

Improved methods for the annotation and processing
of Cryo-EM images and for atomic model validation

Mateusz Olek

Doctor of Philosophy

University of York

Chemistry

September 2022

Abstract

Cryogenic electron microscopy was introduced in the 1980s as an innovative technique to determine protein structures and has become more popular in recent years. With almost 7800 maps released last year, it is the second most used method after X-ray crystallography. The technological advancements and the constant improvements in automated data collection create a strong need for software tools to automatically assess the quality of the collected data.

In this work, we identified the parts of the cryo-EM processing pipeline, from data collection to atomic model building and validation, that can significantly benefit from automated tools. The evaluation of automated particle picking tools revealed a common issue when the picking is less effective from micrographs with non-uniform ice distribution. We developed a software tool which could mitigate this problem and equalise the contrast locally. Additionally, we introduce a new parameter for the processing, which can be used to associate the particles' coordinates with the estimated ice thickness levels, which can also be calibrated to the measured thickness. This software is added to the ISpyB data collection pipeline at the Electron Bio-Imaging Centre in Diamond Light Source.

A software tool was developed for atomic model validation based on the False Discovery Rate approach which allows scoring each residue in the model by checking if they are placed within the cryo-EM map density or in the background noise. This tool is now available from the Collaborative Computer Project for Cryo-EM (CCPEM) software suite.

Table of Contents

Abstract	2
List of Tables.....	5
List of Figures	5
Acknowledgements	8
Declaration	9
1 Introduction	10
1.1 Motivation	10
1.2 Overview	14
1.3 Thesis structure	15
2 Background	17
2.1 Cryogenic Electron Microscopy	21
2.1.1 Electron sources	21
2.1.2 Image formation	23
2.1.3 Electron detectors	28
2.1.4 Fourier space operations in cryo-EM data processing	33
2.1.5 Sample preparation.....	34
2.1.5.1 Specimen behaviour and ice thickness evaluation in the cryo-EM samples	41
2.1.6 Data collection	43
2.1.7 Single Particle Reconstruction	45
2.1.7.1 Beam-induced motion correction.....	45
2.1.7.2 Contrast Transfer Function estimation.....	47
2.1.7.3 Particle Picking	48
2.1.7.4 2D Classification.....	53
2.1.7.5 3D Reconstruction.....	54
2.1.7.6 Resolution of the Cryo-EM maps	57
2.1.8 Atomic model building and validation.....	59
3 Development of a software tool for processing the cryo-EM data with non-uniform ice	63
3.1 Introduction	63

3.2	Methods.....	65
3.2.1	Average pooling	65
3.2.2	K-Means Clustering	67
3.2.3	IceBreaker algorithm description	67
3.3	Results.....	69
3.3.1	The number of clusters and the execution time	69
3.3.2	Assessing the quality of the micrographs	71
3.3.3	Estimated and measured ice-thickness	73
3.3.4	Considerations and limitations of the proposed method for ice thickness estimation related to microscope and image collection setup.....	75
3.3.5	Ice distribution estimated from cryogenic electron tomography dataset	81
3.3.6	Mapping particle coordinates to the estimated ice thickness levels.....	83
3.3.7	Local contrast improvement based on the histogram equalization.....	84
3.3.8	Adaptive histogram equalisation for cryo-EM micrographs.....	85
3.4	Conclusions and future works.....	90
4	Implementation of the atomic model validation tool based on the False Discovery Rate approach.....	94
4.1	Introduction.....	94
4.1.1	Raw and sharpened cryo-EM maps	95
4.1.2	False Discovery Rate approach.....	96
4.2	Methods.....	97
4.2.1	Processing stages description:.....	97
4.2.2	Metrics for the evaluation of classifier performance	98
4.3	Results.....	99
4.3.1	FDR score and cryo-EM map sharpening.....	99
4.3.2	FDR score evaluation.....	103
4.3.3	FDR score and Local Resolution	105
4.3.4	Number of particles vs final resolution.....	110
4.3.5	FDR Score compared to Atom Inclusion.....	112
4.4	Conclusions and future works.....	116
5	Conclusion	118
5.1	Summary	118

5.2	Limitations and future work.....	120
5.3	Closing remarks	121
6	Bibliography.....	124
	Appendix A	
	Appendix B	

List of Tables

Table 2.1	Comparison of different methods for imaging biological samples.....	20
Table 2.2	Comparison of the parameters of different electron guns.....	22
Table 2.3	Different techniques used for the cryo-EM sample preparation.....	40
Table 2.4	Summary of validation tools used for cryo-EM atomic model evaluation.....	62
Table 3.1	Effect of different numbers of patches and clusters on the image segmentation	70
Table 3.2	Pixel intensity profiles estimated from tomography datasets	82
Table 3.3	Number of particles picked from micrographs with different defocus with and without bandpass filter	89
Table 4.1	Performance metrics for FDR score.....	103
Table 4.2	Confusion matrix for the FDR score with 0.65 threshold and 1 Å RMSD ...	104

List of Figures

Figure 2.1	Electron beam interactions with the atoms in the sample.....	23
Figure 2.2	Electron beam scattered through the sample.	25
Figure 2.3	Effect of different defocus values on the collected images and the CTF	26
Figure 2.4	A) Parameters used for astigmatism correction,.....	28
Figure 2.5	Characteristics of the detectors commonly used in cryo-EM.....	31
Figure 2.6	Different operating modes of electron sensors in cryo-EM.....	32
Figure 2.7	An overview general workflow of Single Particle Analysis with cryo-EM. .	32
Figure 2.8	Cryo-EM micrograph and it's power spectrum.....	33
Figure 2.9	cryo-EM sample support.....	36
Figure 2.10	A) cryoChip sample deposition device.....	38
Figure 2.11	A) Self-wicking nanowire grid used with Spotiton/Chameleon foil	39
Figure 2.12.	42

Figure 2.13 Different magnification levels for cryo-EM data collection	44
Figure 2.14 Contrast Transfer Function.....	48
Figure 2.15 Representation of the cryo-EM micrograph with some particles in the field of view.....	51
Figure 2.16 3D reconstruction from the 2D views obtained using Fourier slice theorem	56
Figure 2.17 Summary of the structural features recognisable at different resolution levels of cryo-EM maps	58
Figure 2.18 Beta-galactosidase map coloured according to the local resolution.....	59
Figure 2.19 Flowchart of the map optimisation, model building and validation procedure	60
Figure 3.1 Example of 2x2 kernel applied to 4x4 matrix for average pooling, with stride 2.....	66
Figure 3.2 Voronoi diagram.....	67
Figure 3.3 Segmentation parameters vs execution time	71
Figure 3.4 Micrograph quality assessment using median of pixel intensity.....	72
Figure 3.5 Relationship between the pixel intensity values from IceBreaker and ice thickness measured with energy filter.....	74
Figure 3.6 Defocus values distribution estimated with CTFFIND4	76
Figure 3.7 Comparison of pixel intensities in different ice thickness regions.....	77
Figure 3.8 A) Differences in intensity values between dose weighted and not dose weighted micrographs.....	79
Figure 3.9 Micrographs with crystalline contamination	80
Figure 3.10 Segment from clustered image used as a local mask.....	85
Figure 3.11 Effect of contrast equalization.....	86
Figure 3.12 Example of image with coarse ice gradient.....	87
Figure 3.13 Comparison of picks from a 3.2 micron underfocus micrograph.....	90
Figure 3.14 Segmentation based on the intensity for lower magnification level.....	92
Figure 4.1 FