by Shack-Hartmann sensors[40,41]. This transformation is not new in cryo-EM as it has been proposed previously to facilitate particle screening[42], CTF estimation[43] and local and directional resolution determination[34,44]. In refs. [34,44], the authors used the Riesz transform to obtain amplitude maps, which is similar to the spiral phase transform.

Cryo-EM reconstructions of different types of macromolecules have been used to test the performance of these algorithms. Specifically, we have used a membrane protein (TRP channel), immature ribosomes affected by high compositional heterogeneity, the spliceosome that shows high conformational heterogeneity, recent SARS-CoV-2 reconstructions exhibiting dynamic regions and high-resolution apoferritin reconstructions. In all cases, our proposed approaches show excellent results, improving the analysis and the interpretability of the processed maps. The proposed methods are also highly efficient. For example, the processing of EMD-21457 (map size 400 px^3) using our local enhancement approach took only 12 min on a standard laptop using 4 cores.

2.5 Methods

The proposed methods are based on a 3D generalisation of the 2D spiral phase transform. In the following, we present the 3D spiral phase transform and its application to map enhancement, local B-factor determination, and estimation of local map occupancies. 3D spiral phase transform. The spiral phase transform is a Fourier operator that can factorise a 3D map into its amplitude and phase terms in real space at different resolutions. We assume without loss of generality that a given 3D map can be modelled as a 3D phase modulated signal given by

$$V(\mathbf{r}) = \sum_{\omega} V_{\omega}(\mathbf{r}) = \sum_{\omega} \left(b_{\omega}(\mathbf{r}) + m_{\omega}(\mathbf{r})\cos\left(\varphi_{\omega}(\mathbf{r})\right)\right) \tag{3}$$

where $V(\mathrm{r})$ is the cryo-EM map, $V_{\omega}(\mathbf{r})$ is a band-passed map filtered at frequency ω, $b_{\omega}(\mathbf{r})$ the 3D background or DC term, $m_{\omega}(\mathbf{r})$ the 3D amplitude map, φ_{ω} the 3D modulating phase and $\mathbf{r} = (x, y, z)$. Assuming that we are interested in spatial frequencies higher than 1/50–1/30 1/Å and that the background is usually a low frequency signal, we can approximate the map by a high-passed filtered map VHP for resolutions higher than 50–30 Å by

$$V_{\mathrm{HP}}(\mathbf{r}) = \sum_{\omega} m_{\omega}(\mathbf{r})\cos\left(\varphi_{\omega}(\mathbf{r})\right) \tag{4}$$

For convenience, Eq. (4) can be expanded into its corresponding analytic signal as

$$\tilde{V}_{\mathrm{HP}}(\mathbf{r}) = \sum_{\omega} m_{\omega}(\mathbf{r}) e^{j\varphi_{\omega}(\mathbf{r})} \tag{5}$$

This analytic signal relates to our high-passed filtered map by

$$\begin{aligned} V_{\mathrm{HP}}(\mathbf{r}) &= \textstyle\sum_{\omega} \operatorname{Re}\{\tilde{V}_{\mathrm{HP}}(\mathbf{r})\} = \sum_{\omega} \operatorname{Re}\{m_{\omega}(\mathbf{r})e^{j\varphi_{\omega}(\mathbf{r})}\} = \\ V_{\mathrm{HP}}(\mathbf{r}) &= \textstyle\sum_{\omega} \operatorname{Re}\{(m_{\omega}(\mathbf{r})\cos(\varphi_{\omega}(\mathbf{r})) + j m_{\omega}(\mathbf{r})\sin(\varphi_{\omega}(\mathbf{r})))\} \end{aligned} \tag{6}$$

with Re{·} an operator that takes the real part and j is the imaginary unit ($j^2 = -1$). Note from the analytic signal defined in Eq. (5) that $m_\omega(\mathbf{r})$ and $\varphi_\omega(\mathbf{r})$ clearly represent amplitude and phase terms. The quadrature transformation of Eq. (4) is given by

$$\mathrm{Q}\{V_{\mathrm{HP}}(\mathbf{r})\} = -\textstyle\sum_{\omega} m_{\omega}(\mathbf{r})\sin(\varphi_{\omega}(\mathbf{r})) \tag{7}$$

Then, Eq. (5) may be rewritten as

$$\tilde{V}_{\mathrm{HP}}(\mathbf{r}) = \textstyle\sum_{\omega} (V_{\mathrm{HP}}(\mathbf{r}) - j\mathrm{Q}\{V_{\mathrm{HP}}(\mathbf{r})\}) \tag{8}$$

Assuming that m_ω is a low varying map compared to φ_ω, the gradient of V_{HP} is approximated by

$$\nabla V_{\mathrm{HP}}(\mathbf{r}) \cong -\textstyle\sum_{\omega} m_{\omega}(\mathbf{r}) \sin(\varphi_{\omega}(\mathbf{r}))\nabla\varphi_{\omega}(\mathbf{r}) \tag{9}$$

Rearranging terms, we obtain

$$\mathrm{Q}\{V_{\mathrm{HP},\omega}(\mathbf{r})\} = \frac{\nabla\varphi_{\omega}(\mathbf{r})}{|\nabla\varphi_{\omega}(\mathbf{r})|} \cdot \frac{\nabla V_{\mathrm{HP},\omega}(\mathbf{r})}{|\nabla\varphi_{\omega}(\mathbf{r})|} = -\mathbf{n}_{\varphi}(\mathbf{r}) \cdot \frac{\nabla V_{\mathrm{HP},\omega}(\mathbf{r})}{|\nabla\varphi_{\omega}(\mathbf{r})|} \tag{10}$$

Equation (10) shows that the quadrature term is composed of two terms. The first is an orientation map $\mathbf{n}_\varphi$ and the second corresponds to a non-linear operator that can be interpreted as a 3D generalisation of the 1D Hilbert transform, which can be efficiently calculated using the Fourier transform. As shown in[45], the operator $\nabla V_{\mathrm{HP},\omega}(\mathbf{r})/|\nabla\varphi_{\omega}(\mathbf{r})|$ corresponds to the 3D Hilbert transform applied to our band-passed maps $V_{HP,\omega}(\mathbf{r})$, then

$$\mathbf{H}\{V_{\mathrm{HP},\omega}(\mathbf{r})\} = \mathrm{FT}^{-1}\left\{\frac{-j\mathbf{q}}{\mathbf{q}}\mathrm{FT}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}\right\} \cong \frac{\nabla V_{\mathrm{HP},\omega}(\mathbf{r})}{|\nabla\varphi_{\omega}(\mathbf{r})|} \tag{11}$$

Thus, Eq. (10) can be rewritten as

$$\mathrm{Q}\{V_{\mathrm{HP},\omega}(\mathbf{r})\} = \frac{\nabla\varphi_{\omega}(\mathbf{r})}{|\nabla\varphi_{\omega}(\mathbf{r})|} \cdot \frac{\nabla V_{\mathrm{HP},\omega}(\mathbf{r})}{|\nabla\varphi_{\omega}(\mathbf{r})|} \cong -\mathbf{n}_{\varphi}(\mathbf{r}) \cdot \mathrm{FT}^{-1}\left\{\frac{-i\mathbf{q}}{\mathbf{q}}\mathrm{FT}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}\right\} \tag{12}$$

Note that $\boldsymbol{n}_{\varphi}$ is a unit vector pointing in the same direction that $\nabla V_{\mathrm{HP},\omega}(\mathbf{r})$ (remember that m_{ω} is a low varying map compared to φ_{ω}), but maybe with different orientation because a possible change of sign introduced by the cosine term in Eq. (4). We can rewrite Eq. (12) as

$$\begin{aligned} \mathrm{Q}\{V_{\mathrm{HP},\omega}(\mathbf{r})\} \quad &\cong -\mathbf{n}_{\varphi}(\mathbf{r})\left|\mathrm{FT}^{-1}\left\{\frac{-i\mathbf{q}}{|\mathbf{q}|}\mathrm{FT}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}\right\}\right|\mathbf{n}_{V_{\mathrm{HP},\omega}}(\mathbf{r}) \\ &= -s(\mathbf{r})\left|\mathrm{FT}^{-1}\left\{\frac{-i\mathbf{q}}{|\mathbf{q}|}\mathrm{FT}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}\right\}\right| \end{aligned} \tag{13}$$

where $s(\mathbf{r})$ is a function with range $+1$ or -1 considering that $\mathbf{n}_{\varphi}(\mathbf{r})$ and $\mathbf{n}_{V_{\mathrm{HP.}}}$ can be parallel or antiparallel only. From Eq. (13), we can obtain an estimation of $\varphi_{\omega}(\mathbf{r}$ affected by an indetermination in its sign by

$$\begin{aligned} &\varphi_{\omega}(\mathbf{r}) \cong \arctan\left[\frac{\mathrm{Q}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}}{V_{\mathrm{HP},\omega}(\mathbf{r})}\right] = \\ &-s(\mathbf{r})\arctan\left[\frac{\left|\mathrm{FT}^{-1}\left\{\frac{-i\mathbf{q}}{|\mathbf{q}|}\mathrm{FT}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}\right\}\right|}{V_{\mathrm{HP},\omega}(\mathbf{r})}\right] \end{aligned} \tag{14}$$

However, we can use Eq. (14) to obtain the modulation and cosine terms in Eq. (4) separately without sign ambiguity as

$$\begin{aligned} \cos\left(\varphi_{\omega}(\mathbf{r})\right) &= \cos\left(\arctan\left[\frac{\mathrm{Im}\{\tilde{V}_{\mathrm{HP}}(\mathbf{r})\}}{\mathrm{Re}\{\tilde{V}_{\mathrm{HP}}(\mathbf{r})\}}\right]\right) \\ &\cong \cos\left(\arctan\left[\frac{\left|\mathrm{FT}^{-1}\left\{\frac{-i\mathbf{q}}{|\mathbf{q}|}\mathrm{FT}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}\right\}\right|}{V_{\mathrm{HP},\omega}(\mathbf{r})}\right]\right) \\ m_{\omega}(\mathbf{r}) = \left(\tilde{V}_{\mathrm{HP}}(\mathbf{r})\cdot\tilde{V}_{\mathrm{HP}}(\mathbf{r})^{+}\right)^{1/2} &= \left(\left(V_{\mathrm{HP},\omega}(\mathbf{r})\right)^{2} + \left(\mathrm{Q}\{V_{\mathrm{HP},\omega}(\mathbf{r})\}\right)^{2}\right)^{1/2} \end{aligned} \tag{15}$$

Using these expressions, we can obtain for each frequency ω the terms $\cos\left(\varphi_{\omega}(\mathbf{r})\right)$ and $m_{\omega}(\mathbf{r})$.

2.5.1 Local enhanced map (LocSpiral)

We are proposing here a robust local map enhancement method that only requires as input a binary mask of the macromolecule and a resolution range. The approach works for both high and moderate resolution maps. In the following, we provide details of the proposed method. As explained before, each band-pass filtered map can be factorised into an amplitude and phase term by the spiral phase transform. Then, given a user defined solvent mask, the method obtains the empirical noise amplitude probability distribution (m_ω^N) at frequency ω, selecting the density values of voxels not included in the solvent mask. From this distribution, the approach determines the noise amplitude value corresponding to the 90–95% quantile, given by $m_\omega^N(q = 95\%)$. This value is used to locally normalise map amplitudes in real space along with different frequencies and remove local signals that are below this amplitude threshold as they are likely noise at this given frequency and position. After this non-linear amplitude transformation, the enhanced map at a given frequency ω is given by

$$\check{V}_\omega(\mathbf{r}) = (m_\omega(\mathbf{r}) > m_\omega^N(q = 95\%))\cos(\varphi_\omega(\mathbf{r})) \tag{16}$$

and the map

$$\check{V}(\mathbf{r}) = \sum_\omega \check{V}_\omega(\mathbf{r}) = \sum_\omega (m_\omega(\mathbf{r}) > m_\omega^N(q = 95\%))\cos(\varphi_\omega(\mathbf{r})) \tag{17}$$

The method allows as an option the use of an SNR weighting parameter to weight the contribution of the different amplitudes in the final map. In this case, Eq. (17) is rewritten as

$$\check{V}(\mathbf{r}) = \sum_\omega C_{\mathrm{ref},\omega}(\mathbf{r})(m_\omega(\mathbf{r}) > m_\omega^N(q = 95\%))\cos(\varphi_\omega(\mathbf{r})) \tag{18}$$

with $C_{\mathrm{ref},\omega}(r)$ the SNR weighting parameter given by

$$C_{\mathrm{ref},\omega}(\mathbf{r}) = \frac{m_\omega(\mathbf{r})}{m_\omega(\mathbf{r}) + m_\omega^N(q=95\%)} \tag{19}$$

2.5.2 Local B-factor determination (LocBFactor)

The factorisation of a 3D map into its amplitude and phase terms in real space for different frequencies allows the efficient determination of local B-factor maps. To this end, LocBFactor method first obtains the local map amplitudes $m_\omega(\mathbf{r})$ for resolutions between 15–10 Å to the estimated global map resolution. These amplitude maps are then used to obtain SNR-weighted log-amplitudes of structure factors locally as

$$\log\left(F_{\omega}(\mathbf{r})\right) = \log\left(C_{\mathrm{ref},\omega}(\mathbf{r}) m_{\omega}(\mathbf{r})\right) \tag{20}$$

with $C_{\mathrm{ref},\omega}(\mathbf{r})$ a SNR weighting parameter defined in (19). This expression can be used to fit $\log(F_{\omega}(\mathbf{r}))$ versus ω^2 within the resolution rage defined between 15 and 10 Å to the estimated global map resolution. Thus, finally we have

$$\log\left(F_{\omega}(\mathbf{r})\right) \cong B(\mathbf{r})(\omega^2 - \omega_0^2) + A(\mathbf{r}) \tag{21}$$

with $B(\mathbf{r})$ the local B-factor map or B map, and $\mathrm{A}(\mathbf{r})$ the log-amplitude map at ω_0 (A map). In Eq. (21) the approach typically does not take into consideration in the linear fit amplitude values $(m_{\omega}(\mathbf{r}))$ that are below the noise level $(m_{\omega}^{N}(q = 95\%))$. Additionally, local Guinier plots without at least two points above the noise level are filtered out from the B map. Note that ω_0 corresponds to the lowest frequency within the used resolution range (typically $1/15 - 1/10\text{Å}^{-1}$).

2.5.3 Local B-factor sharpened map (LocBSharpen)

The spiral phase transform can be used to obtain local B-factor sharpened maps. Note that Expression (4) can be modified for frequencies higher than ω_0 as

$$\check{V}(\mathbf{r}) = \sum_{\omega} \check{V}_{\mathrm{HP},\omega}(\mathbf{r}) = \begin{cases} \sum_{\omega} \left(C_{\mathrm{ref},\omega}(\mathbf{r}) m_{\omega}(\mathbf{r}) \cos\left(\varphi_{\omega}(\mathbf{r})\right)\right), \omega < \omega_0 \\ \sum_{\omega} \left(C_{\mathrm{ref},\omega}(\mathbf{r}) A(\mathbf{r}) \cos\left(\varphi_{\omega}(\mathbf{r})\right)\right), \omega \geq \omega_0 \end{cases} \tag{22}$$

With $A(\mathbf{r})$ the log-amplitude map at ω_0 (A map).

2.5.4 Local occupancy map (LocOccupancy)

Low occupancy map regions correspond to parts of the macromolecule where map amplitudes of the reconstruction are significantly smaller when compared to other regions of the macromolecule. Keeping this in mind, we define the occupancy map as

$$O(\mathbf{r}) = \frac{\sum_{\omega} \left(m_{\omega}(\mathbf{r}) > m_{\omega}^{M}(q=25\%)\right)}{\sum_{\omega} \left(m_{\omega}(\mathbf{r}) \geq m_{\omega}^{M}(q=0\%)\right)} \tag{23}$$

where $m_{\omega}^{M}(q = 25\%)$ and $m_{\omega}^{M}(q = 0\%)$ are obtained from the empirical macromolecule amplitude probability distribution (m_{ω}^{M}) at frequency ω. This amplitude probability distribution is calculated from map density values corresponding to voxels that are included in the solvent mask. From this distribution, the approach determines the macromolecule amplitude values corresponding to the 25 and 0% quantiles, given by $m_{\omega}^{M}(q = 25\%)$ and $m_{\omega}^{M}(q = 0\%)$ that are used as thresholds. To calculate local occupancy maps, a typical resolution range between 30 and $10 - 8$Å is used to obtain density occupancies of complete secondary structure motifs, while ranges between 5 and 3-1.5 Å are used for high-resolution cryo-EM maps to obtain occupancies of residues.

2.5.5 Maturity level index

In the analysis of the immature 50S ribosomes, we have proposed a maturity level index. This index can be extended to the analysis of any maturing macromolecule and is useful to place immature macromolecules into a maturing timeline. The calculation of this index requires reconstructions of immature and mature macromolecules. The mature reconstruction is used to obtain a binary solvent mask, while the immature reconstructions are used to calculate occupancy maps. These occupancy maps allow us to determine highly occupied regions (occupancy >0.75) and calculate occupancy masks. Then, the index is obtained comparing the number of voxels activated in the solvent mask of the mature reconstruction with the ones in the occupancy masks.

As can be seen from Figure 2.4, the larger are the regions that are not folded in the immature maps, the smaller is the maturity level.

2.5.6 Cryo-EM image processing of the spliceosome data

The dataset is composed of 327,490 particle images of a spliceosomal B-complex from yeast (EMPIAR-10180)[4]. The particles were polished with Relion, downsampled to 1.699 Å/px and windowed to a size of 320 × 320 pixels. A set of 30 initial volumes were obtained by RANSAC (15 maps) and Eman2 (15 maps) and processed by volume selector approach[22] producing two different initial volumes. Then, Relion 3D classification was used to compute two classes providing both volumes as reference initial maps (class 1 and class 2 composed by 201,407 and 126,083 particles respectively). The resulting classes were refined by Relion autorefine using the maps obtained in the previous 3D classification. Finally, Relion postprocessing provided maps at 4.28 and 4.58 Å for class 1 and class 2, respectively. Lastly, a local resolution was calculated using Relion for both classes.

2.5.7 Data availability

Previously published datasets used for testing are available from the Electron Microscopy Data Bank (https://www.ebi.ac.uk/pdbe/emdb/) under accession codes EMD-10418, EMD-8440, EMD-8441, EMD-8445, EMD-8450, EMD-8434, EMD-21375, EMD-21374, EMD-21452 and EMD-21457. Data that support the findings of this study have been deposited in http://t.ly/XKQa.

2.5.8 Code availability

The source code for the presented methods is freely available under the terms of an opensource software license and can be downloaded from https://github.com/1aviervargas/ LocSpiral-LocBSharpen-LocBFactor-LocOccupancy[21].

2.6 References

1. Wandzik, J. M., Kouba, T., Karuppasamy, M., Pflug, A., Drncova, P., Provaznik, J., ... & Cusack, S. A structure-based model for the complete transcription cycle of influenza polymerase. *Cell* **181**, 877-893 (2020).
2. Ge, P., Scholl, D., Prokhorov, N. S., Avaylon, J., Shneider, M. M., Browning, C., ... & Zhou, Z. H. Action of a minimal contractile bactericidal nanomachine. *Nature* **580**, 658-662 (2020).
3. Davis, J. H. et al. Modular assembly of the bacterial large ribosomal subunit. *Cell* **167**, 1610–1622 (2016).
4. Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617–621 (2017).
5. Razi, A., Davis, J. H., Hao, Y., Jahagirdar, D., Thurlow, B., Basu, K., ... & Ortega, J. Role of Era in assembly and homeostasis of the ribosomal small subunit. *Nucleic acids research* **47**, 8301-8317 (2019).
6. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

7. Fernandez, J. J., Luque, D., Caston, J. R. & Carrascosa, J. L. Sharpening high resolution information in single particle electron cryomicroscopy. *J. Struct. Biol*. **164**, 170–175 (2008).

8. Scheres, S. H. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189**, 114–122 (2015).

9. Terwilliger, T. C., Sobolev, O. V., Afonine, P. V. & Adams, P. D. Automated map sharpening by maximization of detail and connectivity. *Acta Crystallogr. D Struct. Biol.* **74**, 545–559 (2018).

10. Murshudov, G. N. Refinement of atomic structures against cryo-EM maps. *Methods Enzymol.* **579**, 277–305 (2016).

11. Jakobi, A. J., Wilmanns, M. & Sachse, C. Model-based local density sharpening of cryo-EM maps. *eLife* **6**, (2017).

12. Ramirez-Aportela, E. et al. Automatic local resolution-based sharpening of cryo-EM maps. *Bioinformatics* **36**, 765–772 (2020).

13. Sherwood, D., Cooper, J. & Sherwood, D. Crystals, X-rays, and Proteins: Comprehensive Protein Crystallography (2011).

14. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).

15. Winn, M. D., Murshudov, G. N. & Papiz, M. Z. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol.* **374**, 300–321 (2003).

16. Penczek, P. A. Image restoration in cryo-electron microscopy. *Methods Enzymol.* **482**, 35–72 (2010).

17. Liao, H. Y. & Frank, J. Definition and estimation of resolution in singleparticle reconstructions. *Structure* **18**, 768–775 (2010).

18. Afonine, P. V. et al. New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr. D Struct. Biol.* **74**, 814–840 (2018).

19. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryoEM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, (2016).

20. Wang, Q. et al. Lipid interactions of a ciliary membrane trp channel: simulation and structural studies of polycystin-2. *Structure* **28**, 169–184.e165 (2020).

21. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods* **13**, 387–388 (2016).

22. Gomez-Blanco, J., Kaur, S., Ortega, J. & Vargas, J. A robust approach to ab initio cryo-electron microscopy initial volume determination. *J. Struct. Biol.* **208**, 107397 (2019).

23. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

24. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).

25. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D Struct. Biol*. **74**, 531–544 (2018).

26. Lawson, C. L. et al. EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* **39**, D456–D464 (2011).

27. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).

28. Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281-292 (2020).

29. Khalifa, A. A. Z. et al. The inner junction complex of the cilia is an interaction hub that involves tubulin post-translational modifications. *eLife* **9**, e52760 (2020).

30. Ichikawa, M. et al. Tubulin lattice in cilia is in a stressed form regulated by microtubule inner proteins. *J. Proc. Natl Acad. Sci*. **116**, 19930-19938 (2019).

31. Yang, M. et al. Cryo-electron microscopy structures of ArnA, a key enzyme for polymyxin resistance, revealed unexpected oligomerizations and domain movements. *J. Struct. Biol.* **208**, 43–50 (2019).

32. Gutmann, T. et al. Cryo-EM structure of the complete and ligand-saturated insulin receptor ectodomain. *J. Cell Biol*. **219**, 1 (2020).

33. Jahagirdar, D. et al. Alternative conformations and motions adopted by 30S ribosomal subunits visualized by cryo-electron microscopy. *RNA* **26**, 2017–2030 (2020).

34. Vilas, J. L. et al. MonoRes: automatic and accurate estimation of local resolution for electron microscopy maps. *Structure* **26**, 337–344. e334 (2018).

35. Vargas, J., Restrepo, R., Quiroga, J. A. & Belenguer, T. High dynamic range imaging method for interferometry. *Opt. Commun*. **284**, 4141–4145 (2011).

36. Larkin, K. G., Bone, D. J. & Oldfield, M. A. Natural demodulation of twodimensional fringe patterns. I. General background of the spiral phase quadrature transform. *J. Opt. Soc. Am. A* **18**, 1862–1870 (2001).

37. Antonio Quiroga, J. & Servin, M. Isotropic n-dimensional fringe pattern normalization. *Opt. Commun.* **224**, 221–227 (2003).

38. Vargas, J., Quiroga, J. A., Sorzano, C. O., Estrada, J. C. & Carazo, J. M. Twostep interferometry by a regularized optical flow algorithm. *Opt. Lett.* **36**, 3485–3487 (2011).

39. Vargas, J., Quiroga, J. A., Sorzano, C. O., Estrada, J. C. & Servin, M. Multiplicative phase-shifting interferometry using optical flow. *Appl. Opt.* **51**, 5903–5908 (2012).

40. Vargas, J., González-Fernandez, L., Quiroga, Juan, A. & Belenguer, T. Shack–Hartmann centroid detection method based on high dynamic range imaging and normalization techniques. *Appl. Opt.* **49**, 2409–2416 (2010).

41. Vargas, J. et al. Shack-Hartmann centroid detection using the spiral phase transform. *Appl. Opt.* **51**, 7362–7367 (2012).

42. Vargas, J. et al. Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *J. Struct. Biol.* **183**, 342–353 (2013).

43. Vargas, J. et al. FASTDEF: fast defocus and astigmatism estimation for highthroughput transmission electron microscopy. *J. Struct. Biol*. **181**, 136–148 (2013).

44. Vilas, J. L., Tagare, H. D., Vargas, J., Carazo, J. M. & Sorzano, C. O. S. Measuring local-directional resolution and local anisotropy in cryo-EM maps. *Nat. Commun*. **11**, 55 (2020).

45. Servin, M., Quiroga, J. A. & Marroquin, J. L. General n-dimensional quadrature transform and its application to interferogram demodulation. *J. Opt. Soc. Am. A* **20**, 925–934 (2003).

46. Kaur, S. et al. Local computational methods to improve the interpretability and analysis of cryo-EM maps. *Nat Commun* **12,** 1240 (2021).

2.6 Supplemental information

2.6.1 Supplementary note 1: Polycystin-2 (PC2) TRP channel

We compared the performance of LocSpiral with other methods, including LocalDeblur, our proposed local B-factor correction method (LocBSharpen) and the global B-factor correction approach as implemented in Relion. To compare the different results, we used metrics proposed in[1]. The results are shown in Supplementary Figure 2.2. In this case, we used a relatively high threshold value to visually compare the different maps. From Supplementary Figure 2.2, we can see that the map obtained by LocSpiral shows good connectivity and is less affected by broken or missed densities. EMRINGER[2] and cross-correlation scores (obtained using PDB 6t9n as reference) show approximately similar results for all cases, though, the highest scores are provided by LocSpiral and LocBSharpen approaches. For the sake of comparison, we also provide FSC curves calculated by comparing the different maps with the reference atomic model (PDB 6t9n). In this case, the best results at high resolutions are provided by LocalDeblur and by LocSpiral approaches.

In addition, we provide results of LocBFactor and LocOccupancy methods. In Supplementary Figure 2.2B, we show the obtained local B-factor map (left map) to be used for sharpening (slope of the local Guinier plot multiplied by a factor 4) and the A-map (middle map) corresponding to the local values of the logarithm of structure factors amplitudes at 15 Å. The resolution range used to estimate these maps was between 15 Å to the FSC resolution (2.96 Å). The average value inside the solvent mask of local B-factors obtained from amplitude values above the noise level (signal B-factors) gives a value of -129.76 $Å^2$, which is smaller than the value provided by Relion (-84.56 $Å^2$) computed from the unmasked reconstruction. The A-map provides the fitted local amplitudes at 15 Å, showing the local "amount" of signal at this resolution. As expected, local B-factor map

(B-map) shows that the inner parts of the protein show lower Bfactors than the outer regions. Supplementary Figure 2.2B shows additionally the obtained local occupancy map (right map). Interesting, both the occupancy and A maps show low values in the regions occupied by detergent densities, lipid densities and cholesterol densities (please see Figure 2 in[3]), indicating the presence of compositional variability in these regions and low signal at 15 Å.

2.6.2 Supplementary note 2: Immature prokaryote ribosomes

In Supplementary Figure 2.4 we show maps with improved contrast at high-resolution obtained after processing EMD-8441 by LocSpiral and Relion methods[4,5]. The same soft mask was applied to both maps. In the figure, we show the maps at low and high threshold values. When a low threshold value is used, it is not possible to see details in the Relion map, while at high threshold values many regions of this map are not visible. Conversely, LocSpiral approach shows high resolution features at both high and low thresholds without losing appreciable map densities.

Finally, we also show in Supplementary Figure 2.4 the local B-factor map (B map given by the slopes of the local Guinier plot) and the local values of the logarithm of structure factor's amplitudes at 15 Å (A map in the figure) obtained by LocBFactor approach. The average value of the local signal B-factors to be used for sharpening (slope of the local Guinier plot multiplied by 4) within a solvent mask is -394.28 Å^2. We obtained the B-factor estimations within a resolution range between 15 Å to the FSC resolution given by 3.7 Å. The B-factor map shows higher B-factors at the outer part of the macromolecule, corresponding to regions that are partially folded and show compositional and conformational heterogeneity and lower local resolutions, as can be seen from Supplementary Figure 2.4, Class 3 in[6]. As shown in the previous Spliceosome case, regions dominated by the noise signal within the used resolution range present artefactual low Bfactors (noise B-factors) which describe the noise fall off inside the resolution range. These noise

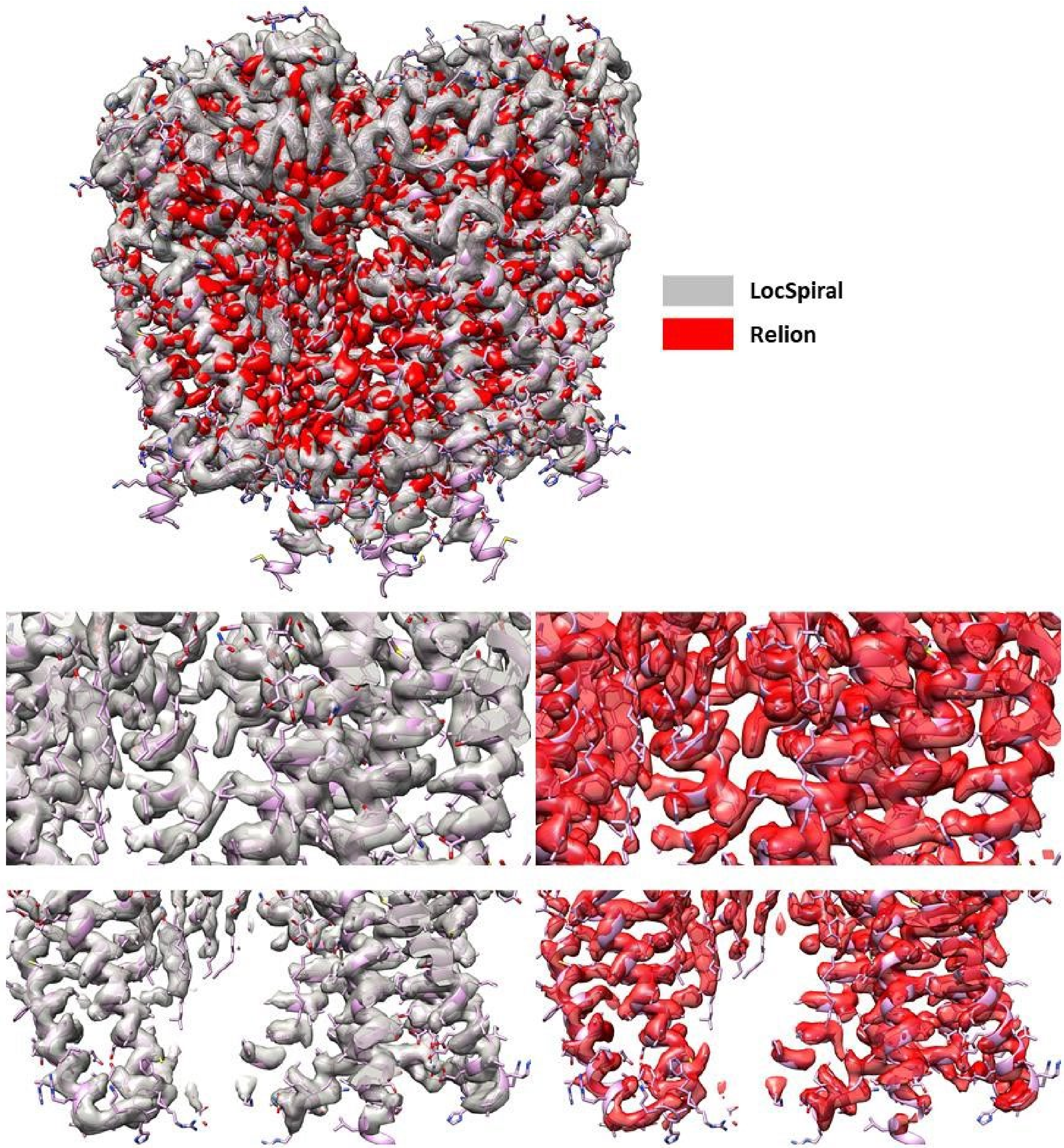
B-factors appear in Supplementary Figure 2.4 with black colour and correspond to the lowest amplitudes (below the noise level) in the A map at 15 Å resolution.

2.6.3 Supplementary note 3: B-factor analysis of low and high resolution maps

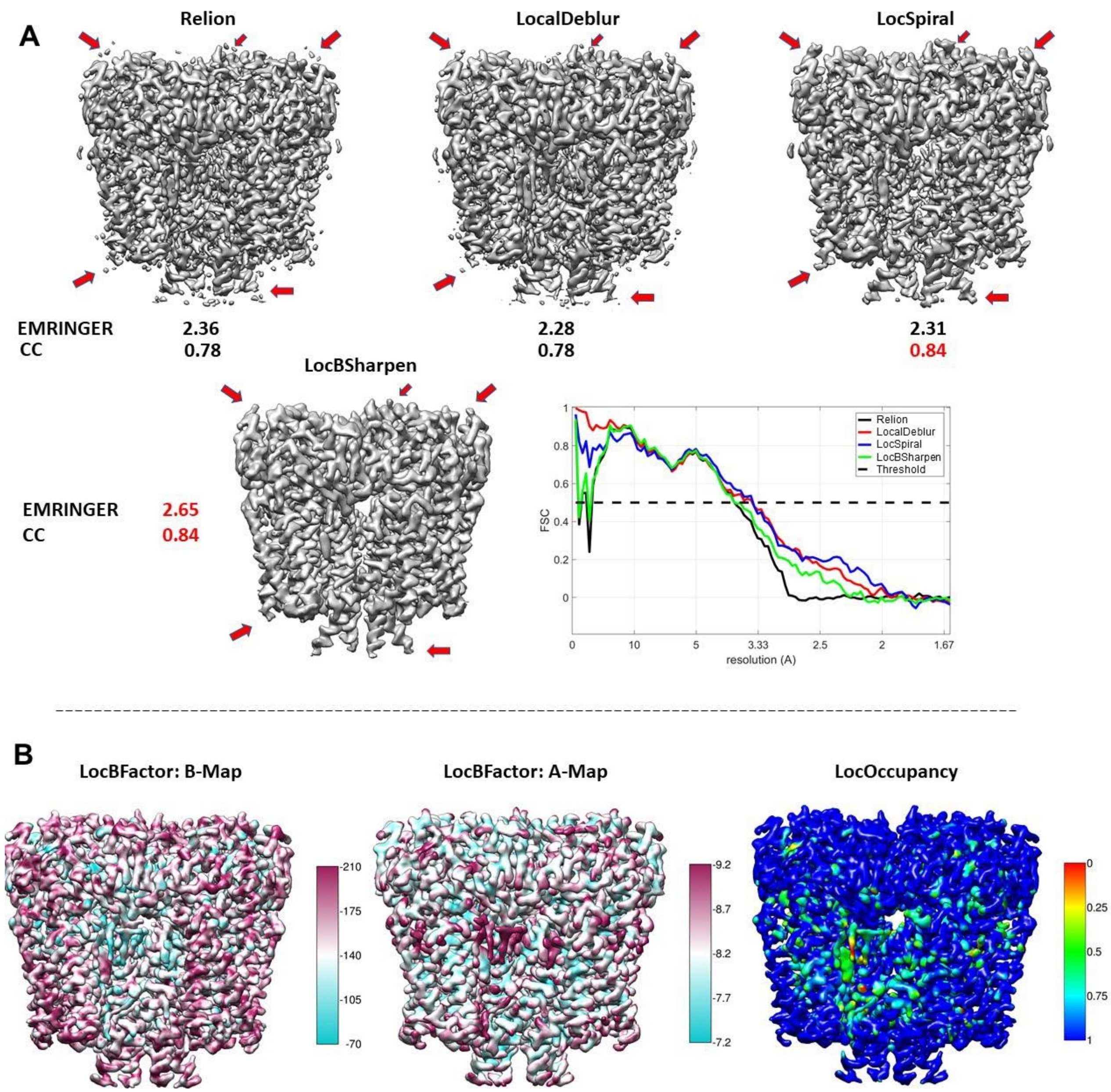
We have performed additional B-factor analysis of approximately homogeneous low- and highresolution maps corresponding to EMD-20671 and EMD-21024. The Gold standard FSC

resolution and the estimated B-factors to be used for sharpening (slope of the local Guinier plot multiplied by 4) as determined by Relion postprocessing are 16.01 Å and -97.70 $Å^2$ for EMD20671, and 1.77 Å and -50.81 $Å^2$ for EMD-21024, respectively. Note that the B-factors for sharpening obtained by Relion are very similar for maps showing very different resolutions. Additionally, we have computed local B-factor maps from LocBFactor approach. For both maps, we used a resolution range of [15, 4] Å. The obtained averages of signal B-factors used for sharpening inside solvent masks are -1172 $Å^2$ and -78 $Å^2$ for EMD-20671 and EMD-21024, respectively. Note that the values obtained by Relion and LocBFactor for EMD-21024 are similar. Oppositely, the average B-factor obtained by LocBFactor for EMD-20671 is much lower and consistent with a map at 16.01 Å resolution than the one reported by Relion. We believe that the reason of this discrepancy is because LocBFactor filters out noise B-factors (B-factors obtained from amplitudes below the noise level for the used resolution range) while Relion does not filter regions dominated by noise within the used resolution range. Supplementary Figure 2.7 shows obtained B-factor maps, FSC curves and respective Guinier plots at noise and signal regions for both cases.

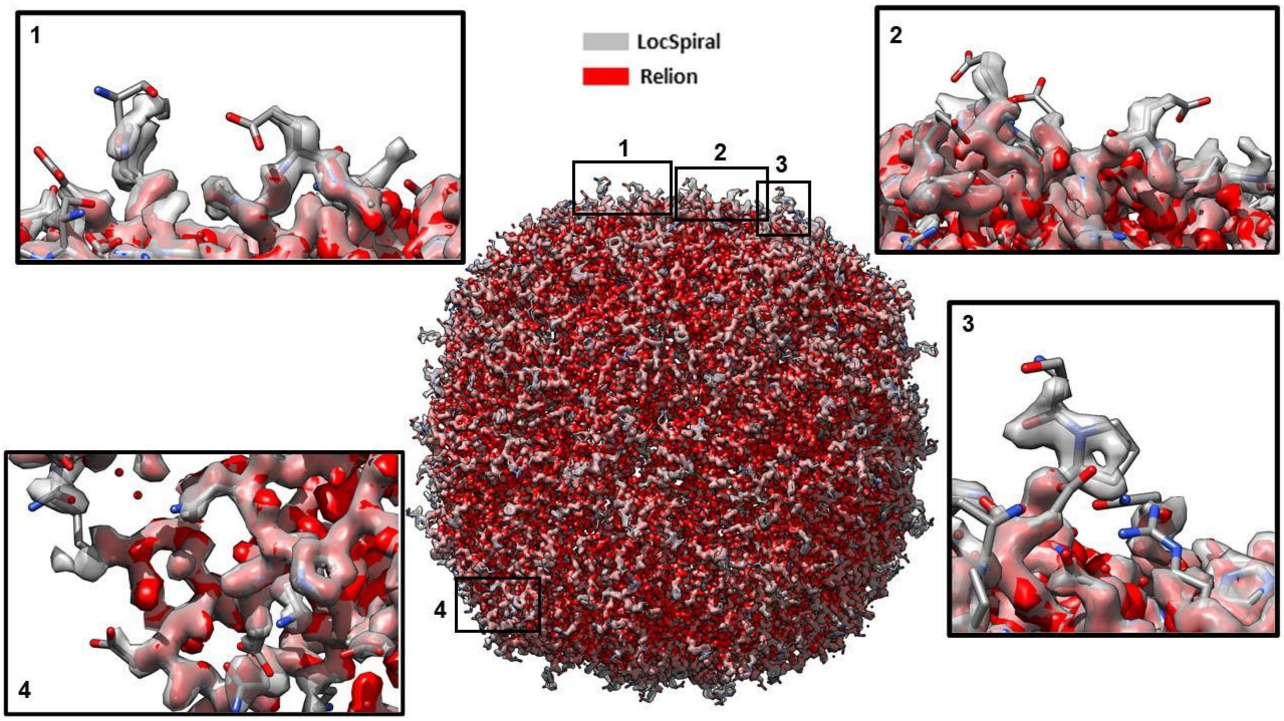
2.6.4 Supplementary figures



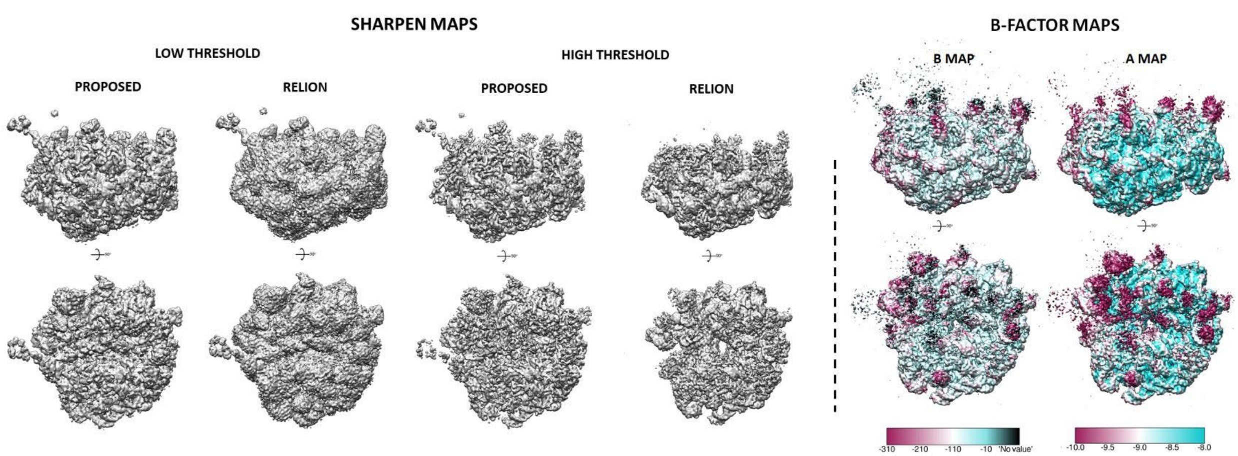
Supplementary Figure S2.1 Comparison between LocSpiral and Relion postprocessing maps for the TRP channel. *A) Complete and overlapping LocSpiral and Relion maps shown with the corresponding atomic structure (PDB 6t9n). B) Reconstructions of corresponding regions at the core and bottom outer region of the TRP channel obtained from LocSpiral (left) and Relion (right) approaches.*



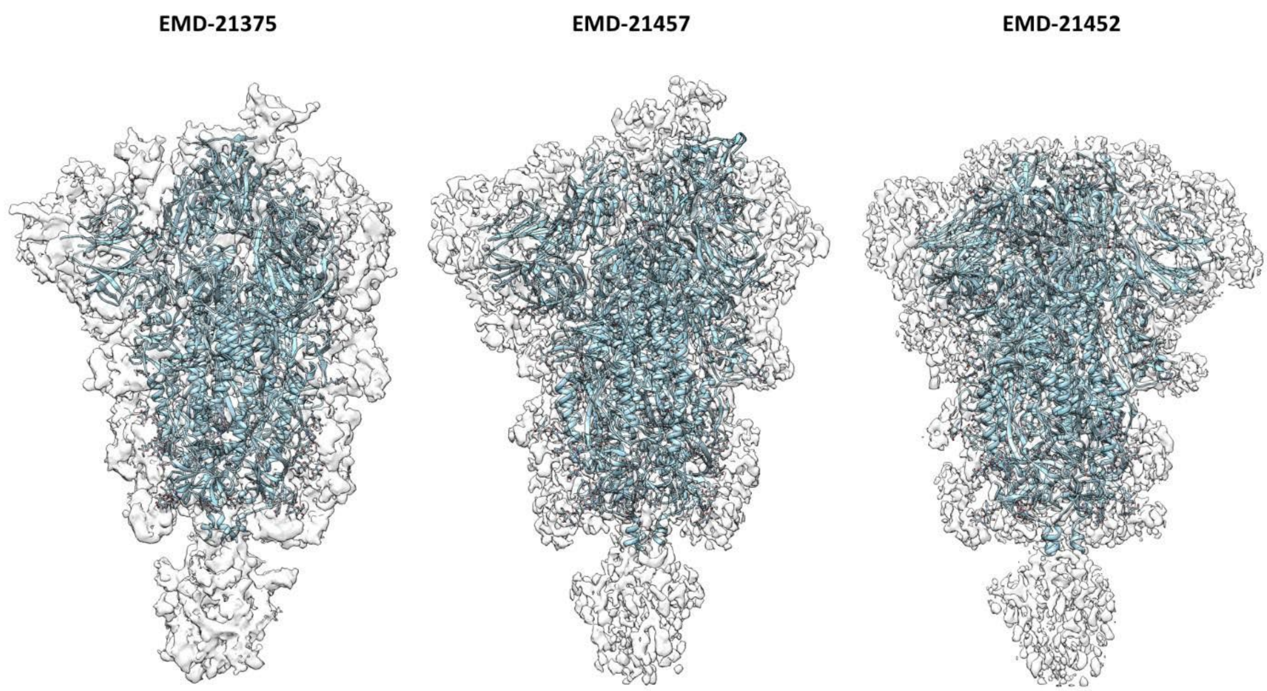
Supplementary Figure S2.2 Results and comparisons between different methods over the TRP channel. *A) Comparison between maps obtained by different sharpening approaches: Relion, LocalDeblur, LocSpiral and LocBSharpen. Red arrows show broken or missed densities that could be seen from LocSpiral and LocBSharpen maps. Below each map, EMRINGER and* *crosscorrelation (CC) scores calculated between obtained maps and the atomic model (PDB 6t9n) are provided. We also show FSC curves comparing the different maps with the reference atomic model (PDB 6t9n). B) Results obtained by LocBFactor (B and A maps) and LocOccupancy for the TRP channel.*



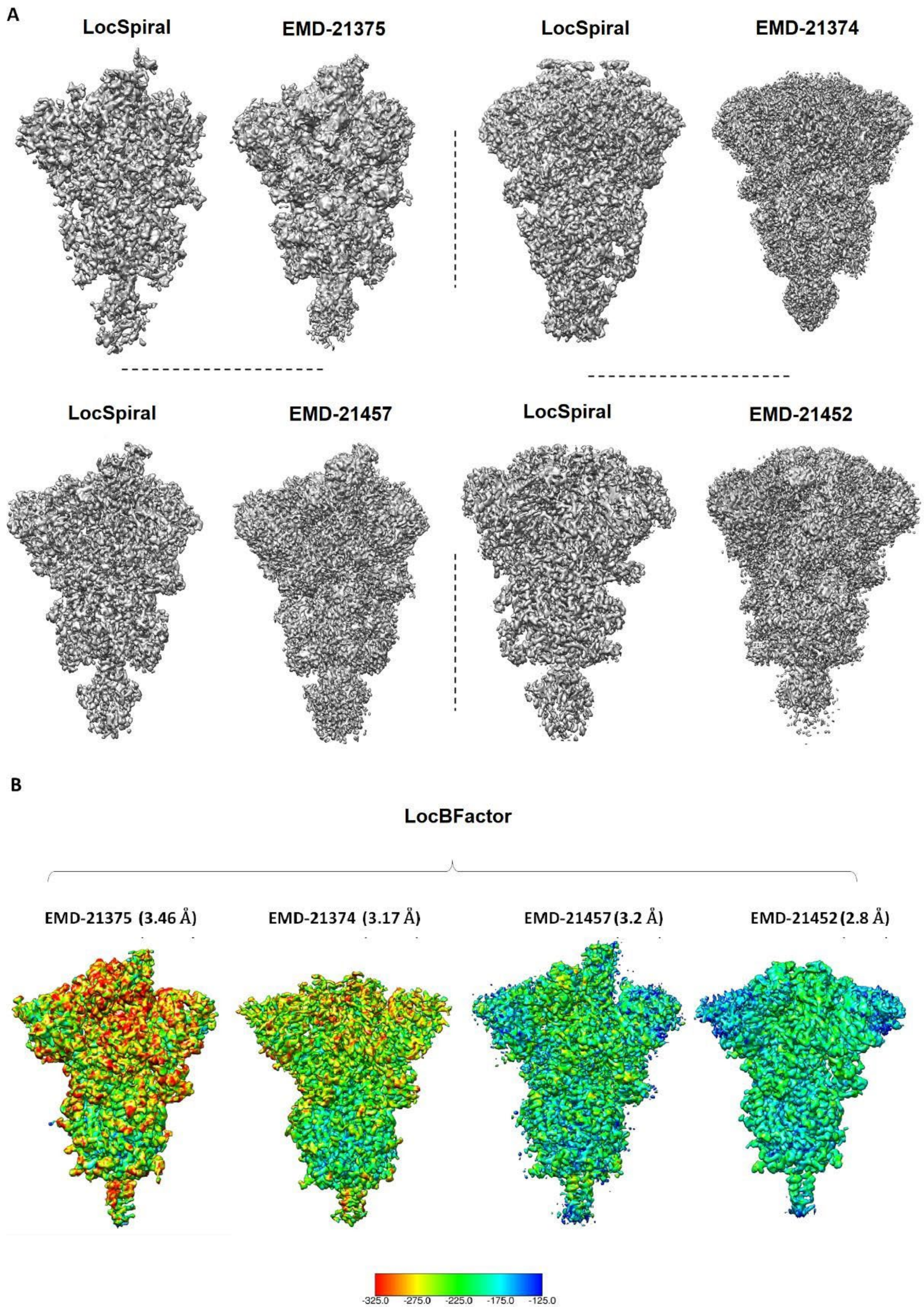
Supplementary Figure S2.3 Complete and superimposed sharpened maps. *LocSpiral (gray colour) and Relion (red colour) for EMD-9865 with the corresponding atomic structure (PDB 6v21). In the black rectangles are shown zoomed views of the regions labelled with the same index.*



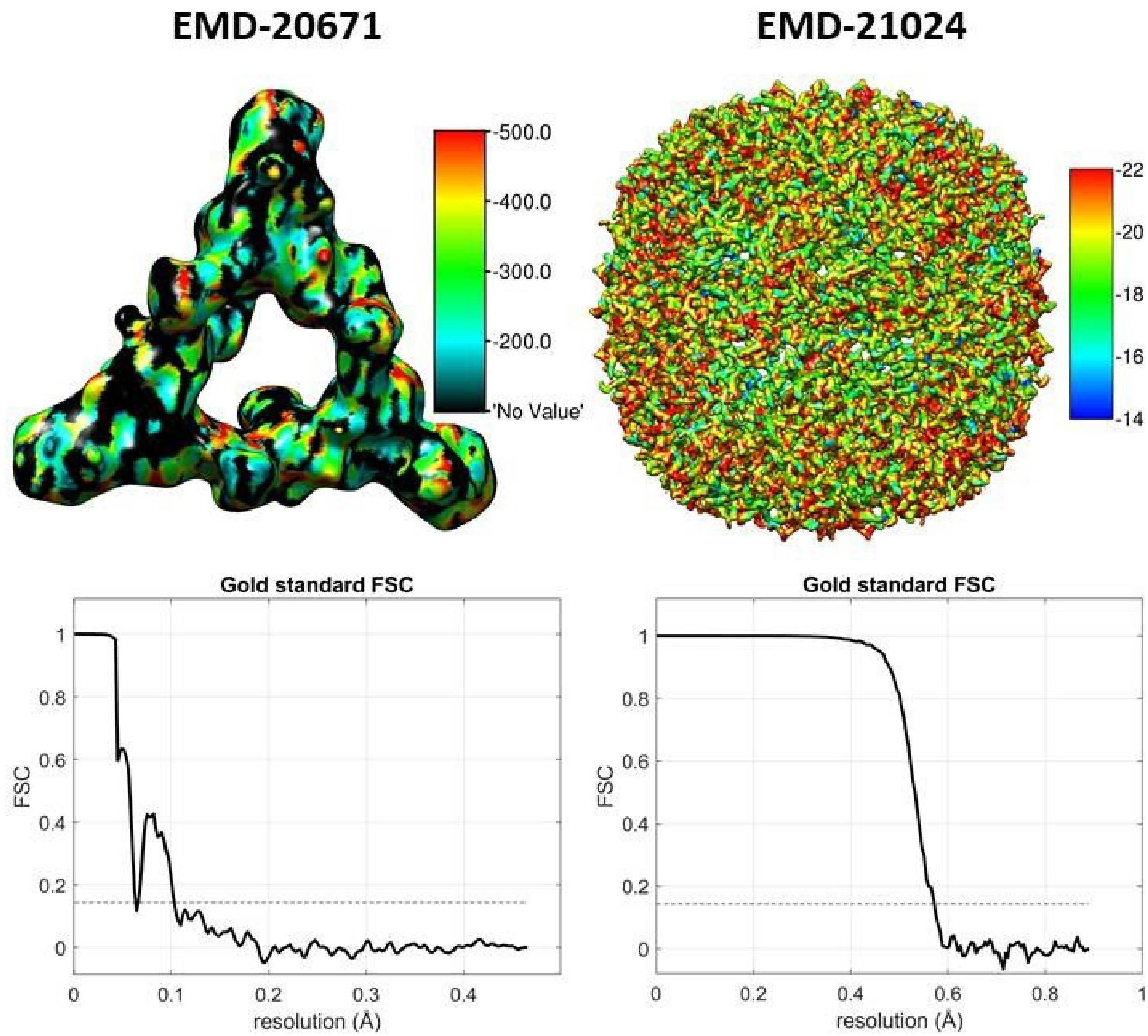
Supplementary Figure S2.4 Results and comparisons between different methods for the EMD-8441 immature ribosome. *(Left) Obtained sharpened maps for EMD-8441 by LocSpiral and Relion. These maps are shown at low and high thresholds. (Right) Obtained B-maps (local B-factor maps corresponding to the slope of the local Guinier plots) and A-maps (local values of the logarithm of structure factor amplitudes at 15 Å) of EMD-8441 at different orientations. In these figures, noise B-factors (B-factors obtained from amplitudes below the noise level for the used resolution range) are filtered out and appear with black color.*

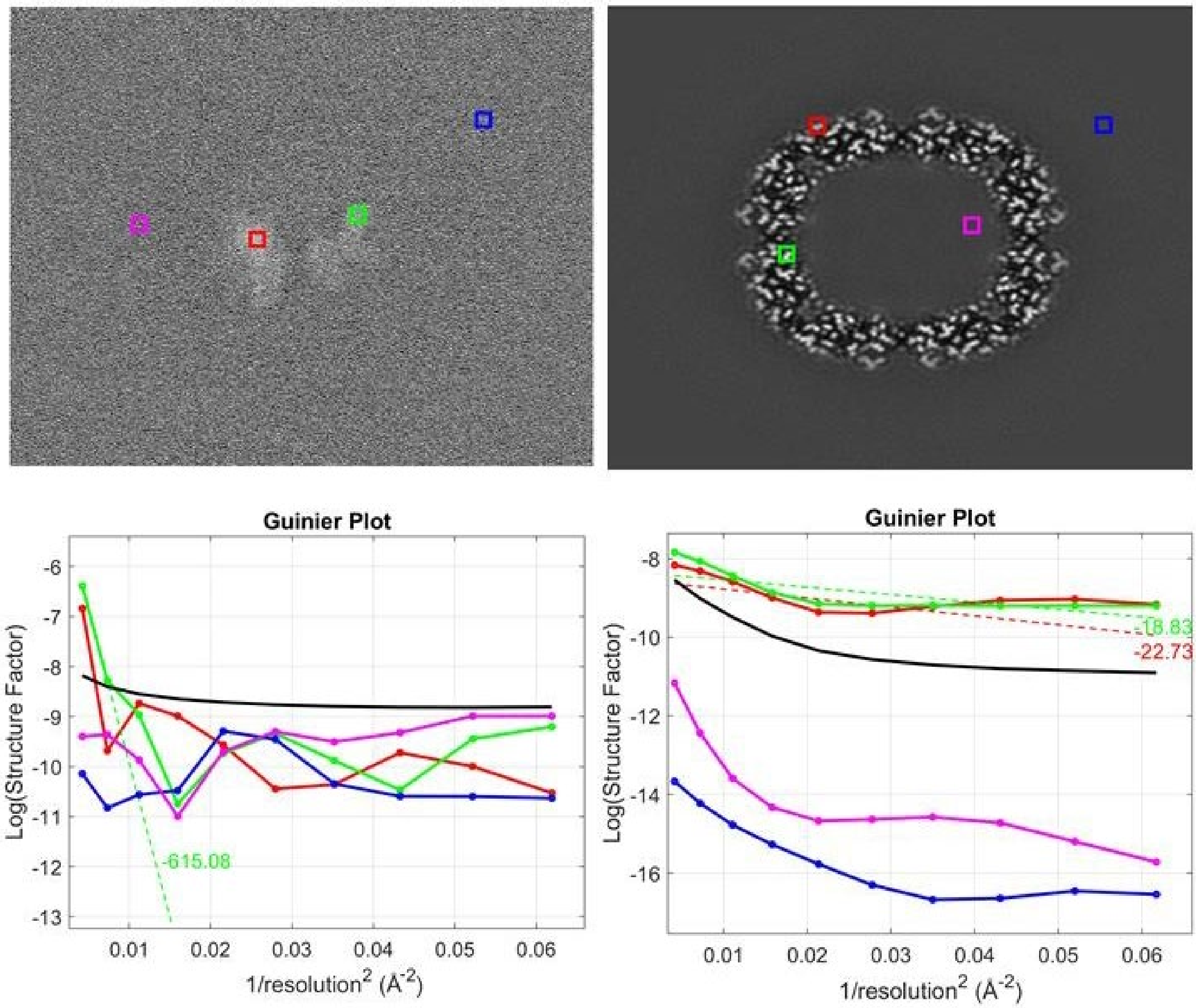


Supplementary Figure S2.5 Improved maps obtained by LocSpiral from EMD-21375, EMD21457, EMD-21452 and corresponding fitted atomic models.


A
LocSpiral
EMD-21375
LocSpiral
EMD-21374
LocSpiral
EMD-21457
LocSpiral
EMD-21452
B
LocBFactor
EMD-21375 (3.46 Å)
EMD-21374 (3.17 Å)
EMD-21457 (3.2 Å)
EMD-21452 (2.8 Å)
-325.0
-275.0
-225.0
-175.0
-125.0

Supplementary Figure S2.6 Results obtained by LocSpiral, LocBFactor for SARS-CoV-2 samples. *A) Maps obtained by LocSpiral approach (left) compared with maps as deposited in EMDB with accessing codes (EMD-21375, EMD-21374, EMD-21457, EMD-21452). B) Obtained B-factor maps to be used for sharpening (slope of the local Guinier plot multiplied by 4) by LocBFactor approach for EMD-21375, EMD-21374, EMD-21457, EMD-21452.*





Supplementary Figure S2.7 Obtained B-factor maps (slope of the local Guinier plot) by LocBFactor approach. *EMD-20671 and EMD-21024, corresponding FSC curves and Guinier plots of macromolecule and noise/background points indicated with coloured points shown in corresponding central slices.*

2.6.5 Supplementary table

		TRP channel (EMD-10418) (PDB 6t9n)	Apoferritin (EMD-9865) (PDB 6v21)	SARS-CoV-2 (EMD-21375) (PDB 6vsb)
EMRINGER	**EMRINGER LocSpiral**	2.31	8.63	2.31
	EMRINGER Relion	2.36	2.51	2.27
	Rotamer-ratio LocSpiral	0.70	0.97	0.70
	Rotamer-ratio Relion	0.72	0.67	0.73
	Max Z-score LocSpiral	7.93	48.83	9.47
	Max Z-score Relion	8.11	14.22	9.06
	Model Length LocSpiral	1184	3200	1683
	Model Length Relion	1184	3200	1598
MOLPROBITY	All-atom Clashscore LocSpiral	6.44	5.90	13.66
	All-atom Clashscore Relion	6.12	5.27	14.34
	Ramachandran Plot LocSpiral	Outliers:0.00% Allowed:4.38% Favored:95.62%	Outliers:0.00% Allowed:1.90% Favored:98.10%	Outliers:0.00% Allowed:8.36% Favored:91.46%
	Ramachandran Plot Relion	Outliers:0.00% Allowed:3.12% Favored:96.88%	Outliers:0.00% Allowed:2.33% Favored:97.67%	Outliers:0.00% Allowed:8.54% Favored:91.64%
	Rotamer Outliers LocSpiral	7.24 %	1.39 %	13.63 %
	Rotamer Outliers Relion	2.77 %	1.49 %	9.96 %
	Cbeta Deviations LocSpiral	0.00 %	0.00 %	0.00 %
	Cbeta Deviations Relion	0.00 %	0.00 %	0.00 %
	Peptide Plane LocSpiral	Cis-proline:0% Cis-general:0% Twisted Proline:0% Twisted-General: 0%	Cis-proline:25% Cis-general:0% Twisted Proline:0% Twisted-General:0%	Cis-proline:0.67% Cis-general:0% Twisted Proline:0.67% TwistedGeneral:0.03%
	Peptide Plane Relion	Cis-proline:0% Cis-general:0% Twisted Proline:0% Twisted-General: 0%	Cis-proline:25% Cis-general:0% Twisted Proline:0% Twisted- General:0%	Cis-proline:0% Cis-general:0% Twisted Proline:0% Twisted-General: 0%

Supplementary Table S2.1 EMRINGER and Molprobity modeling scores. *Obtained between sharpened maps by Relion postprocessing and LocSpiral, and corresponding atomic models after refining the structure against corresponding maps by Phenix real_space_refine approach using 5 refining iterations*

2.6.6 Supplementary references

1. Afonine, P. V. *et al.* New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr D Struct Biol* **74**, 814-840, (2018).
2. Barad, B. A. *et al.* EMRinger: side chain-directed model and map validation for 3D cryoelectron microscopy. *Nat Methods* **12**, 943-946, (2015).
3. Wang, Q. *et al.* Lipid Interactions of a Ciliary Membrane TRP Channel: Simulation and Structural Studies of Polycystin-2. *Structure* **28**, 169-184, (2020).
4. Fernandez, J. J., Luque, D., Caston, J. R. & Carrascosa, J. L. Sharpening high resolution information in single particle electron cryomicroscopy. *J Struct Biol* **164**, 170-175, (2008).
5. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, (2016).
6. Davis, J. H. *et al.* Modular Assembly of the Bacterial Large Ribosomal Subunit. *Cell* **167**, 1610-1622 (2016).

Connecting text: Chapter 2 to 3

In the previous chapter, we demonstrated that cryo-EM maps with heterogenous SNR can be processed using local improvement methods, which present better map connectivity while enhancing the resolution of the resultant maps. The goal is to improve the interpretability and visualization of the resultant cryo-EM maps. However, when macromolecules undergo large conformational changes, high resolution structures are dependent on 3D classification methods within the image processing workflow. These will determine the number of output conformations showing the heterogeneity of the input dataset, and reveal fine structural details without the artefactual blurring of multiple superimposed conformations. Therefore, the next chapter is dedicated to the approaches to process dynamic macromolecules while exposing their various conformation states.

CHAPTER 3

HIERARCHICAL AUTOCLASSIFICATION OF CRYO-EM SAMPLES AND MACROMOLECULAR ENERGY LANDSCAPE DETERMINATION

3.1 Abstract

3.1.1 Background and objective

Cryo-electron microscopy using single particle analysis is a powerful technique for obtaining 3D reconstructions of macromolecules in near native conditions. One of its major advances is its capacity to reveal conformations of dynamic molecular complexes. Most popular and successful current approaches to analyzing heterogeneous complexes are founded on Bayesian inference. However, these 3D classification methods require the tuning of specific parameters by the user and the use of complicated 3D re-classification procedures for samples affected by extensive heterogeneity. Thus, the success of these approaches highly depends on the user experience. We introduce a robust approach to identify many different conformations presented in a cryo-EM dataset based on Bayesian inference through Relion classification methods that does not require tuning of parameters and reclassification strategies.

3.1.2 Methods

The algorithm allows both 2D and 3D classification and is based on a hierarchical clustering approach that runs automatically without requiring typical inputs, such as the number of conformations present in the dataset or the required classification iterations. This approach is applied to robustly determine the energy landscapes of macromolecules.

3.1.3 Results

We tested the performance of the methods proposed here using four different datasets, comprising structurally homogeneous and highly heterogeneous cases. In all cases, the approach provided excellent results. The routines are publicly available as part of the CryoMethods plugin included in the Scipion package.

3.1.4 Conclusions

Our results show that the proposed method can be used to align and classify homogeneous and heterogeneous datasets without requiring previous alignment information or any prior knowledge about the number of co-existing conformations. The approach can be used for both 2D and 3D autoclassification and only requires an initial volume. In addition, the approach is robust to the "attractor" problem providing many different conformations/views for samples affected by extensive heterogeneity. The obtained 3D classes can render high resolution 3D structures, while the obtained energy landscapes can be used to determine structural trajectories.

3.2 Introduction

Cryo-electron microscopy (cryo-EM) using single particle analysis is a powerful technique for obtaining high-resolution three-dimensional (3D) reconstructions of macromolecular structures in a close-to-native state[1–3]. This structural information is essential to infer the biological processes driven by macromolecular machines. Recently, this approach has broken the atomic resolution barrier for the apoferritin sample thanks to improved hardware and software tools[4,5]. Moreover, this technique holds the promise of revealing the complete conformational variability of dynamic macromolecular complexes at equilibrium.

The most popular and successful approaches to analyze heterogeneous complexes are currently founded on Maximum-likelihood or Bayesian inference[6–9]. These methods simultaneously refine multiple 3D reconstructions by marginalizing particle image likelihoods over missing information such as particle's orientation and conformation, and their robustness has been demonstrated by numerous groups, such as[10–12]. Moreover, the major advances of Bayesian methods are highlighted by their capacity to infer the particle's orientation and conformation simultaneously, and their proven robustness to challenging cases, including particle images

showing low signal-to-noise-ratios (SNRs) and/or low contrast. Conversely, these approaches require information from users, i.e., the number of classification iterations and the number of conformations that are coexisting in the dataset (number of classes). Additionally, they are usually affected by the "attractor problem"[13,14], the phenomenon where particles correlation is bias towards classes containing more particles. Thus, Maximum-likelihood/Bayesian based methods may tend to classify single particles according to the SNRs of the different classes rather than by the actual conformation of the particles. This issue limits the capacity of these approaches to 1) provide many different 3D classes showing different macromolecular conformations (they are usually limited to deliver between 3-5 different reconstructions); 2) capture minoritarian conformations; and 3) provide homogeneous classes that may produce high-resolution reconstructions. Several methods have been proposed to alleviate these issues. Examples of these approaches include particle "pruning" methods to filter incorrectly 3D classified single particles[15–20] and robust methods for 2D classification, as in CL2D[21] or GTM clustering[22]. Unfortunately, extensions of these automatic 2D classification approaches to 3D are still pending. Thus, users pursuing to find an unknown number of (many) different conformations from the particle set are required to perform difficult and tedious reclassification procedures, which require deep expertise on the technique and sample[23–26].

More recently, approaches have been proposed to process macromolecular assemblies showing continuous flexibility. For example, Relion models these continuous changes in the macromolecular structure by means of a number of user-defined rigid bodies that can change their relative pose[27]. In HEMNMA, macromolecular conformational changes are modeled using a set of user-defined normal mode-based deformations[28]. Continuous heterogeneity may be also described using linear approaches as principal components analysis[29–32], however, these methods may show artifacts when large conformational deformations are poorly approximated by linear approximations along a volume basis[33]. Deep learning approaches based on generative networks were also proposed to capture the continuous motions of flexible macromolecules[34,35]. These approaches have shown their capacity to resolve continuous and discrete heterogeneity for large complexes. However, the robustness of these methods to process smaller and lower contrast samples has not yet been demonstrated. Moreover, these approaches require as input the orientation of the single particles, a high-resolution consensus reconstruction, a large dataset to train the network, and input parameters including the dimension of the latent space or the number of epochs to train the network.

Methods that can provide many different conformations of a macromolecule are being used to generate energy landscapes of the sample[36–38]. Energy landscapes allow to describe the macromolecular motion in a comprehensive and quantifiable manner, determining, for example, sorted conformational trajectories along different states. In[39,40], the authors determine many different conformations by extensive manual 3D classifications with Relion to map the conformational landscape into an energy landscape. This mapping is done by use of principal component analysis and the Boltzmann distribution that establishes a link between the energy of a system, its temperature, and its population. This method requires the user to determine the two PCA eigenvectors corresponding to the reaction coordinates to be analyzed. Note, however, that reaction coordinates (and structural variability) may not be directly factorizable by PCA decomposition using merely two principal components. Additionally, as explained above, PCA decomposition may show artifacts when large conformational deformations are poorly approximated by linear approximations along a volume basis. In[41] authors empirically show that latent encoding of particles in their software, CryoDRGN, is done such that structurally related particles are in proximity. Thus, the particle's latent space may be interpreted as a pseudo-conformational landscape. However, as the authors pointed out, this interpretation is not mathematically guaranteed, as the objective function aims to reproduce the distribution of structures only.

In this work, we present a robust approach to obtain many different conformations from a heterogeneous dataset without requiring previous alignment information or any prior knowledge about the number of co-existing conformations. The approach is based on Bayesian inference classification but works hierarchically. Thus, for each classification step, the data is classified into a small number of classes, typically two, using Relion as our classification engine[42]. The proposed method automatically determines the number of required classification iterations to be performed for each classification step. Moreover, the approach uses automated stop conditions to provide as many 3D classes as possible, while avoiding unnecessary classifications. Consequently, for each intermediate class, the approach automatically determines whether this data should be reclassified again or not, continuing this workflow iteratively. Once this tree-based processing is finished, the obtained structures can be used to build conformational and energy landscapes. To this end, we employ Isomap[43], which is a nonlinear dimensionality reduction method to compute a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. Other non-linear dimensionality reduction methods such as LTSA[44], Hessian LLE[45] or t-SNE[46] may be used as well. Finally, the obtained raw 3D classes can be clustered into final 3D classes where similar conformations are grouped together automatically by the use of the affinity propagation algorithm[47].

Our results show that the proposed method can be used to align and classify homogeneous and heterogeneous datasets. The approach can be used for both 2D and 3D autoclassification and only requires an initial volume. In addition, the approach is robust to the "attractor" problem providing many different conformations/views. The obtained 3D classes can render high resolution 3D structures, while the obtained energy landscapes can be used to determine structural trajectories.

3.3 Methods

In this work, we describe two main approaches. First, we present robust 2D and 3D autoclassification methods. Second, we determine energy landscapes from the output of the proposed hierarchical 3D autoclassification method.

3.3.1 Hierarchical 3D autoclassification

We present a robust and automated 3D classification approach that aims to determine many different conformations presented in the dataset without compromising particle alignment quality and resolution of the captured conformers. We term this approach as 3D autoclassification, and it consists of the following processes: particle alignability estimation; hierarchical 3D classification and 3D clustering. In Figure 3.1, we show a scheme of the proposed approach.

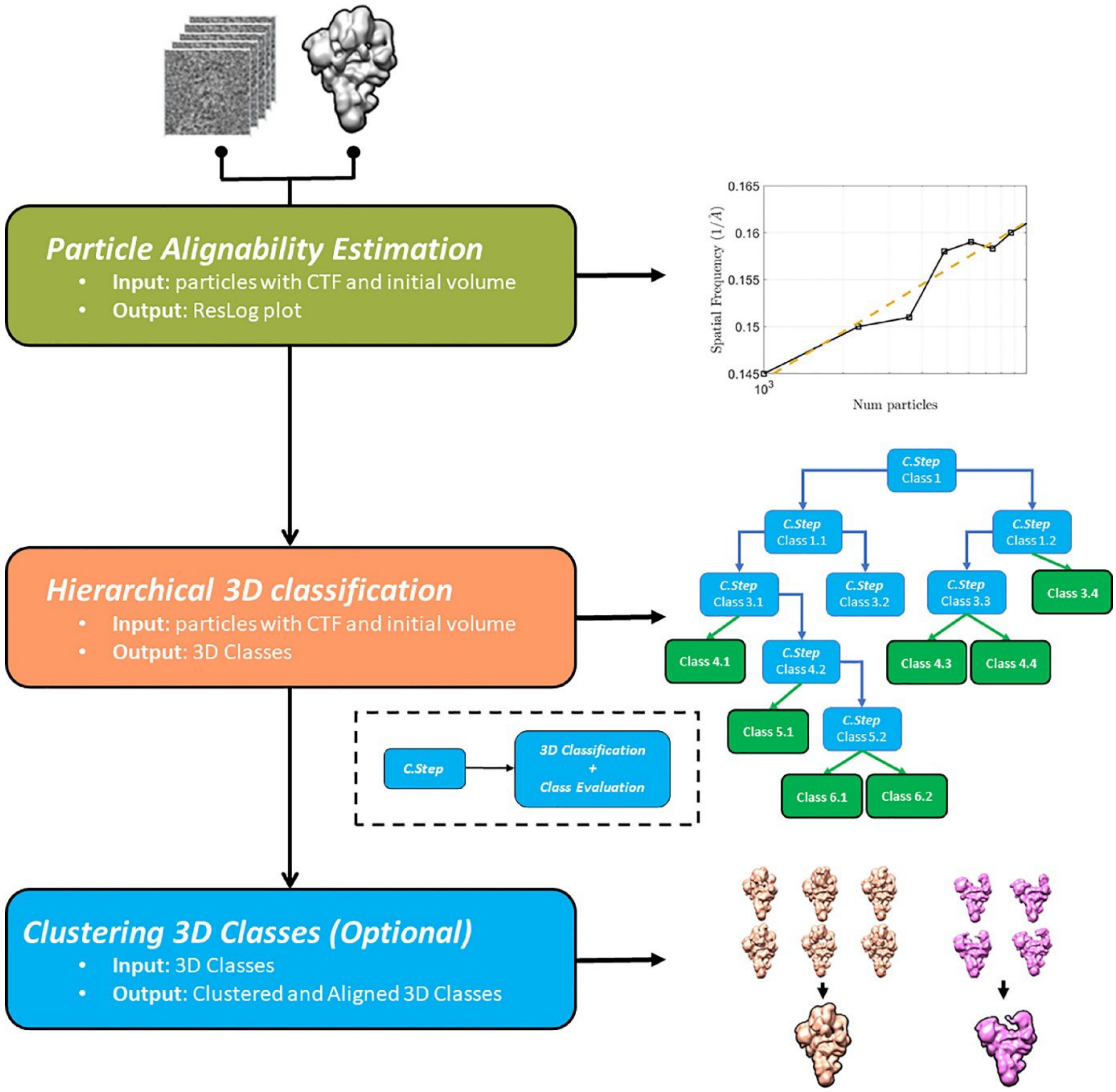


Figure 3.1 Scheme of the proposed 3D autoclassification approach. *From top to bottom: first, the particle alignability is estimated for different subsets of particles by a ResLog plot (plot of inverse resolution vs. the logarithm of the number of particles). Then, 3D classifications and evaluation steps are performed in a hierarchical manner. When a 3D classification is finished, the resulting classes are evaluated to determine if further reclassifications are required. Once all 3D classifications are finished, the raw 3D classes obtained are aligned and clustered to produce final 3D classes.*

3.3.2 Particle alignability estimation

The approach first computes a ResLog plot (plot of inverse resolution vs. the logarithm of the number of particles)[48,49] by randomly taking ten sets of particles between 10,000 and 100,000 images. The plot is obtained using the same initial model in all cases and performing 3D classifications in one class with Relion to align particle images only. The ResLog plot is a quality indicator of how well the particles are aligned and classified in a 3D reconstruction. The hierarchical 3D classification step uses this ResLog plot as a baseline in posterior classification actions to determine if it is convenient to continue reclassifying particle subsets or not. Note that as we perform particle alignment and classification in each classification step, the alignment accuracy may be compromized depending on the number of particles to be processed. Thus, alignment/classification quality tests after each classification are performed using the ResLog linear relation between the inverse of the obtained map resolution and the logarithm of number of particles, so further reclassification of particle images is carried out only if the resolution obtained is above the ResLog baseline. In cases where the number of particles is less than 100,000, the ResLog plot is estimated using 10 random sets between 10% and 90% of the images.

3.3.3 Hierarchical 3D autoclassification

The dataset is then 3D classified using a hierarchical approach. Our method uses Relion 3D classification[50] as its classification engine. In each classification step, the particles are divided into a small number of classes, typically 2 or 3 at most, to avoid the "attractor problem"[51]. After each classification, the method does an evaluation step to assess if the class may be further classified or not. The evaluation uses a two-fold criterion: (1) the class number of particles should be higher than a threshold provided by the user (typically 5000–10,000 particles) and (2) the class resolution provided by the classification method should be higher than the value provided by the ResLog plot with this number of particles. If both conditions are met, then the class is further classified.

All classification steps run automatically without requiring any additional input from the user. The classification iterations are determined on-the-fly by the following approach. In each classification step, ten classification iterations are initially performed. Then, the average of the maximum Bayesian probabilities (P_{max}) between particles and classes, as outputted by Relion 3D classification method with respect to possible particle´s orientation and conformation, are analyzed for the last six iterations. Note that P_{max} refers to the average value over all particles of the maxima of their computed probability distributions. If P_{max} values show a plateau curve along these last iterations, it means that the classification n step converged, and the process is automatically finished. If not, five more classification iterations are included and checked again for convergence. In case a classification step does not converge after 75 iterations, the process is automatically stopped.

3.3.4 3D clustering (optional)

The aim of this step is to merge obtained 3D classes that represent similar conformations. Note that the proposed Hierarchical 3D classification algorithm does not guarantee that the classes obtained will be structurally distinct, so they may represent similar conformations. Indeed, conformations that have a deep well in the energy landscape are likely to have multiple near-identical class averages.

Each obtained 3D class has a class representative, which is the volume or map mimicking the macromolecular conformation captured by this class. The first step of the 3D clustering process consists of aligning all obtained class representatives or maps. This alignment procedure is done effectively by spherical harmonics decomposition[52]. After all volumes are aligned, the approach computes truncated principal component analysis (*t*PCA) capturing 90% of the observed variance of the maps. Each map is then projected into the resulting *t*PCA base, obtaining a feature vector. These vectors are used to determine structural differences between maps by calculating the Euclidean distance (Euclidean norm) between them. Then, the Affinity propagation algorithm[53] is used to generate 3D clusters. Affinity propagation is based on the concept of "message passing" between data points, and its main advantage is that it clusters data points automatically, determining the number of clusters or classes presented in the dataset.

3.3.5 Hierarchical 2D autoclassification

This approach focusses on obtaining as many different 2D classes as possible and is based on the same hierarchical classification approach outlined before in Section 2.1.2. In this case, the method uses Relion 2D classification[54] as its classification engine, and hierarchically and iteratively divides the dataset into a small number of 2D classes at each classification step, typically between 2 and 4. After each classification, the method automatically evaluates if the resultant classes may be further classified according to the number of particles alone. Typically, further classifications are automatically stopped, when the number of particles belonging to a given class is below 200.

3.3.6 Energy landscape determination

Once the dataset is processed by the Hierarchical 3D autoclassification approach presented above, a large number of 3D classes are obtained. These 3D classes can be used to build an energy (or conformation) landscape of the sample. To this end, first, the different maps obtained representing different confirmations are realigned and encoded by means of feature vectors using the same approach outlined in Section 2.1.3. As a summary, this process consists of (1) fast alignment of maps by spherical harmonics decomposition[52], and (2) map dimensionality reduction by truncated principal component analysis (tPCA), where 90% of the observed variance is captured. The captured PCA basis represents orthogonal conformational coordinates controlling collective conformational changes. The projection of maps onto this base (feature vectors) constitutes tPCA coefficients determining the conformational state of each map. Note that these same feature vectors could be used to directly build energy (or conformation) landscapes as previously proposed by Haselbach et al.[55,56]. These works assume thermal equilibrium to connect energy, population, and conformation via the Boltzmann relation as

$$n_k \propto e^{-\frac{G_k}{k_B T}} \tag{3.1}$$

Where, G_k and n_k represent the Gibbs free energy and population at conformation k, k_B is the Boltzmann constant and T is the temperature. For a large ensemble of particles in equilibrium, Eq. (3.1) can be used to obtain differences in the Gibbs free energy (ΔG) between states with populations n_1 and n_2 as

$$\Delta G = -k_B T \log\left(\frac{n_1}{n_2}\right) \tag{3.2}$$

Note that the structural information of each conformation can be approximated by the first two PCA coefficients, while variations in the Gibbs free energy (ΔG) between conformations may be computed from equation 3.2 using respective number of particles at each conformation (or 3D class). In Eq. (3.2), the conformation (3D class) attracting the highest number of particles is usually used as the reference (n_2). Thus, a 2D energy landscape may be computed by respective 2D PCA coordinates and obtained variations in the Gibbs free energy.

This approach assumes that conformational heterogeneity shown by particles in the dataset can be well-approximated by linear approximations along a volume PCA basis and using two PCA coefficients only. However, samples exhibiting large conformational deformations may be poorly represented by this method, thus, showing artifacts in the resulting energy landscapes[57]. To overcome this limitation, we project the n-dimensional feature vectors (capturing 90% of the variance) into a 2D space using a non-linear dimensionality reduction approach. One possibility to realize this mapping is to maintain the geodesic distances between pairs of feature vectors in the reduced 2D space. This isometric mapping can be efficiently computed by the Isomap algorithm[58]. Other non-linear approaches that can be used by our method are Local Tangent Space Alignment (LTSA)[59], Hessian Eigenmapping[60] and t-distributed stochastic neighbor embedding (t-SNE)[61].

3.4 Results

Four experimental datasets were used to test the proposed autoclassification methods. These datasets, which are publicly available from the EMPIAR database[62], are composed of cryo-EM single particles of (1) L17-depleted 50S ribosomal assembly intermediates (EMPIAR-10,076); (2) spliceosomal B-complex from yeast[63],(EMPIAR-10,180); (3) beta-galactosidase in complex with a cell-permeant inhibitor[64], (EMPIAR-10,061); and (4) AP-1:Artf1:tetherin-HIV-Nef complex[65] (EMPIAR-10,178). This last dataset is used to analyze our hierarchical 2D autoclassification approach only.

The selected datasets comprise two different kinds of samples. On the one hand, the beta-galactosidase shows low structural heterogeneity, while immature ribosomes and the spliceosome show extensive heterogeneity. The goal of our methods is to obtain many different conformations (or views for the 2D classifications) allowing us to capture underlying low populated classes. This information may be used to build energy landscapes, trace potential structural trajectories of the samples, and to detect and remove false positive particles (particle screening).

3.4.1 L17-depleted 50S ribosomal intermediates

This dataset consists of 131,899 particle images of immature *E. coli* large ribosomal subunits[66] and is publicly available from EMPIAR database (EMPIAR-10,076). The particles have a pixel size of 1.31 Å/px and a window size of 320 px. The dataset was collected on a Titan Krios using a K2 direct electron detector (Gatan). This dataset shows massive heterogeneity, and the authors of the original publication[66] were able to identify 15 different conformations by extensive 3D classifications and reclassifications with Relion and deep expertise with the sample.

We processed this dataset with our proposed 3D autoclassification approach producing 65 raw 3D classes by hierarchical 3D autoclassification, which were automatically clustered into 18 final 3D classes. In each classification step, the particles were divided into 2 classes. The 'number of particles' threshold, used to stop the automatic reclassifications, was set to 5000 particles to produce a large number of 3D classes. Figure 3.2A and D show the obtained final (18) and raw (65) 3D classes, respectively. Note that in Figure 3.2, we have labeled each 3D class with a letter ('A','B','C','D','E') by visual identification with the original labelling done in the original work of Davis et al.[66]. Additionally, although our automatic method can provide more different conformations (18) than the manual approach followed in[66], the 3D autoclassification method could not provide the 30S ribosomal subunit as a separate class (Class F in[66]). We believe that the 30S subunit was not distinguished as a separate class because the used initial volume. Note that each 3D classification step uses as initial volume the previously obtained 3D reconstruction (or the input one in the first iteration). Thus, if the employed initial volume is very different from the target 3D class, the particles could not be aligned and classified properly. For the sake of comparison, we include here the resolution reported in the original work[66] for the different 3D classes: Class A 6.5 Å; Class B 4.5 Å; Class C1 4.5 Å; Class C2 4.6 Å; Class C3 4.0 Å; Class D1 4.7 Å; Class D2 4.9 Å; Class D3 4.6 Å; Class D4 4.7 Å; Class E1 4.2 Å; Class E2 4.5 Å; Class E3

5.0 Å; Class E4 4.5 Å; Class F 7.9 Å. Note that authors in[66] employed a different reconstruction workflow, using Frealign instead of Relion for refinement of particle alignments and map postprocessing. Additionally, note that to accurately compare our results with the ones reported in[66], we would have to use the same solvent masks when postprocessing the maps. Thus, these comparisons should be used only qualitatively to show that most of the obtained 3D classes are similar in terms of resolution. Moreover, in Figure 3.2B, we show the obtained Gold-standard FSCs computed after refining the different final 3D classes with Relion autorefine. As it can be seen from this figure, the resolution of 3D reconstructions ranges between 4.2 and 6.9 Å. Finally, in Figure 3.2C, we show the energy landscape in $k_B T$ units, computed from the raw 3D classes. Note that this energy landscape is similar to previous landscapes obtained for this sample by other approaches[57]. Figure 3.2C also shows a structural trajectory using the computed energy landscape and the corresponding sorted structures along this trajectory.

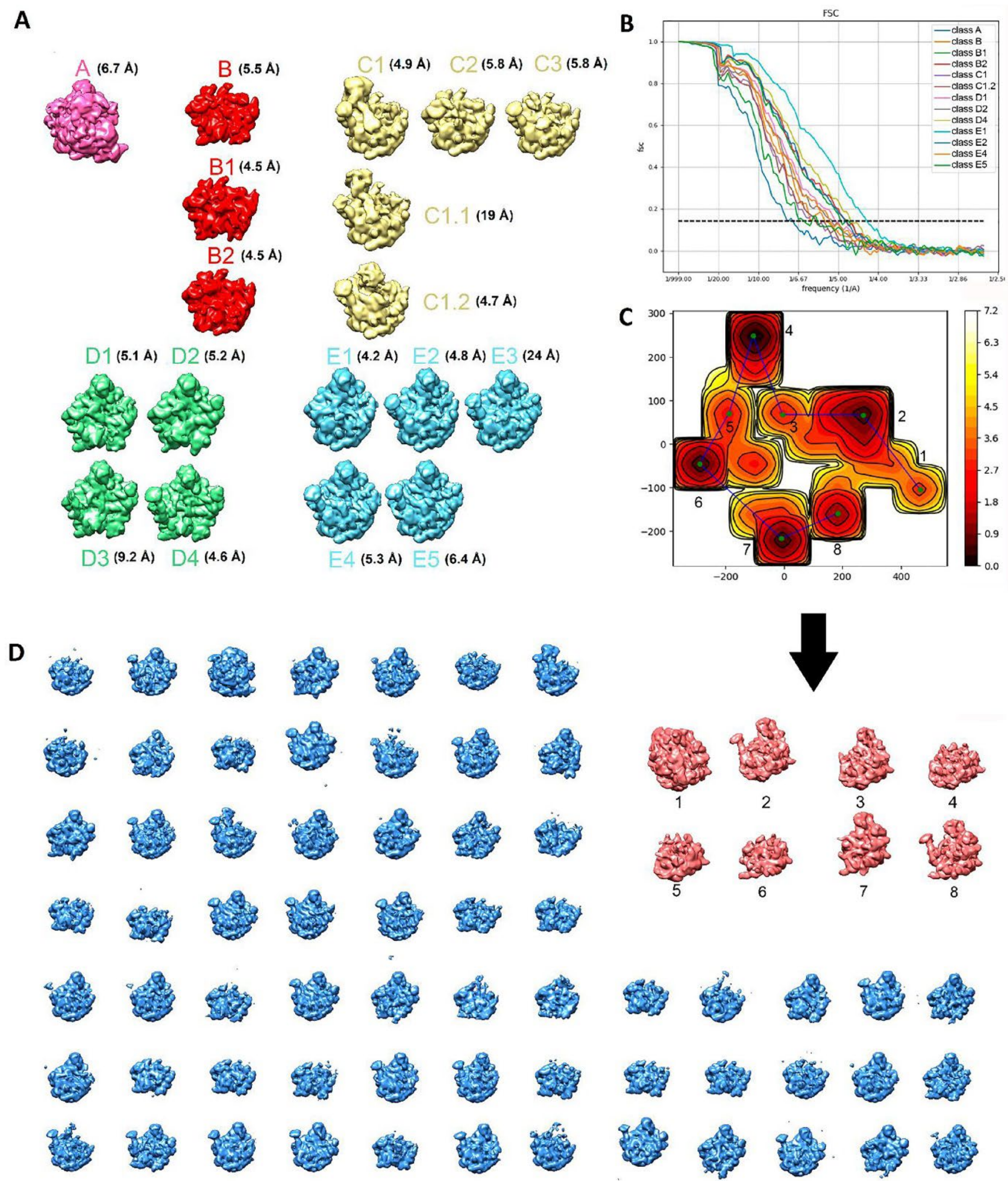


Figure 3.2 3D autoclassification and energy landscape results for L17-depleted 50S ribosomal intermediates. *A) final 3D classes obtained after 3D clustering of raw 3D classes (shown in Figure 3.2D) with the obtained resolution after refining final 3D classes with Relion. The classes are labelled and colored according to their similarity by visual inspection with classes presented in the original publication. B) Gold-standard FSCs determined after refining final 3D classes with Relion autorefine. C) Energy landscape in* $k_B T$ *units computed from raw 3D classes. D) Raw 3D classes obtained in the Hierarchical 3D autoclassification step.*

3.4.2 Spliceosomal B-complex dataset

This dataset is composed of 327,490 particle images of a spliceosomal B-complex from yeast[63] and is deposited and publicly available from EMPIAR database (EMPIAR-10,180). The deposited particle images correspond to "shiny" particles after Relion particle polishing that were downsampled to 1.699 Å/px and windowed to 320 px. The dataset was collected on a Titan Krios using a K2 direct electron detector (Gatan).

This particle set was processed using the proposed 3D autoclassification approach producing 102 raw 3D classes that were automatically clustered into 14 final classes. In each classification step, the particles were divided into 2 classes and the 'number of particles' threshold to stop the automatic reclassifications was set to 5000 particles to produce a large number of 3D classes. Figure 3.3A shows the obtained final classes, while in Figure 3.3B we show the raw 3D classes for classes 1, 2, 3, 7, 9, 13 and 14, which are labeled with different colors. In Figure 3.3C, we show the obtained energy landscape for the spliceosomal B-complex, computed using the 102 raw 3D classes previously obtained by the 3D autoclassification approach. Finally, Figure 3.3D shows obtained sorted conformations along the structural trajectory displayed with a blue line in the energy landscape in Figure 3.3C.

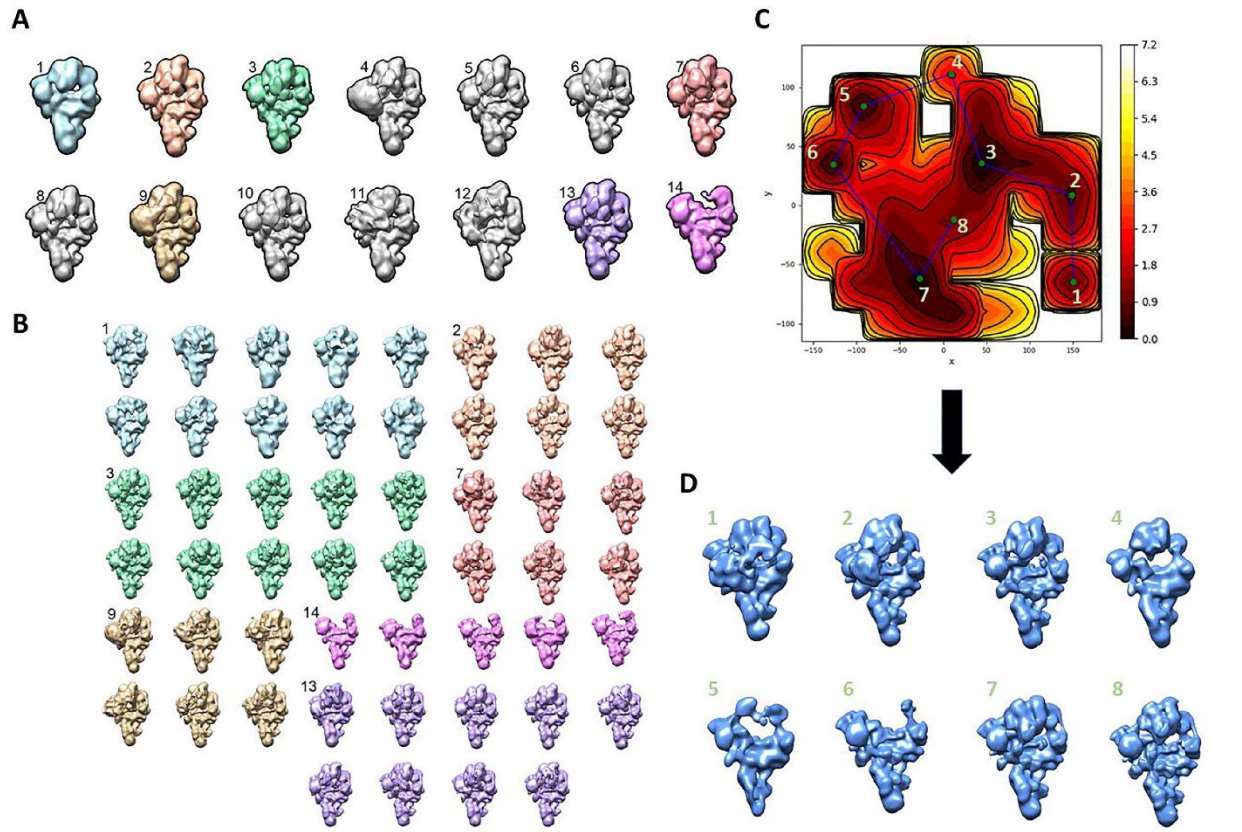


***Figure 3.3 3D autoclassification and energy landscape results for the spliceosomal B-complex**. A) Final 3D classes labelled with class numbers after clustering of obtained raw 3D classes. B) Subset of raw 3D classes obtained, labelled by colours for class numbers C) Calculated energy landscape in $k_B T$ units from computed raw 3D classes with a defined structural trajectory marked with blue lines. D) Corresponding conformations for the structural trajectory shown in C).*

3.4.3 Beta-galactosidase in complex with a cell-permeant inhibitor dataset

We use beta-galactosidase single particles to show the capacity of the method to process structurally homogeneous datasets. The dataset consists of 1539 micrographs with a pixel size of 0.32 Å/px collected on a Titan Krios microscope using a K2 direct electron detector (Gatan). This dataset is publicly available from EMPIAR database (EMPIAR-10,061). The CTF parameters were estimated using CTFFIND4[67] and we use the same 2D coordinates employed in the original work[64] for particle extraction, which are available from EMPIAR. The extracted particles were downsampled four times and classified with Relion to obtain 2D class averages with improved SNRs. These 2D class averages were used then to obtain 8 *de novo* initial maps by the RANSAC approach[68]. The volume selector method[69] processed these maps, using subsets of 5000 particles randomly selected, providing a single initial volume suitable to be used by our 3D autoclassification approach as reference map.

The particle set was then processed using the proposed 3D autoclassification approach producing four raw 3D classes after the hierarchical 3D classification step, which were clustered into two final 3D classes by the 3D clustering step. In each classification step, the particles were divided into 2 classes and the 'number of particles' threshold was set to 10,000 particles. In Figure 3.4, we show the obtained intermediate results by the 3D autoclassification approach and the final 3D classes. As can be seen from Figure 3.4A, the output final classes consist of one major 3D class (Class 7), which accumulates 37,154 particles (90% of the dataset) and a minor one (Class 8) containing 3958 particles. This minor class may be composed by a large percentage of low-quality particles, as the refinement of the complete dataset with Relion (41,122 particles) produces slightly lower quality results than the reconstruction obtained with Class 7 particles only (37,154 particles) (see corresponding FSC curves in Figure 3.4C). In Figure 3.4B, we show the refined 3D map produced by Class A after sharpening with Relion. Finally, we show in Figure 3.4D and E obtained 2D class averages by the 2D autoclassification approach and by Relion 2D classification method respectively. As can be seen from these figures, our approach is able to produce a larger number of meaningful 2D class averages as it is less affected by the attractor problem. Finally, we have removed particles belonging to 2D classes capturing a low number of particles after 2D autoclassification to show the capacity of the autoclassification method to improve the homogeneity of cryo-EM datasets. This results into a dataset composed of 36,879 particles, thus, we removed 4243 particles (~10% of the dataset). This dataset produces a reconstruction at 2.81 Å resolution that was automatically sharpened by Relion with an apply B-factor of -72.5 Å^2. Note that the original dataset composed of 41,122 particles produced a postprocessed reconstruction at 2.81 Å, automatically sharpened by Relion with a B-factor of -72.9 Å^2. Moreover, the major 3D class (Class 7) obtained by the 3D autoclassification method and composed by 37,154 particles produced a postprocessed reconstruction at 2.78 Å^2 automatically sharpened by Relion with a B-factor of -71.8 Å^2. These results show that the proposed autoclassification methods can produce improved 3D reconstructions as measured by resolution and B-factors.

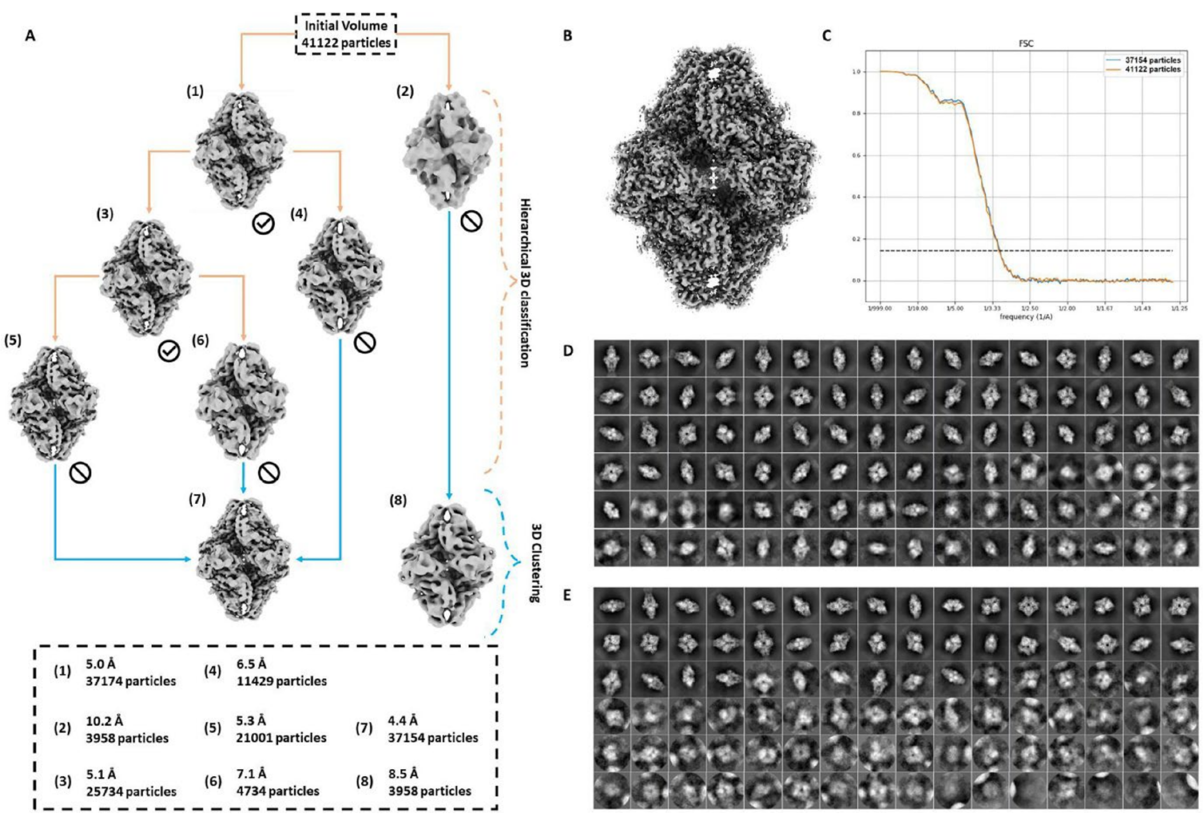


Figure 3.4 3D and 2D autoclassification results for beta-galactosidase in complex with a cell-permeant inhibitor. *A) Intermediate 3D classes obtained in the hierarchical 3D classification and clustering steps with corresponding populations and resolutions. B) 3D map computed by refinement of Class A particles only with Relion autorefine. C) Resulting Gold-standard FSCs after refining Class A particles (blue curve) and Class A + Class B particles (orange*

curve) with Relion autorefine. D) 2D class averages obtained by the 2D autoclassification approach. E) 2D class averages obtained by 2D classification method of Relion.

3.4.4 Complex AP-1:Artf1:tetherin-HIV-Nef

This dataset comprises 53,841 particles collected on a Titan Krios using a K2 direct electron detector (Gatan) in super-resolution counting mode. The nominal magnification is 22,500X providing a calibrated pixel size of 1.07 Å/pixel and is available from EMPIAR-10,178. Each particle image is 224 × 224 pixels in size and the CTF parameters are available from EMPIAR. We used this dataset to test our proposed 2D autoclassification approach. Thus, we performed 2D classification using both the proposed approach and Relion 2D classification. The 'number of particles' threshold to stop the automatic reclassifications in our approach was set to 200 particles and we asked for 220 classes in Relion 2D classification. The 2D autoclassification approach yielded 198 2D classes (Figure 3.5A). As can be seen from Figure 3.5A, among these 198 2D classes, only 23 show low SNRs and should be removed for further processing steps. In Figure 3.5A, we mark these low SNR classes with red squares. On the other hand, Relion 2D classification generates many blurred and low quality 2D classes, with only 70 classes out of 220 showing appropriate SNRs. Figure 3.5B renders the 2D class averages computed by Relion, where good classes with high SNRs are marked with green squares.

A

B

__Figure 3.5 2D autoclassification for the complex AP-1:Artf1:tetherin-HIV-Nef.__ A) 2D class averages obtained using the proposed 2D autoclassification approach. Class averages enclosed in *red squares represent "junk" classes and should be removed from the workflow. B) 2D class averages obtained by Relion 2D classification. Class averages enclosed in the green box represent good 2D classes with acceptable signal-to-noise ratio.*

3.5 Discussion

Cryo-EM using single particle analysis is a mature and powerful technique to obtain 3D reconstructions of macromolecules in near native conditions at atomic resolution. One of its major advances is its capacity to reveal conformations of dynamic molecular complexes. Most popular and successful current approaches to analyze heterogeneous complexes are founded on Bayesian inference[47,7,17,2]. These methods can classify cryo-EM particles according to their conformation, obtain their 3D orientation simultaneously, and have shown their robustness and capacity to process challenging datasets. These challenging cases include small and large samples, datasets affected by large numbers of false positive particles and/or particle images showing low signal-to-noise-ratios (SNRs) or low contrast. According to the Electron Microscopy Databank, the Relion 3D classification approach accumulates 2594 records for final 3D classification out of 3362 released entries (77%). However, these 3D classification methods require the tuning of specific parameters by the user, as the number of classes or the refinement iterations. Moreover, these 3D classification approaches are affected by the 'attractor problem' that limits the number of different classes that these methods can discern within the data, irrespective of the number of classes selected by the user. This limitation imposes the use of complicated 3D re-classification procedures, whose success highly depends on the user's experience. Recently, deep-learning approaches to 3D classification of single particles have been implemented, for example[57,70]. These approaches have shown their capacity to resolve continuous and discrete heterogeneity for large complexes. However, the robustness of these methods to process more challenging cases as smaller and lower contrast samples has not yet been demonstrated. Moreover, these approaches require as input the orientation of the single particles, a consensus high resolution reconstruction, a large set of single particles to train the network (note that for cryoDRGN typically both the decoder and the encoder requires 3938,306 parameters to be trained), the dimension of the latent space and the number of epochs to train the network.

In this work, we propose automatic 3D and 2D classification approaches based on Bayesian inference through the Relion classification methods but working in a hierarchical fashion. These methods, which use Relion as the classification engine, allow unsupervised and automatic classification of single particles and do not require users to provide information such as the number of coexisting conformations in the dataset, or the number of classification iterations to be performed. The approaches described here prevent the attraction of different particles into classes with high SNRs and the use of complicated manual reclassification approaches. Thus, these methods will reduce the level of expertise required to process cryo-EM data, especially for difficult projects with samples showing massive heterogeneity. Therefore, users with reduced expertise in single particle image processing will be able to generate large sets of classes (either 2D or 3D) using a single and simple protocol without difficult-to-tune input parameters or reclassification approaches. In addition, our method will be useful in high-throughput streaming approaches, reducing the necessity of user intervention. Also, it can identify conformations of low population in the dataset. The proposed method is more computationally expensive than a normal 3D classification process in Relion. However, users processing datasets affected by extensive heterogeneity that have no clue about the number of coexisting conformations, are required to run multiple rounds of 3D classification using typically different inputs (number of 3D classes and initial volume), and intense further reclassification procedures. This workflow is based on trial and error, and it is driven by the user´s experience, thus, it could be very time consuming, subjective, and difficult to reproduce. As reference, we provide our processing times for the 3D autoclassification of 131,899 immature ribosomal particles (~ 18 h), 41,122 beta-galactosidase particles (~ 7 h) and 327,490 spliceosomal particles (~ 40 h). In all cases, we used a processing server equipped with 6 Nvidia Titan X GPUs.

The proposed approach has been tested with four different datasets comprising both homogeneous and heterogeneous samples. The first case is the L17-depleted 50S ribosomal intermediates that represent immature ribosomal 50S subunits showing extensive compositional heterogeneity. The proposed 3D autoclassification approach provides 65 raw 3D classes that were automatically clustered into 18 final 3D classes. These final classes could be refined to 3D reconstructions ranging between 4.3 and 6.9 Å. Moreover, the 65 raw 3D classes were used to compute an energy landscape of macromolecules. Secondly, we processed single particles of the spliceosomal B-complex from yeast, which also shows massive conformational heterogeneity. As in the previous case, we could compute many different raw 3D classes (102 3D classes) that were automatically clustered into 14 final 3D classes. The raw 3D classes were used to build an energy landscape and to trace sorted structural trajectories. Then, we processed the beta-galactosidase in complex with a cell-permeant inhibitor, which represents a homogeneous set. We obtained 4 raw 3D classes, that were automatically clustered in two final 3D classes, one major (90% of the dataset) and one minor (10% of the dataset). The minor class seems to be composed of a large percentage of low-quality particles, as the refinement of the complete dataset with Relion (41,122 particles) produces slightly lower quality results than the reconstruction obtained by the major class only (90% of the dataset). This shows that the proposed approach can be used for particle screening as well. Finally, we processed the Complex AP-1:Artf1:tetherin-HIV-Nef to test the 2D autoclassification method. In the case of the 2D autoclassification, we have shown that the proposed 2D autoclassification approach can provide more meaningful 2D class averages than the Relion 2D classification method, including low-populated views and reducing the number of empty classes as it is less affected by the attractor problem. Thus, the 2D autoclassification approach can provide large numbers of 2D classes, which can be used, for example, as input for *ab initio* initial map algorithms to improve the accuracy of these methods and to obtain initial maps of low populated conformations.

3.5.1 Software and data availability

The presented methods are included in our software package CryoMethods under the names 3D autoclassifier, 2D autoclassifier and ML_landscape. The source code is freely available under the terms of an open-source software license and can be downloaded from https://github.com/mcgill-femr/cryomethods and https://github.com/mcgill-femr/scipion-em-cryomethods. These methods can be used through Scipion platform https://github.com/I2PC/scipion. All data used in this work is publicly available from EMPIAR database under the following codes EMPIAR-10,076, EMPIAR-10,180, EMPIAR-10,061 and EMPIAR-10,178.

3.5.2 Acknowledgments

Authors want to acknowledge economical support from the Spanish Ministry of Science and Innovation through the call 2019 Proyectos de I+D+i - RTI Tipo A (PID2019-108850RA-I00).

3.6 References

1. Merk, A. *et al.* Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* **165**, 1698–1707 (2016).

2. Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination

in RELION-3. *Elife* **7**, (2018).

3. Danev, R., Yanagisawa, H. & Kikkawa, M. Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions. *Trends Biochem. Sci.* **44**, 837–848 (2019).
4. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).
5. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).
6. Scheres, S. H. W., Núñez-Ramírez, R., Sorzano, C. O. S., Carazo, J. M. & Marabini, R. Image processing for electron microscopy single-particle analysis using XMIPP. *Nat. Protoc.* **3**, 977 (2008).
7. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
8. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, 377–388 (2013).
9. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
10. Khoshouei, M., Radjainia, M., Baumeister, W. & Danev, R. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat. Commun.* **8**, 1–6 (2017).
11. Lyumkis, D. *et al.* Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science* **342**, 1484–1490 (2013).
12. Zhang, K. *et al.* Cryo-EM and antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *Nat. Struct. Mol. Biol.* **28**, 747–754 (2021).
13. Sorzano, C. O. S. *et al.* A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.* **171**, 197–206 (2010).

14. Wu, J. *et al.* Massively parallel unsupervised single-particle cryo-EM data clustering via statistical manifold learning. *PLoS One* **12**, e0182130 (2017).

15. Vargas, J. *et al.* Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *J. Struct. Biol.* **183**, 342–353 (2013).

16. Vargas, J., Otón, J., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. Particle alignment reliability in single particle electron cryomicroscopy: a general approach. *Sci. Reports* **6**, 1–11 (2016).

17. Vargas, J., Melero, R., Gómez-Blanco, J., Carazo, J. M. & Sorzano, C. O. S. Quantitative analysis of 3D alignment quality: its impact on soft-validation, particle pruning and homogeneity analysis. *Sci. Reports* **7**, 1–14 (2017).

18. Sanchez-Garcia, R., Segura, J., Maluenda, D., Carazo, J. M. & Sorzano, C. O. S. Deep Consensus, a deep learning-based approach for particle pruning in cryo-electron microscopy. *IUCrJ* **5**, 854–865 (2018).

19. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. Sect. D Struct. Biol.* **73**, 496–502 (2017).

20. Zhou, Y., Moscovich, A., Bendory, T. & Bartesaghi, A. Unsupervised particle sorting for high-resolution single-particle cryo-EM. *Inverse Probl.* **36**, 044002 (2020).

21. Davis, J. H. *et al.* Modular Assembly of the Bacterial Large Ribosomal Subunit. *Cell* **167**, 1610–1622 (2016).

22. Razi, A. *et al.* Role of Era in assembly and homeostasis of the ribosomal small subunit. *Nucleic Acids Res.* **47**, 8301–8317 (2019).

23. Haselbach, D. *et al.* Long-range allosteric regulation of the human 26S proteasome by 20S proteasome-targeting cancer drugs. *Nat. Commun.* **8**, 1–8 (2017).

24. Haselbach, D. *et al.* Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell* **172**, 454–464 (2018).

25. Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S. H. W. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *Elife* **7**, (2018).

26. Harastani, M., Sorzano, C. O. S. & Jonić, S. Hybrid Electron Microscopy Normal Mode Analysis with Scipion. *Protein Sci.* **29**, 223–236 (2020).

27. Frank, J. & Liu, W. Estimation of variance distribution in three-dimensional reconstruction. I. Theory. *JOSA A* **12**, 2615–2627 (1995).

28. Penczek, P. A., Kimmel, M. & Spahn, C. M. T. Identifying Conformational States of Macromolecules by Eigen-Analysis of Resampled Cryo-EM Images. *Structure* **19**, 1582–1590 (2011).

29. Tagare, H. D., Kucukelbir, A., Sigworth, F. J., Wang, H. & Rao, M. Directly reconstructing principal components of heterogeneous particles from cryo-EM images. *J. Struct. Biol.* **191**, 245–262 (2015).

30. Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021).

31. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. Methods* **18**, 176–185 (2021).

32. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).

33. Zhang, Z. & Zha, H. Principal Manifolds and Nonlinear Dimensionality Reduction via

Tangent Space Alignment. *J. of Shanghai Univ.* **26**, 313–338 (2006).

34. Donoho, D. L. & Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5591–5596 (2003).

35. L. Van der Maaten & G. Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

36. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).

37. Rosenthal, P. B. & Henderson, R. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

38. Stagg, S. M., Noble, A. J., Spilman, M. & Chapman, M. S. ResLog plots as an empirical metric of the quality of cryo-EM reconstructions. *J. Struct. Biol.* **185**, 418–426 (2014).

39. Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M. & Förster, F. Fast and accurate reference-free alignment of subtomograms. *J. Struct. Biol.* **182**, 235–245 (2013).

40. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods* **13**, 387–388 (2016).

41. Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nat.* **546**, 617–621 (2017).

42. Bartesaghi, A. *et al.* 2.2 å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science* **348**, 1147–1151 (2015).

43. Morris, K. L. *et al.* HIV-1 Nefs Are Cargo-Sensitive AP-1 Trimerization Switches in Tetherin Downregulation. *Cell* **174**, 659–671 (2018).

44. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron

micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).

45. Vargas, J., Álvarez-Cabrera, A. L., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. Efficient initial volume determination from electron microscopy images of single particles. *Bioinformatics* **30**, 2891–2898 (2014).

46. Gomez-Blanco, J., Kaur, S., Ortega, J. & Vargas, J. A robust approach to ab initio cryo-electron microscopy initial volume determination. *J. Struct. Biol.* **208**, 107397 (2019).

47. Scheres, S. H. W. Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods Enzymol.* **579**, 125–157 (2016).

CHAPTER 4

OVERALL DISCUSSION AND SUMMARY

The discussion chapter is structured in two sections. In the first one, I present a discussion of each of the included papers in this thesis, in the same order in which they appear in chapters 2 and 3. The second section addresses the challenges overcome by the described methods as well as the contribution of this thesis work in the field of cryo-EM using SPA, reflecting upon the biological significance and future expectations.

4.1 New methods to improve cryo-EM map locally at high-resolution for building an accurate atomic model

In cryo-EM, attaining a high resolution is very important to build an atomic model[202], thereby allowing us to achieve the structural understanding of a macromolecule of interest, and proferring clues as to its function *in vivo*. However, cryo-EM maps with different local SNR or heterogeneity can hinder our understanding. A significant impact of such heterogeneity is the contrast loss at high-resolution in cryo-EM maps due to the delocalization of density features as domains move or manifest a different occupancy of binding partners.

The publication included in this thesis[1] of chapter 2 shows a new way to use the 3D spiral phase transform (having been heretofore employed in particle screening[235], CTF estimation[236] and local/directional resolution determination[237,238]) to recover continuity in broken map densities and provide better map connectivity for heterogenous cryo-EM maps for effective atomic modeling. These methods are incorporated in the cryo-EM workflow as part of sharpening techniques (used as a post-processing step), which is a well-established practice used at the end of reconstruction to enhance interpretability of the map (i.e. improve map contrast), and therefore to trace the atomic model.

Several published methods have already used our robust techniques to achieve their respective goals as discussed in section 4.3.2. All these results from our publication of chapter 2 and citations[239–241] indicate that the proposed sharpening methods increase the interpretability of reconstructed maps. It has been observed that most of the EMDB deposited cryo-EM maps are sharpened maps[242]. This allows sharpening methods to be a currently active area of methods development[239–241].

However, sometimes when the main focus is sharpening the cryo-EM maps, experimental map signals can be easily lost due to oversharpening, which leads to losing the true structural information of the macromolecule. Therefore, one of our goals here was to maintain the original experimental signal properties when performing map enhancement, which has been achieved by boosting the local map amplitude values, while not using inputs such as atomic models as a reference or local resolution estimation. As methods improvement is an ongoing task in cryo-EM, our approach (LocSpriral and LocOccupancy) depends on reliable solvent masks, to differentiate the macromolecule from the noise.

A surge in cryo-EM-generated atomic models has occurred due to the increased deposited number of cryo-EM maps with high resolution. For building an atomic model, cryo-EM maps are generally post-processed to increase the contrast of their high-resolution features, instead of starting from the raw maps, but there is no standardized approach for how this is done. Since our proposed method directly contributes to high-quality post-processing operations by local sharpening, it is our hope that this work will augment the suite of tools that can be used for a community consensus approach for map sharpening and validation.

4.2 Analysing heterogeneous data in the form of automatic 3D classification and trajectory

Cryo-EM has the potential to analyse heterogeneous data of a biomolecule for uncovering its various conformations. The publication presented in chapter 3 introduces a new automatic method to perform 2D or 3D classification for heterogenous cryo-EM data and to generate a free-energy landscape, without requiring knowledge of the number of conformations present in the dataset or the number of iterations over which to classify. Using controlled parameters, we have discovered that the type of 2D and 3D classification method and choice of linear or non-linear manifold embedding are very important factors to analyze the conformations of a given macromolecule.

The aim of our proposed 2D autoclassification approach (in chapter 3) is to provide many different 2D views, including low populated ones. According to this main objective, it is clear from our results that our approach performs better than Relion 2D classification. However, it is also true that Relion yields superior screening particle images that originate from artifacts, pure background or very low SNR. We cannot assure that particle picking will not generate false positives. Even so, our method shows meaningful improvement as compared to Relion, e.g., our 2D autoclassification approach yielded 198 2D classes (as can be seen from Fig. 3.5A), of which only 23 harbour low SNRs (marked with red squares and should be removed for further processing steps). On the other hand, Relion 2D classification generates many blurred and low-quality 2D classes, with only 70 classes out of 220 showing appropriate SNRs. From this perspective, our approach emerges as a refreshing triumph, even if it scores less successfully vis-à-vis particle screening.

To be sure, our approach is based on Bayesian inference through Relion classification, albeit working in a hierarchical fashion. This means that some obtained raw 3D classes could represent the same conformation or very similar conformations. To help the interpretation of the results and avoid cases where the 3D classes are very similar or just the same, our approach can run a final clustering test, which is optional. Thus, both the raw 3D classes and the obtained classes after clustering are important and complementary information that should be analysed.

The output of our approach is not biased towards major classes as per standard Bayesian methods[91,155,228,243], i.e. is not affected by the "attractor problem", as our propped method classifies particles based on their actual conformations instead of their SNRs. Therefore, the resultant classes are high- as well as low-populated (in terms of particle population).

The free energy landscape provides a system where the potential energy is defined as a function of all coordinates[244]. Ergo, the "story of the sample" can be discovered by analyzing the trajectory of conformation changes for a dynamic macromolecule. The purpose of using different machine learning techniques in our method (PCA, non-linear manifold embedding) is described in section 4.3.1.

On a free-energy landscape, it is important to recognize the meta-stable states, activation barriers as well as transition states for a macromolecule, which are key to biomolecule function. Generally, the lower energy path of the free-energy landscape decides the trajectory of changing conformations for a macromolecule, which includes lower energy meta-stable states separated by high energy transition states and activation barriers. If the experimental sample information is previously known, such as number of transitioning, active or inactive states present, it can be compared with obtained energy landscape to confirm the trajectory. However, a method to determine the optimal trajectory path remains to be developed.

Although not discussed explicitly in this thesis, it is important to note that free-energy landscapes obtained directly from experimental cryo-EM datasets are reflections of the sample preparation methods used before freezing the sample on a grid. These preparative methods can have a dramatic influence on the output of the free-energy landscapes. For example, temperature plays a key role in determining the free-energy landscape as per Eqs. 3.1, 3.2. These equations are based on the assumption that cryo-EM samples undergo instant vitrification and therefore various states of a macromolecule are trapped in the ensemble just before freezing. It matters whether the sample was at 4 ºC or 22 ºC prior to vitrification. However, in the real world, vitrification is a slow process and hence can generate artificial energy barrier conditions. As such, this freezing process itself can affect the outcome, and our equations require further evaluation in the cryo-EM field to generate effective free-energy landscapes.

The landscape that we obtain in chapter 3 represents the energy landscape of the reaction catalysed by the removed ribosomal protein rather than showing a molecular free-energy landscape (Fig.3.2). These landscapes are very interesting to trace trajectories in ribosome biogenesis for determining the sequence of conformations towards ribosome maturation.

Our approach's robustness has been proven over various samples and has shown noteworthy results, as conclusively demonstrated in chapter 3. Shortcomings in the form of future goals of this method, along with others proposed in this thesis, are discussed in section 4.5.

4.3 General discussion and contribution

3D-EM has advanced progressively since its first application to the T4 bacteriophage tail[245]. Interest in the cryo-EM for structural analysis has spiked over the last years by virtue of its ability to achieve atomic details, which were previously observed only in X-ray crystallography. Around the world, institutions and drug companies are setting up cryo-EM facilities and providing employment at a significant rate. This all happened because of new instrumentation, robust image analysis methods, and automation in data collection and analysis. To achieve atomic details in a macromolecule, cryo-EM had to overcome numerous obstacles. Some of these are outlined below:

Automation of image processing methods is essential and a major challenge to increase their throughput. Using an automatic algorithm can (i) reduce the errors in structural determination workflow introduced by the user's lack of right judgment or less expertise on the given technique; (ii) reduce the time and effort required to achieve the desired goal; and (iii) provide high-throughput results. Our approach (described in chapter 3) serves this goal of automation but does not require input information from the user, such as number of conformations present in cryo-EM data and the number of iterations per classification step.

Cryo-EM faces major challenges when dealing with heterogeneous samples. Heterogeneity can be caused by a macromolecule's domain flexibility, which can hinder achievement of high-resolution 3D reconstructions. When the resolution is globally measured, small changes in the map generally go unnoticed as they do not affect the resolution value. Therefore, a close inspection of the cryo-EM density map is required which can show the blurred regions which do not possess the same resolution as the global value and can be caused by unaccounted sample heterogeneity. Ergo, this local heterogeneous distribution of resolution is vital to recognize superior 3D reconstruction quality and ultimately evaluate the atomic model. I have already discussed several methods in chapter 2, section 2.2, which describes the need for a sophisticated method to analyse these local resolutions in a map, and also introduces methods to improve the interpretability of a heterogenous cryo-EM map. Chapter 3 furnishes details about analysing various functionally relevant conformations of a structure, which generally requires classification of the particle images into structurally homogeneous subsets to achieve most of the conformations present in cryo-EM.

Some conformational heterogeneity of a macromolecule can go unnoticed, due to lack of advanced 2D or 3D classification approaches, which requires user input to provide number of final output conformation and iteration, thus might lead to a misinterpretation of the resultant structural heterogeneity. Such a limitation can hinder the comprehensive interpretation of biological function concerning the conformational changes. Both the proposed approaches, described in chapters 2 and 3 explain in the description, the heterogeneous samples and the methods to deal with such data.

In case of a model-building challenge, it can be difficult to find an algorithm designed to fit an atomic model to density maps of various structures. Software tools are available to generate high-quality models, however, these methods require further analysis and generate new approaches to better assess model correctness, because, for example, 50% of the EMD database deposited maps were generated using a single software tool, missing chances to perform a broader assessment of algorithms[246]. Occupancy values (defined in chapter 2, section 2.2, along with B-factor values) are the additional important parameters to refine an atomic model for adequately representing the experimental map. Additionally, there is no package currently available among cryo-EM techniques that can measure occupancy maps locally. Presence of unrealistic occupancy values, which has been reported in 31% of all models deposited in PDB and EMDB, resulted in very low correlation coefficient values between the model and the corresponding experimental maps. This issue can be addressed using an automated or semi-automated technique. Our method introduced in chapter 2 describes a semi-automated technique that generates better quality atomic models along with local occupancy maps calculations for heterogeneous data without the need of input parameters such as atomic models or local resolution maps.

Our proposed methods (chapter 2 and 3) not only deal with these limitations of single-particle cryo-EM but also provide excellent results. The motives of introduced approaches are to:

- ***improve de novo model building***
- ***process cryo-EM map with different SNRs***

- ***design automated or semi-automated cryo-EM image processing methods***
- ***process heterogeneous data by the automatic Bayesian-based classification approach***
- ***analyse the trajectory for a dynamic macromolecule***

Moreover, by combining algorithms from both of our publications, high resolution conformations can be achieved. For example, if the input is maps affected by inhomogeneous local resolutions/SNRs and broken densities, our local map enhancement approach (LocSpiral)[1] (mentioned in chapter 2, section 2.3.1) can be used to obtain maps with better connectivity. These high-quality maps can be harnessed as initial volumes to reclassify the cryo-EM particles according to our automatic hierarchical clustering approach for performing 2D/3D classification (described in chapter 3)[2], which will then be able to process massive heterogeneous data. This higher or N dimension output can be further reduced using a non-linear dimensionality reduction algorithm as explained in chapter 3, thereby obtaining the free-energy landscape representing the conformational changing trajectory of a macromolecule at a comparatively better resolution.

These approaches employ various easy-to-use tools and reliable algorithms, which have already been tested in various publications, as mentioned below.

4.3.1 Machine learning algorithm

The proposed methods in the thesis make excellent use of machine learning tools, such as the specific machine learning implemented in chapter 3 to generate a free-energy landscape. By reducing high-dimension data to lower dimensions, these tools facilitate the quest for meaningful visual perceptions. Such reduction works analogously to how the human brain converts input sensory information from $3*10^4$ auditory nerve fibres or 10^6 optic nerve fibres into the minimum required number of relevant features. The same holds true in other fields such as global climate patterns or human gene distributions.

There are several classical linear approaches employed for this purpose, such as PCA[247] and MDS[248], which can accurately reduce the datasets on or near a linear subspace of the high-dimensional input space. However, complex datasets contain non-linear structures, which both PCA and MDS fail to detect. A non-linear manifold can be described as heterogenous cryo-EM data.

In a simplified form, it can be shown as in Figure 4.1. Classical methods (Figure 4.1(a)), determine the Euclidean structure; thus, they fail to detect the intrinsic geometry of the data. By contradistinction, non-linear methods (Figure 4.1 (b)) use the geodesic distance, or short path, by calculating manifold distances between all pairs of data points, as it captures the non-linearity of the manifold. There are several tools for nonlinear dimensionality reduction such as LTSA[249], Hessian LLE[250] or t-SNE[251], and isomap[252]. Our method in chapter 3 used isomap for generating a free energy landscape, while the other mentioned nonlinear manifold approaches can be implemented as well. Isomap reduces the dimension by preserving the geodesic distance in lower dimension. Here, isomap is not taking the Euclidean distance between data points into account, because the Euclidean distance neglects the shape of the manifold while reducing the dimension, whereas geodesic distances are measured according to the manifold shape, as explained in Figures 4.1A and 1B. The output of non-linear manifold embedding, used in our method, is in the form of vectors y_i in the lower dimensional space that preserves the intrinsic geometry of the input high-dimension data[252].

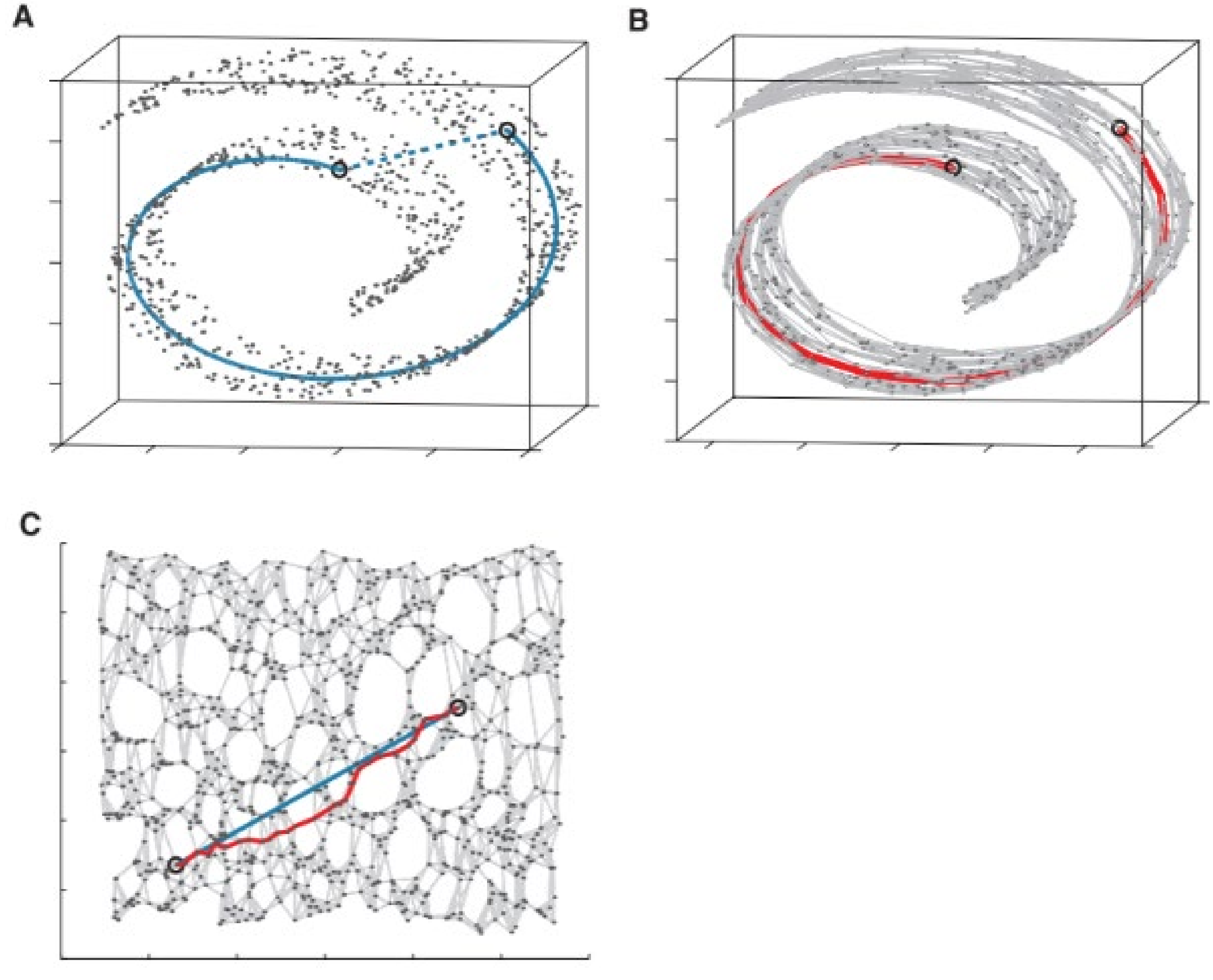


Figure 4.1*: Non-linear dimensionality reduction for isomap.* *Various forms of distance measurements methods in high dimension data are shown for non-linear data or in this case 'swiss-roll' form: a) Euclidean distance (dotted line) and geodesic distance (solid curved line). b) 2D data recovered preserving the shortest distance. c) True geodesic distance (red line) is now shown in a simple and clean approximation (blue line). (Reproduced with permission from Tenenbaum J.B. et al. 2000 from[252]).*

4.3.2 Map enhancement approaches (LocSpiral)

Our LocSpiral approach (Chapter 2) has been used in several publications and proven to be successful with efficient results as follows:

- This method helped to recover the broken density for segment M89-K101 due to flexibility or 16 nm averaging of the cryo-EM data, in the case of the model of *C. reinhardtii* PACRG (B1B601)[253].
- For high-resolution cryo-EM map of *Tetrahymena* doublet microtubule, LocSpiral method helped to achieve *de novo* model building[254].
- Tetrameric ArnA cryo-EM map had broken densities due to carboxylase domain movement. Better map connectivity was observed after using our algorithm of LocSpiral[255].
- Enhanced the structure analysis and interpretability for ligand-bound IR-ECD[256].
- Fragmented densities of 30S subunit due to flexible conformation were improved to achieve an informative molecular model[257].

4.4 Biological significance

The methods identified in this thesis have the purpose of dealing with heterogeneity as well as producing better-quality atomic models by using automated or semi-automated methods to process cryo-EM data. The final goal is to understand the molecular structural and structural dynamics of a biomolecule. To that end, the higher the structure quality, the better is the interpretation of its functioning.

In structural analysis techniques, cryo-EM has gained an exemplary reputation over the years to provide near-atomic resolution maps for a diverse range of macromolecules between tens to thousands of kilodaltons. An atomic model of cryo-EM maps elucidates the molecular interaction of macromolecules in chemical and physical terms, in addition to revealing other information such as 3D structure analysis, comparison between different macromolecule structures, details on complex formation between different biomolecules and prediction of structures of new related macromolecules. Being a strong graphic tool, the atomic model also helps to uncover side-chain interactions, binding pockets, and catalytic regions, which generally lays the foundation for drug discoveries. Along with improving atomic model quality, our proposed method (in chapter 2) has enhanced the analysis of the cryo-EM map. For example, in the case of the SARS-CoV-2 structure (chapter 2, section 2.3.6), our technique was not only able to improve the visualization of deposited EMD maps but also helped to include additional motifs.

Among various structural biology techniques, cryo-EM is the one to hold potential for processing both sample heterogeneity and inherent flexibility. Macromolecules as dynamic machines undergo conformational rearrangement during the cellular process to perform several biological functions[258]. These movements can be classified as discrete or as occurring along a continuum of several conformations, where either one subunit or several subunits of a macromolecule complex moves independently of each other. Tracing these conformations is of key importance to understanding a high-resolution structure and hence the function of the protein complex of interest. Many of these protein complexes are in therapeutics development studies. Researchers can even visualize the effect of drug binding on macromolecular complex structures in the form of energy landscapes. In Haselbach et al, 2017[259], cancer drugs are shown to impact the structure of the 26S proteasome on a free-energy landscape. Methods (proposed in chapter 3) to gain information like this have the potential to unravel the mystery of various unanswered questions regarding the trajectories for a germ's key macromolecules, e.g., how drug binding (like the above-mentioned example) can modify the energy landscape of conformational motion in a germ's macromolecule and will play a key role in therapeutics production.

4.5 Future Goals

4.5.1 Normal Mode Analysis

As discussed in the publication of chapter 3, the free-energy landscape consists of high energy regions, where few to no cryo-EM maps are present, as well as lower energy regions that represent most of the maps. It is possible that during all image processing algorithms, some of the information of a macromolecule can be lost because the image processing pipeline is not ideally perfect. In the case of heterogeneous data, after the 3D classification step of the cryo-EM image processing pipeline (chapter 1 section 1.4.4.7), the resolution obtained may be limited by the particle count belonging to the 3D homogenous classes. Therefore, some minor classes do not have enough observations to show the secondary structures, which leads to untraced conformations. While generating the free-energy landscape in chapter 3, these untraced conformations leave visible gaps in the conformational space, thereby hampering the trajectory analysis of a flexible macromolecule.

The future optimistic goal is to fill these gaps or lower energy spaces on the free-energy landscape as much as possible. To that effect, NMA[260,261] can be a major contribution to the free-energy landscape algorithm. It is a powerful computational tool that helps to analyse the large-amplitude motions as a collection of simple harmonic oscillations vibrating around an energy minimum as shown in Figure 4.2. This method can have an algorithm as followS:

1. 3D classes obtained from the automated 3D classification method (explained in chapter 3, section 3.3.1) can be used as an input, which will represent the various states of conformations present in a given cryo-EM data.
2. This obtained set of 3D classes is considered as a sampling point in a corresponding free-energy landscape. To fill the gap in the conformational space of the obtained free-energy

landscape, information around each sampling point is extended by considering each class as pseudo-atoms, which will be synthetically perturbed using NMA[260,261].

3. NMA works by taking each input class as a reference class to predict its movements and then aligning (with elastic 3D-to-2D alignment procedure) with the EM images to verify whether the predicted movements actually transpired in the sample. Finally, the reference structure is deformed into a set of conformations using possible directions of the conformational change predicted by NMA.
4. A common resolution will be achieved using a low pass filter for both originally obtained 3D classes and those obtained by NMA perturbation. These resultant classes are used by the same automated 3D classification method, used initially in point 1, as initial maps.
5. This resultant large conformational space can be used to generate free-energy landscape as per our method described in chapter 3.

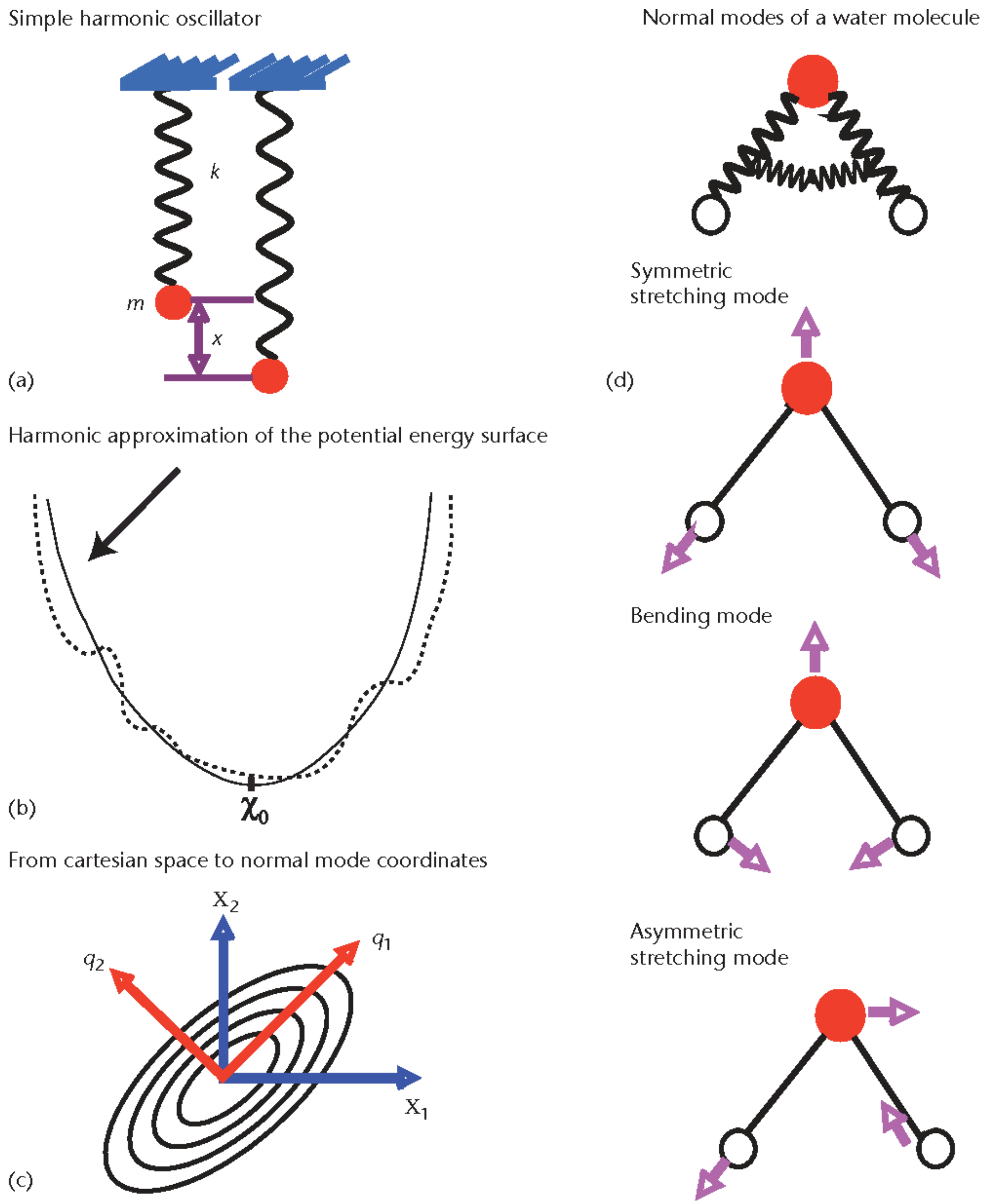


Figure *4.2*: ***Description of NMA.*** *a) Simple harmonic oscillator: m represents the particle, k is force constant of a spring attached to m and x is the displacement. b) Rugged surface (dotted) is the real energy of biological sytems and harmonic surface (plain line) represents the NMA. c) Equipotential points of a parabolic force field in a 2D space are described by contour lines here. Blue axis represents the cartesian coordinates while red axes are for NM coordinates. Particle's motion is explained according to which coordinates are pertinent (one axis if on one of the NM axes, both axes if on any of the Cartesian axes). d) Three types of motion for water molecule as predicted by NMA (symmetric stretching mode, asymmetric stretching mode, bending mode),* *with arrows showing the direction of motion of an atom measured by normal-mode theory. This figure is reproduced with permission from López-Blanco JR et al, 2014 from*[262].

Some of the macromolecules such as ribosomes exist in various intermediate conformations before the mature state. Such a large-scale conformational pathway cannot be achieved by current computational algorithms. In the typical cryo-EM 3D reconstruction process, most of the structures are not further analysed because of the low particle abundance. This limitation hinders the correct structural analysis of macromolecular complexes. However, with the aid of NMA, various conformational states, representing major as well as minor classes, can be accessed. This information can then be mapped to our multidimensional free energy landscape algorithm (chapter 3, section 3.3.3) for visualization and interpretation.

4.5.2 To deal with macromolecules with large conformational changes

While using the algorithms in chapter 2, to perform sharpening and evaluating atomic model, it should be noted that:

- In case of measuring the map signal and occupancy maps[1], it is possible that when a sample is showing large conformational changes, proposed methods can observe lower density values and give a close-to-zero occupancy value.
- These proposed sharpening methods work well if 3D classification has been carefully executed. Otherwise, macromolecules with compositional heterogeneity can show density values of a significant amount on output 3D maps, which should be empty in actuality.

These concerns point out that dealing with high amount of heterogeneity and 3D-classification step of the cryo-EM workflow are the crucial aspects to consider to refine our proposed algorithms[1] in the future. Therefore, if there will be a 3D classification method in the future, that can analyse the heterogeneity to generate high-resolution cryo-EM maps, finer than the existing standard methods[91,243,263,264] (which cannot trace 3D classes with lower-particle count), then our technique will prove invaluable to analysing map signals and density values. In this case, one of the solutions to consider for processing massive heterogeneous data can be our automatic hierarchical clustering-based 2D/3D classification approach introduced in chapter 3, which can render a significant number of conformations, even including the ones with low particle population because this method is not affected by the attractor problem[22,23], the latter being commonly present in standard Bayesian classification methods (as discussed in chapter 3).

4.6 Concluding Remark

In toto, all the algorithms introduced in this thesis have shown promising results and space for progression in the future, to further improve the structure details in cryo-EM. Our findings indicate that i) cryo-EM maps with different SNRs can be enhanced with better connectivity and without broken densities, ii) *de novo* model building can be improved using local B-factors and local occupancy maps, iii) a significant number of conformations can be extracted from input sample data for homogenous as well as heterogeneous macromolecules, without the “attractor” problem, iv) the free-energy landscape can elaborately explain the conformational changing trajectory for flexible macromolecules. Such methods represent the finest advancements in the cryo-EM image processing workflow.

References

1. Kaur, S. *et al.* Local computational methods to improve the interpretability and analysis of cryo-EM maps. *Nat. Commun.* **12**, 1–12 (2021).

2. Gomez-Blanco, J., Kaur, S., Strauss, M. & Vargas, J. Hierarchical autoclassification of cryo-EM samples and macromolecular energy landscape determination. *Comput. Methods Programs Biomed.* **216**, 106673 (2022).
3. Renaud, J. P. *et al.* Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).
4. De Rosier, D. J. & Klug, A. Reconstruction of Three Dimensional Structures from Electron Micrographs. *Nature* **217**, (1966).
5. Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods* **13**, 24 (2016).
6. The Nobel Prize in Physics 1986 - Perspectives: Life through a Lens - NobelPrize.org. https://www.nobelprize.org/prizes/physics/1986/perspectives/.
7. Porter, K. R., Claude, A. & Fullam, E. F. A study of tissue culture cells by electron microscopy: Methods and preliminary observations. *J. Exp. Med.* **81**, 233–246 (1945).
8. Tanaka, K., Mitsushima, A., Fukudome, H. & Kashima, Y. Three-dimensional architecture of the Golgi complex observed by high resolution scanning electron microscopy. *J. Submicrosc. Cytol.* **18**, 1–9 (1986).
9. Song, J., Lee, C., Lin, C. H. S. & Chen, L. B. Electron microscopic studies of the endoplasmic reticulum in whole-mount cultured cells fixed with potassium permanganate. *J. Struct. Biol.* **107**, 106–115 (1991).
10. Masters, B. R. History of the electron microscope in cell biology. *eLS* (2009).
11. Bennett, M. R. The early history of the synapse: From plato to sherrington. *Brain Research Bulletin* **50**, 95–118 (1999).
12. Weber, A. & Franzini-Armstrong, C. Hugh E. Huxley: Birth of the filament sliding model of

muscle contraction. *Trends in Cell Biology* vol. **12**, 243–245 (2002).

13. Huxley, H. E. The double array of filaments in cross-striated muscle. *J. Biophys. Biochem. Cytol.* **3**, 631–648 (1957).

14. Kausche, G. A., Pfankuch, E. & Ruska, H. Die Sichtbarmachung von pflanzlichem Virus im Übermikroskop. *Naturwissenschaften* **27**, 292–299 (1939).

15. von Borries, B., Ruska, E. & Ruska, H. Bakterien und Virus in Übermikroskopischer Aufnahme - Mit einer Einführung in die Technik des Übermikroskops. *Klin. Wochenschr.* **17**, 921–925 (1938).

16. Glaeser, R. M. Limitations to significant information in biological electron microscopy as a result of radiation damage. *J. Ultrasructure Res.* **36**, 466–482 (1971).

17. Renaud, J. P. *et al.* Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).

18. Amos, L. A. & Finch, J. T. Aaron Klug and the revolution in biomolecular structure determination. *Trends in Cell Biology* vol. **14**, 148–152 (2004).

19. Teng, T. Y. & Moffat, K. Radiation damage of protein crystals at cryogenic temperatures between 40 K and 150 K. *J. Synchrotron Radiat.* **9**, 198–201 (2002).

20. Taylor, K. A. & Glaeser, R. M. Electron microscopy of frozen hydrated biological specimens. *J. Ultrasructure Res.* **55**, 448–456 (1976).

21. Henderson, R. & Unwin, P. N. T. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257**, 28–32 (1975).

22. Dubochet, J., Lepault, J., Freeman, R., Berriman, J. A. & Homo, J. -C. Electron microscopy of frozen water and aqueous solutions. *J. Microsc.* **128**, 219–237 (1982).

23. Henderson, R. *et al.* Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213**, 899–929 (1990).

24. Transmission Electron Microscope (TEM) - Bioscience Notes. http://www.biosciencenotes.com/transmission-electron-microscope-tem/.

25. Kastner, B. *et al.* GraFix: sample preparation for single-particle electron cryomicroscopy. *Nat. Methods* **5**, 53–55 (2008).

26. Brenner, S. & Horne, R. W. A negative staining method for high resolution electron microscopy of viruses. *Biochim. Biophys. Acta* **34**, 103–110 (1959).

27. Thompson, R. F., Walker, M., Siebert, C. A., Muench, S. P. & Ranson, N. A. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods* **100**, 3–15 (2016).

28. Burgess, S. A., Walker, M. L., Sakakibara, H., Oiwa, K. & Knight, P. J. The structure of dynein-c by negative stain electron microscopy. *J. Struct. Biol.* **146**, 205–216 (2004).

29. Bakker, S. E. *et al.* The respiratory syncytial virus nucleoprotein-RNA complex forms a left-handed helical nucleocapsid. *J. Gen. Virol.* **94**, 1734–1738 (2013).

30. Taylor, K. A. & Glaeser, R. M. Electron Microscopy of Frozen Hydrated Biological Specimens. *J. Ultrastruct. Res.* **55**, 448–456 (1976).

31. Dubochet, J., Lepault, J., Freeman, R., Berriman, J. A. & Homo, J.-C. Electron microscopy of frozen water and aqueous solutions. *J. Microsc.* **128**, 219–237 (1982).

32. Saibil, H. R. Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallographica Section D: Biological Crystallography* **56**, 1215–1222 (2000).

33. Miller, J. L. *et al.* Three-dimensional reconstruction of Heterocapsa circularisquama RNA virus by electron cryo-microscopy. *J. Gen. Virol.* **92**, 1960–1970 (2011).

34. Fujiyoshi, Y. Low dose techniques and cryo-electron microscopy. *Methods Mol. Biol.* **955**, 103–118 (2013).

35. Thürmer, K. & Nie, S. Formation of hexagonal and cubic ice during low-temperature growth. *Proc. Natl. Acad. Sci.* **110**, 11757–11762 (2013).

36. Bokstad, M., & Medalia, O. Correlative Light Electron Microscopy as a Navigating Tool for Cryo-Electron Tomography Analysis. *Fluorescence Microscopy*, 121-131, Academic Press (2014).

37. Murata, K. & Wolf, M. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim. Biophys. Acta - Gen. Subj.* **1862**, 324–334 (2018).

38. Glaeser, R. M. Retrospective: radiation damage and its associated 'information limitations'. *J. Struct. Biol.* **163**, 271–276 (2008).

39. Bammes, B. E., Jakana, J., Schmid, M. F. & Chiu, W. Radiation damage effects at four specimen temperatures from 4 to 100 K. *J. Struct. Biol.* **169**, 331–341 (2010).

40. Steven, A. & Belnap, D. Electron microscopy and image processing: an essential tool for structural analysis of macromolecules. *Curr. Protoc. protein Sci.* **Chapter 17**, (2005).

41. Carter, M., & Shieh, J. C. Guide to research techniques in neuroscience. *Academic Press* (2015).

42. Otegui, M. S. Electron tomography and immunogold labeling of plant cells. *Methods Cell Biol.* **160**, 21–36 (2020).

43. Kaplan, M. *et al.* In Situ Imaging and Structure Determination of Biomolecular Complexes Using Electron Cryo-Tomography. *Methods Mol. Biol.* **2215**, 83–111 (2021).

44. Zhang, P. Advances in cryo-electron tomography and subtomogram averaging and classification. *Curr. Opin. Struct. Biol.* **58**, 249–258 (2019).

45. Turk, M. & Baumeister, W. The promise and the challenges of cryo-electron tomography. *FEBS Lett.* **594**, 3243–3261 (2020).

46. Briggs, J. A. G. Structural biology in situ--the potential of subtomogram averaging. *Curr. Opin. Struct. Biol.* **23**, 261–267 (2013).

47. Zhang, Y. *et al.* Molecular architecture of the luminal ring of the Xenopus laevis nuclear pore complex. *Cell Res.* **30**, 532–540 (2020).

48. Mahamid, J. *et al.* Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science* **351**, 969–972 (2016).

49. Cassidy, C. K. *et al.* CryoEM and computer simulations reveal a novel kinase conformational switch in bacterial chemotaxis signaling. *Elife* **4**, (2015).

50. Bykov, Y. S. *et al.* The structure of the COPI coat determined within the cell. *Elife* **6**, (2017).

51. Pfeffer, S. *et al.* Structure of the native Sec61 protein-conducting channel. *Nat. Commun.* **6**, (2015).

52. Tegunov, D., Xue, L., Dienemann, C., Cramer, P. & Mahamid, J. Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. *Nat. Methods* **18**, 186–193 (2021).

53. Qu, K. *et al.* Structure and architecture of immature and mature murine leukemia virus capsids. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11751–E11760 (2018).

54. Schur, F. K. M. *et al.* An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. *Science* **353**, 506–508 (2016).

55. Dick, R. A. *et al.* Structures of immature EIAV Gag lattices reveal a conserved role for IP6 in lentivirus assembly. *PLoS Pathog.* **16**, (2020).

56. Mattei, S., Glass, B., Hagen, W. J. H., Kräusslich, H. G. & Briggs, J. A. G. The structure and flexibility of conical HIV-1 capsids determined within intact virions. *Science* **354**, 1434–1437 (2016).

57. Dodonova, S. O., Prinz, S., Bilanchone, V., Sandmeyer, S. & Briggs, J. A. G. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. *Proc. Natl. Acad. Sci. U. S.*

A. **116**, 10048–10057 (2019).

58. von Kügelgen, A. *et al.* In Situ Structure of an Intact Lipopolysaccharide-Bound Bacterial Surface Layer. *Cell* **180**, 348–358.e15 (2020).

59. Wan, W. & Briggs, J. A. G. Cryo-Electron Tomography and Subtomogram Averaging. *Methods Enzymol.* **579**, 329–367 (2016).

60. Lučić, V., Rigort, A. & Baumeister, W. Cryo-electron tomography: the challenge of doing structural biology in situ. *J. Cell Biol.* **202**, 407–419 (2013).

61. Zhang, P. Advances in cryo-electron tomography and subtomogram averaging and classification. *Curr. Opin. Struct. Biol.* **58**, 249–258 (2019).

62. Turoňová, B., Schur, F. K. M., Wan, W. & Briggs, J. A. G. Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4Å. *J. Struct. Biol.* **199**, 187–195 (2017).

63. Hadani, R. & Singer, A. Representation theoretic patterns in three dimensional Cryo-Electron Microscopy I: The intrinsic reconstitution algorithm. *Ann. Math.* **174**, 1219 (2011).

64. Gomez-Blanco, J., Kaur, S., Ortega, J. & Vargas, J. A robust approach to ab initio cryo-electron microscopy initial volume determination. *J. Struct. Biol.* **208**, 107397 (2019).

65. Penczek, P. A. Fundamentals of three-dimensional reconstruction from projections. *Methods Enzymol.* **482**, 1 (2010).

66. Bai, X. C., McMullan, G., & Scheres, S. H. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40**, 49–57 (2015).

67. Frank, J. Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. *Oxford university press* (2006).

68. Franken, L. E. *et al.* A Technical Introduction to Transmission Electron Microscopy for Soft-

Matter: Imaging, Possibilities, Choices, and Technical Developments. *Small* **16**, 1906198 (2020).

69. Naydenova, K., Jia, P. & Russo, C. J. Cryo-EM with sub-1 Å specimen movement. *Science* **370**, 223–226 (2020).

70. Zhou, H. *et al.* Programming Conventional Electron Microscopes for Solving Ultrahigh-Resolution Structures of Small and Macro-Molecules. *Anal. Chem.* **91**, 10996–11003 (2019).

71. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).

72. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *Elife* **4**, e06980 (2015).

73. Glaeser, R. M. Limitations to significant information in biological electron microscopy as a result of radiation damage. *J. Ultrastruct. Res.* **36**, 466–482 (1971).

74. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).

75. Scheres, S. H. w. Beam-induced motion correction for sub-megadalton cryo-EM particles. *Elife* **3**, e03665 (2014).

76. Campbell, M. G. *et al.* Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *Structure* **20**, 1823–1828 (2012).

77. Abrishami, V. *et al.* Alignment of direct detection device micrographs using a robust Optical Flow approach. *J. Struct. Biol.* **189**, 163–176 (2015).

78. Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., ... & Cheng, Y. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods 2013 106* **10**, 584–590 (2013).

79. Wu, S., Armache, J.-P. & Cheng, Y. Single-particle cryo-EM data acquisition by using direct

electron detection camera. *Microscopy* **65**, 35 (2016).

80. Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. W. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife* **2013**, (2013).

81. Rubinstein, J. L. & Brubaker, M. A. Alignment of cryo-EM movies of individual particles by optimization of image translations. *J. Struct. Biol.* **192**, 188–195 (2015).

82. Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., ... & Cheng, Y. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).

83. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *Elife* **4**, (2015).

84. Zheng, S. Q. *et al.* MotionCor2 - anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331 (2017).

85. Kimanius, D., Dong, L., Sharov, G., Nakane, T. & Scheres, S. H. W. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochem. J.* **478**, 4169–4185 (2021).

86. Ripstein, Z. A., & Rubinstein, J. L. Processing of Cryo-EM Movie Data. *Methods Enzymol.* **579**, 103–124 (2016).

87. Erickson, H. P., & Klug, A. Measurement and compensation of defocusing and aberrations by Fourier processing of electron micrographs. *Philos. Trans. R. Soc. London. B, Biol. Sci.* **261**, 105–118 (1971).

88. Wade, R. H. A brief look at imaging and contrast transfer. *Ultramicroscopy* **46**, 145–156 (1992).

89. Van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).

90. Wiener, N., Wiener, N., Mathematician, C., Wiener, N., Wiener, N., & Mathématicien,

C. Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications. *Cambridge, MA: MIT press* **113**, 1043-54 (1949).

91. Scheres, S. H. W., Núñez-Ramírez, R., Sorzano, C. O. S., Carazo, J. M. & Marabini, R. Image processing for electron microscopy single-particle analysis using XMIPP. *Nat. Protoc.* **3**, 977 (2008).

92. Wang, F. *et al.* DeepPicker: a Deep Learning Approach for Fully Automated Particle Picking in Cryo-EM. *J. Struct. Biol.* **195**, 325–336 (2016).

93. Zhu, Y., Ouyang, Q. & Mao, Y. A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC Bioinforma.* **18**, 1–10 (2017).

94. Heimowitz, A., Andén, J., & Singer, A. APPLE picker: Automatic particle picking, a low-effort cryo-EM framework. *J. Struct. Biol.* **204**, 215–227 (2018).

95. Zhang Software - MRC Laboratory of Molecular Biology. https://www2.mrc-lmb.cam.ac.uk/research/locally-developed-software/zhang-software/.

96. Hoang, T. V, Cavin, X., Schultz, P. & Ritchie, D. W. gEMpicker: a highly parallel GPU-accelerated particle picking tool for cryo-electron microscopy. *BMC Struct. Biol.* **13**, 1–10 (2013).

97. Voss, N. R., Yoshioka, C. K., Radermacher, M., Potter, C. S., & Carragher, B. DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. *J. Struct. Biol.* **166**, 205–213 (2009).

98. Wagner, T. *et al.* SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun. Biol.* **2**, (2019).

99. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* , 779–788 (2015).

100. Heel, M. van & Stöffler-Meilicke, M. Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope. *EMBO J.* **4**, 2389 (1985).

101. Sorzano, C. O. S. *et al.* A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.* **171**, 197 (2010).

102. Scheres, S. H., Valle, M., Nuñez, R., Sorzano, C. O., Marabini, R., Herman, G. T., & Carazo, J. M. Maximum-likelihood multi-reference refinement for electron microscopy images. *J. Mol. Biol.* **348**, 139–149 (2005).

103. Scheres, S. H., Valle, M., Nuñez, R., Sorzano, C. O., Marabini, R., Herman, G. T., & Carazo, J. M. Maximum-likelihood multi-reference refinement for electron microscopy images. *J. Mol. Biol.* **348**, 139–149 (2005).

104. Sigworth, F. J., Doerschuk, P. C., Carazo, J.-M. & Scheres, S. H. W. Maximum-likelihood methods in cryo-EM. Part I: theoretical basis and overview of existing approaches. *Methods Enzymol.* **482**, 263 (2010).

105. Kimanius, D., Forsberg, B. O., Scheres, S. H. W. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUS in RELION-2. *Elife* **5**, (2016).

106. Sorzano, C. O. S. *et al.* A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.* **171**, 197–206 (2010).

107. Yang, Z., Fang, J., Chittuluru, J., Asturias, F. J. & Penczek, P. A. Iterative Stable Alignment and Clustering of 2D Transmission Electron Microscope Images. *Structure* **20**, 237 (2012).

108. Heel, M. van *et al.* Multivariate Statistical Analysis of Large Datasets: Single Particle Electron Microscopy. *Open J. Stat.* **6**, 701–739 (2016).

109. Van Heel, M., & Frank, J. Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy* **6**, 187–194 (1981).

110. Heel, M. Classification of very large electron microscopical image data sets. *undefined* (1989).

111. Van Heel, M. Multivariate statistical classification of noisy images (randomly oriented biological macromolecules). *Ultramicroscopy* **13**, 165–183 (1984).

112. Frank, J., Verschoor, A., & Boublik, M. Computer averaging of electron micrographs of 40S ribosomal subunits. *Science* **214**, 1353–1355 (1981).

113. Frank, J. Classification of macromolecular assemblies studied as 'single particles'. *Q. Rev. Biophys.* **23**, 281–329 (1990).

114. Bartesaghi, A., Matthies, D., Banerjee, S., Merk, A. & Subramaniam, S. Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *Proc. Natl. Acad. Sci.* **111**, 11709–11714 (2014).

115. Jacobsen, C. Relaxation of the Crowther criterion in multislice tomography. *Opt. Lett.* **43**, 4811 (2018).

116. Sorzano, C. O. S. *et al.* A Survey of the Use of Iterative Reconstruction Algorithms in Electron Microscopy. *Biomed Res. Int.* **2017**, (2017).

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

118. Zhao, S. R., & Halling, H. A new Fourier method for fan beam reconstruction. *IEEE Nuclear Science Symposium and Medical Imaging Conference Record* **2**, 1287-1291 (1995).

119. Barnett, A., Greengard, L., Pataki, A. & Spivak, M. Rapid solution of the cryo-EM reconstruction problem by frequency marching. *SIAM J. Imaging Sci.* **10**, 1170–1195 (2017).

120. Gomez-Blanco, J., Kaur, S., Ortega, J. & Vargas, J. A robust approach to ab initio cryo-electron microscopy initial volume determination. *J. Struct. Biol.* **208**, 107397 (2019).

121. Kimanius, D., Forsberg, B. O., Scheres, S. H., & Lindahl, E. Accelerated cryo-EM structure

determination with parallelisation using GPUs in RELION-2. *elife* **5**, e18722 (2016).

122. Punjani, A., Rubinstein, J. L., Fleet, D. J., & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature methods* **14**, 290-296 (2017).

123. Vargas, J., Alvarez-Cabrera, A. L., Marabini, R., Carazo, J. M., & Sorzano, C. O. S. Efficient initial volume determination from electron microscopy images of single particles. *Bioinformatics* **30**, 2891-2898 (2014).

124. Lyumkis, D., Vinterbo, S., Potter, C. S., & Carragher, B. Optimod–an automated approach for constructing and optimizing initial models for single-particle electron microscopy. *Journal of structural biology* **184**, 417-426 (2013).

125. Yan, X., Dryden, K. A., Tang, J., & Baker, T. S. (2007). Ab initio random model method facilitates 3D reconstruction of icosahedral particles. *Journal of structural biology* **157**, 211-225 (2007).

126. Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., & Ludtke, S. J. EMAN2: an extensible image processing suite for electron microscopy. *Journal of structural biology* **157**, 38-46 (2007).

127. Sanz-García, E., Stewart, A. B., & Belnap, D. M. The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry. *Journal of structural biology* **171**, 216-222 (2010).

128. Henderson, R. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18037 (2013).

129. Mao, Y. *et al.* Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12438–12443 (2013).

130. Sorzano, C. O. S. *et al.* Cryo-EM and the elucidation of new macromolecular structures: Random Conical Tilt revisited. *Sci. Reports 2015 51* **5**, 1–6 (2015).

131. Radermacher, M., Wagenknecht, T., Verschoor, A., & Frank, J. Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. *Journal of microscopy* **146,** 113-136 (1987).

132. Leschziner, A. E., & Nogales, E. The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *J. Struct. Biol.* **153**, 284–299 (2006).

133. Thuman-Commike, P. A., & Chiu, W. Improved common line-based icosahedral particle image orientation estimation algorithms. *Ultramicroscopy* **68**, 231–255 (1997).

134. Penczek, P. A., Zhu, J. & Frank, J. A common-lines based method for determining orientations for N > 3 particle projections simultaneously. *Ultramicroscopy* **63**, 205–218 (1996).

135. Ogura, T. & Sato, C. A fully automatic 3D reconstruction method using simulated annealing enables accurate posterioric angular assignment of protein projections. *J. Struct. Biol.* **156**, 371–386 (2006).

136. Liu, X., Jiang, W., Jakana, J., & Chiu, W. Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a Multi-Path Simulated Annealing optimization algorithm. *J. Struct. Biol.* **160**, 11–27 (2007).

137. Elmlund, D., Davis, R., & Elmlund, H. Ab initio structure determination from electron microscopic images of single molecules coexisting in different functional states. *Structure* **18**, 777–786 (2010).

138. Elmlund, D., & Elmlund, H. SIMPLE: Software for ab initio reconstruction of heterogeneous single-particles. *J. Struct. Biol.* **180**, 420–427 (2012).

139. Crowther, R., Amos, L. A., Finch, J. T., De Rosier, D. J., & Klug, A. Three Dimensional Reconstructions of Spherical Viruses by Fourier Synthesis from Electron Micrographs. *Nature* **226**, 421–425 (1970).

140. Castón, J. R., Belnap, D. M., Steven, A. C., & Trus, B. L . A strategy for determining the orientations of refractory particles for reconstruction from cryo-electron micrographs with particular reference to round, smooth-surfaced, icosahedral viruses. *J. Struct. Biol.* **125**, 209–215 (1999).

141. Vargas, J., Alvarez-Cabrera, A. L., Marabini, R., Carazo, J. M., & Sorzano, C. O. S. Efficient initial volume determination from electron microscopy images of single particles. *Bioinformatics* **30**, 2891–2898 (2014).

142. Sorzano, C. O. S. *et al.* A statistical approach to the initial volume problem in Single Particle Analysis by Electron Microscopy. *J. Struct. Biol.* **189**, 213–219 (2015).

143. Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods* **13**, 24 (2016).

144. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).

145. Sigworth, F. J., Doerschuk, P. C., Carazo, J. M. & Scheres, S. H. W. An introduction to maximum-likelihood methods in cryo-EM. *Methods Enzymol.* **482**, 263–294 (2010).

146. Scheres, S. H. W. Classification of structural heterogeneity by maximum-likelihood methods. *Methods Enzymol.* **482**, 295–320 (2010).

147. Kurkcuoglu, Z., Bahar, I. & Doruker, P. ClustENM: ENM-Based Sampling of Essential Conformational Space at Full Atomic Resolution. *J. Chem. Theory Comput.* **12**, 4549–4562 (2016).

148. Gur, M., Madura, J. D. & Bahar, I. Global Transitions of Proteins Explored by a Multiscale Hybrid Methodology: Application to Adenylate Kinase. *Biophys. J.* **105**, 1643 (2013).

149. Costa, M. G. S., Batista, P. R., Bisch, P. M. & Perahia, D. Exploring Free Energy Landscapes of

Large Conformational Changes: Molecular Dynamics with Excited Normal Modes. *J. Chem. Theory Comput.* **11**, 2755–2767 (2015).

150. Haselbach, D., Komarov, I., Agafonov, D. E., Hartmuth, K., Graf, B., Dybkov, O. *et al.* Structure and Conformational Dynamics of the Human Spliceosomal B act Complex. *Cell* **172**, 454–464 (2018).

151. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

152. Sanchez Sorzano, C. O., Alvarez-Cabrera, A. L., Kazemi, M., Carazo, J. M. & Jonić, S. StructMap: Elastic Distance Analysis of Electron Microscopy Maps for Studying Conformational Changes. *Biophys. J.* **110**, 1753 (2016).

153. Penczek, P. A. International Tables for Crystallography, Vol. B, edited by U. Shmueli. *New York Springer. Penczek, PA (2010). Methods Enzym.* **482**, 73–100 (2008).

154. Scheres, S. H. W. Classification of Structural Heterogeneity by Maximum-Likelihood Methods. *Methods Enzymol.* **482**, 295–320 (2010).

155. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

156. Radermacher, M. Three-Dimensional reconstruction of single particles from random and nonrandom tilt series. *J. Electron Microsc. Tech.* **9**, 359–394 (1988).

157. Penczek, P. A., Grassucci, R. A. & Frank, J. The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. *Ultramicroscopy* **53**, 251–270 (1994).

158. Sorzano, C. O. S. *et al.* A new algorithm for high-resolution reconstruction of single particles by electron microscopy. *J. Struct. Biol.* **204**, 329–337 (2018).

159. Penczek, P. A. Resolution measures in molecular electron microscopy. *Methods Enzymol.* **482**, 73–100 (2010).

160. Penczek, P. A. Resolution measures in molecular electron microscopy. *Methods Enzymol.* **482**, 73 (2010).

161. Rosenthal, P. B., & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

162. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

163. Stewart, A. & Grigorieff, N. Noise bias in the refinement of structures derived from single particles. *Ultramicroscopy* **102**, 67–84 (2004).

164. Grigorieff, N. Resolution measurement in structures derived from single particles. *Acta Crystallogr. D. Biol. Crystallogr.* **56**, 1270–1277 (2000).

165. Henderson, R. *et al.* Outcome of the First Electron Microscopy Validation Task Force Meeting. *Struct. England1993)* **20–330**, 205 (2012).

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

167. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

168. Lawson, C. L. *et al.* EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.* **44**, 396–403 (2016).

169. Henderson, R., Chen, S., Chen, J. Z., Grigorieff, N., Passmore, L. A., Ciccarelli, L. *et al.* Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. *J. Mol. Biol.* **413**, 1028–1046 (2011).

170. Rosenthal, P. B. & Henderson, R. Optimal Determination of Particle Orientation, Absolute Hand,

and Contrast Loss in Single-particle Electron Cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

171. Russo, C. J. & Passmore, L. A. Robust evaluation of 3D electron cryomicroscopy data using tilt-pairs. *J. Struct. Biol.* **187**, 112–118 (2014).

172. Wasilewski, S., & Rosenthal, P. B. Web server for tilt-pair validation of single particle maps from electron cryomicroscopy. *J. Struct. Biol.* **186**, 122–131 (2014).

173. Vargas, J., Melero, R., Gómez-Blanco, J., Carazo, J. M. & Sorzano, C. O. S. Quantitative analysis of 3D alignment quality: its impact on soft-validation, particle pruning and homogeneity analysis. *Sci. Reports* **7**, 1–14 (2017).

174. Vargas, J., Otón, J., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. Particle alignment reliability in single particle electron cryomicroscopy: a general approach. *Sci. Reports* **6**, 1–11 (2016).

175. Heymann, J. B. Validation of 3D EM Reconstructions: The Phantom in the Noise. *AIMS Biophys.* **2**, 21–35 (2015).

176. Saxton, W. O. & Baumeister, W. The correlation averaging of a regularly arranged bacterial cell envelope protein. *J. Microsc.* **127**, 127–138 (1982).

177. Frank, J., Verschoor, A., & Boublik, M. Computer averaging of electron micrographs of 40S ribosomal subunits. *Science* **214**, 1353–1355 (1981).

178. Penczek, P. A. Three-dimensional spectral signal-to-noise ratio for a class of reconstruction algorithms. *J. Struct. Biol.* **138**, 34–46 (2002).

179. Unser, M., Trus, B. L. & Steven, A. C. A new resolution criterion based on spectral signal-to-noise ratios. *Ultramicroscopy* **23**, 39–52 (1987).

180. Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184**, 226–236 (2013).

181. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM

density maps. *Nat. Methods* **11**, 63–65 (2014).

182. Vilas, J. L. *et al.* MonoRes: Automatic and Accurate Estimation of Local Resolution for Electron Microscopy Maps. *Structure* **26**, 337–344 (2018).

183. Ramírez-Aportela, E., Mota, J., Conesa, P., Carazo, J. M. & Sorzano, C. O. S. DeepRes: a new deep-learning- and aspect-based local resolution method for electron-microscopy maps. *IUCrJ* **6**, 1054–1063 (2019).

184. Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* **161**, 438 (2015).

185. Deptuch, G., Besson, A., Rehak, P., Szelezniak, M., Wall, J., Winter, M., & Zhu, Y. Direct electron imaging in electron microscopy with monolithic active pixel sensors. *Ultramicroscopy* **107**, 674–684 (2007).

186. Faruqi, A. R., Cattermole, D. M. & Raeburn, C. Direct electron detection methods in electron microscopy. *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.* **513**, 317–321 (2003).

187. Jin, L., Milazzo, A. C., Kleinfelder, S., Li, S., Leblanc, P., Duttweiler, F. *et al.* Applications of direct detection device in transmission electron microscopy. *J. Struct. Biol.* **161**, 352–358 (2007).

188. Marriott, S. J., Jose, S. & Stevenson, R. L. Exhibition 2D Display Technologies 3D Displays and Holography Image and Document Visualization Image Processing Sensors, Capture, and Machine Vision Multimedia Processing and Applications Visual Communications and Image Processing Optical Security and Anti-Counterfeiting Advance Technical Program. 18–22 (2004).

189. Wu, S., Armache, J.-P. & Cheng, Y. Single-particle cryo-EM data acquisition by using direct electron detection camera. *Microscopy* **65**, 35 (2016).

190. Wu, S., Armache, J. P. & Cheng, Y. Single-particle cryo-EM data acquisition by using direct

electron detection camera. *Microscopy* **65**, 35 (2016).

191. Bammes, B. E., Chen, D.-H., Jin, L. & Bilhorn, R. B. Visualizing and Correcting Dynamic Specimen Processes in TEM Using a Direct Detection Device. *Microsc. Microanal.* **19**, 1320–1321 (2013).

192. Bhella, D. Cryo-electron microscopy: an introduction to the technique, and considerations when working to establish a national facility. *Biophys. Rev.* **11**, 515–519 (2019).

193. Joel Ltd. *Glossary of TEM Terms*. [online] Jeol.co.jp. Available at: <https://www.jeol.co.jp/en/words/emterms/search_result.html?keyword=electron gun, JEOL Ltd (1996-2022)>.

194. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

195. Cheng, A. *et al.* Leginon: New features and applications. *Protein Sci.* **30**, 136–150 (2021).

196. Latitude D Software | Gatan, Inc. https://www.gatan.com/products/tem-imaging-spectroscopy/latitude-d-software.

197. Zhang, J. *et al.* JADAS: A Customizable Automated Data Acquisition System and its Application to Ice-embedded Single Particles. *J. Struct. Biol.* **165**, 1 (2009).

198. EPU | EM Software | Single Particle Analysis - CA.

199. Li, X., Zheng, S., Agard, D. A. & Cheng, Y. Asynchronous data acquisition and on-the-fly analysis of dose fractionated cryoEM images by UCSFImage. *J. Struct. Biol.* **192**, 174 (2015).

200. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).

201. Li, X., Grigorieff, N. & Cheng, Y. GPU-enabled FREALIGN: Accelerating single particle 3D reconstruction and refinement in Fourier space on graphics processors. *J. Struct. Biol.* **172**, 407–

412 (2010).

202. Wandzik, J. M. *et al.* A Structure-Based Model for the Complete Transcription Cycle of Influenza Polymerase. *Cell* **181**, 877–893.e21 (2020).

203. Ge, P. *et al.* Action of a minimal contractile bactericidal nanomachine. *Nature* **580**, 658–662 (2020).

204. Bass, R. B., Strop, P., Barclay, M. & Rees, D. C. Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel. *Science* **298**, 1582–1587 (2002).

205. Jacobson, R. A., Wunderlich, J. A., Lipscomb, W. N. & IUCr. The crystal and molecular structure of cellobiose. *Acta Crystallographica* **14**, 598–607 (1961).

206. Wilson, A. J. C. Determination of Absolute from Relative X-Ray Intensity Data. *Nature* **150**, 152–152 (1942).

207. Singer, A. Wilson statistics: derivation, generalization and applications to electron cryomicroscopy. *Acta Crystallogr. Sect. A, Found. Adv.* **77**, 472–479 (2021).

208. Sherwood, D. & Cooper, J. Crystals, X-rays and proteins: comprehensive protein crystallography. *OUP Oxford* (2010).

209. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

210. Terwilliger, T. C., Sobolev, O. V., Afonine, P. V. & Adams, P. D. Automated map sharpening by maximization of detail and connectivity. *urn:issn:2059-7983* **74**, 545–559 (2018).

211. Larkin, K. G., Bone, D. J., & Oldfield, M. A. Natural demodulation of two-dimensional fringe patterns. I. General background of the spiral phase quadrature transform. *JOSA A* **18**, 1862-1870 (2001).

212. Vargas, J., Quiroga, J. A., Sorzano, C. O. S., Estrada, J. C. & Servín, M. Multiplicative phase-

shifting interferometry using optical flow. *Appl. Opt.* **51**, 5903–5908 (2012).

213. Sorzano, C. O. S., Vargas, J., Quiroga, J. A., Estrada, J. C. & Carazo, J. M. Two-step interferometry by a regularized optical flow algorithm. *Opt. Lett.* **36**, 3485–3487 (2011).

214. Vargas, J., Restrepo, R., Quiroga, J. A. & Belenguer, T. High dynamic range imaging method for interferometry. *Opt. Commun.* **284**, 4141–4145 (2011).

215. Vargas, J. *et al.* Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *J. Struct. Biol.* **183**, 342–353 (2013).

216. Vargas, J. *et al.* FASTDEF: Fast defocus and astigmatism estimation for high-throughput transmission electron microscopy. *J. Struct. Biol.* **181**, 136–148 (2013).

217. Vilas, J. L., Tagare, H. D., Vargas, J., Carazo, J. M. & Sorzano, C. O. S. Measuring local-directional resolution and local anisotropy in cryo-EM maps. *Nat. Commun.* **11**, 1–7 (2020).

218. Felsberg, M., & Sommer, G. The monogenic signal. *IEEE transactions on signal processing* **49**, 3136-3144 (2001).

219. Unser, M., & Van De Ville, D. Wavelet steerability and the higher-order Riesz transform. *IEEE Transactions on Image Processing* **19**, 636-652 (2009).

220. Unser, M., Sage, D., & Van De Ville, D. Multiresolution monogenic signal analysis using the Riesz–Laplace wavelet transform. *IEEE Transactions on Image Processing* **18**, 2402-2418 (2009).

221. Dashti, A. *et al.* Trajectories of the ribosome as a Brownian nanomachine. *Proc. Natl. Acad. Sci.* **111**, 17492–17497 (2014).

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

223. Giraldo-Barreto, J. *et al.* A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Sci. Reports* **11**, 1–15 (2021).

224. Mallamace, F. *et al.* Transport properties of glass-forming liquids suggest that dynamic crossover temperature is as important as the glass transition temperature. *Proc. Natl. Acad. Sci.* **107**, 22457–22462 (2010).

225. Yip, S. & Short, M. P. Multiscale materials modelling at the mesoscale. *Nat. Mater.* **12**, 774–777 (2013).

226. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. Navigating the folding routes. *Science* **267**, 1619–1620 (1995).

227. Neupane, K., Manuel, A. P. & Woodside, M. T. Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes. *Nat. Phys.* **12**, 700–703 (2016).

228. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, 377–388 (2013).

229. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. Methods 2021 182* **18**, 176–185 (2021).

230. Haselbach, D. *et al.* Long-range allosteric regulation of the human 26S proteasome by 20S proteasome-targeting cancer drugs. *Nat. Commun.* **8**, 1–8 (2017).

231. De la Rosa-Trevín, J. M. *et al.* Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.* **195**, 93–99 (2016).

232. Elmlund, D., & Elmlund, H. SIMPLE: Software for ab initio reconstruction of heterogeneous single-particles. *J. Struct. Biol.* **180**, 420–427 (2012).

233. Ludtke, S. J., Baldwin, P. R., & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).

234. Frank, J. *et al.* SPIDER and WEB: Processing and visualization of images in 3D electron

microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).

235. Vargas, J. *et al.* Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *Journal of Structural Biology* **183**, 342-353 (2013).

236. Vargas, J., Otón, J., Marabini, R., Jonic, S., De La Rosa-Trevín, J. M., Carazo, J. M., & Sorzano, C. O. S. FASTDEF: fast defocus and astigmatism estimation for high-throughput transmission electron microscopy. *J. Struct. Biol.* **181**, 136–148 (2013).

237. Vilas, J. L., Tagare, H. D., Vargas, J., Carazo, J. M. & Sorzano, C. O. S. Measuring local-directional resolution and local anisotropy in cryo-EM maps. *Nat. Commun. 2020 111* **11**, 1–7 (2020).

238. Vilas, J. L. *et al.* MonoRes: Automatic and Accurate Estimation of Local Resolution for Electron Microscopy Maps. *Structure* **26**, 337–344.e4 (2018).

239. Wu, X. *et al.* CryoETGAN: Cryo-Electron Tomography Image Synthesis via Unpaired Image Translation. *Front. Physiol.* **0**, 128 (2022).

240. Vargas, J., Gómez-Pedrero, J. A., Quiroga, J. A. & Alonso, J. Enhancement of Cryo-EM maps by a multiscale tubular filter. *Opt. Express* **30**, 4515 (2022).

241. Park, C. *et al.* Structural basis of neuropeptide Y signaling through Y1 receptor. *Nat. Commun.* **13**, 1–12 (2022).

242. Vilas, J. L. *et al.* Re-examining the spectra of macromolecules. Current practice of spectral quasi B-factor flattening. *J. Struct. Biol.* **209**, (2020).

243. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).

244. Chong, S.-H. & Ham, S. Folding Free Energy Landscape of Ordered and Intrinsically Disordered Proteins. *Sci. Reports* **9**, 1–9 (2019).

245. Leiman, P. G., Chipman, P. R., Kostyuchenko, V. A., Mesyanzhinov, V. V. & Rossmann, M. G. Three-Dimensional Rearrangement of Proteins in the Tail of Bacteriophage T4 on Infection of Its Host. *Cell* **118**, 419–429 (2004).

246. Challenges for cryo-EM. *Nat. Methods* **15**, 985–985 (2018).

247. Maćkiewicz, A. & Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **19**, 303–342 (1993).

248. Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. Data visualization with multidimensional scaling. *Journal of computational and graphical statistics* **17**, 444-472 (2008).

249. Qiang, Z. Internationalization of Higher Education: Towards a Conceptual Framework. *Policy futures in education* **1**, 248–270 (2003).

250. Donoho, D. L. & Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5591–5596 (2003).

251. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

252. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science (80-.).* **290**, 2319–2323 (2000).

253. Khalifa, A. A. Z. *et al.* The inner junction complex of the cilia is an interaction hub that involves tubulin post-translational modifications. *Elife* **9**, (2020).

254. Ichikawa, M. *et al.* Tubulin lattice in cilia is in a stressed form regulated by microtubule inner proteins. *Proc. Natl. Acad. Sci.* **116**, 19930–19938 (2019).

255. Yang, M. *et al.* Cryo-electron microscopy structures of ArnA, a key enzyme for polymyxin resistance, revealed unexpected oligomerizations and domain movements. *J. Struct. Biol.* **208**, 43–

50 (2019).

256. Gutmann, T. *et al.* Cryo-EM structure of the complete and ligand-saturated insulin receptor ectodomain. *J. Cell Biol.* **219**, (2020).

257. Jahagirdar, D. *et al.* Alternative Conformations and Motions Adopted by 30S Ribosomal Subunits Visualized by Cryo-Electron Microscopy. *bioRxiv* **26**, 2020.03.21.001677 (2020).

258. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).

259. Haselbach, D. *et al.* Long-range allosteric regulation of the human 26S proteasome by 20S proteasome-targeting cancer drugs. *Nat. Commun.* **8**, (2017).

260. Cui, Q. & Bahar, I. Normal Mode Analysis Theory and Applications. *Biomed. Eng. (NY).* 448 (2006).

261. Bahar, I., & Rader, A. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **15**, 586–592 (2005).

262. López-Blanco, J. R., Miyashita, O., Tama, F. & Chacón, P. Normal Mode Analysis Techniques in Structural Biology. *eLS* (2014).

263. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

264. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, 377–388 (2013).

265. Wu, J. *et al.* Massively parallel unsupervised single-particle cryo-EM data clustering via statistical manifold learning. *PLoS One* **12**, e0182130 (2017).

266. Sorzano COS. *et al.* A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.* **171**, 197–206 (2010).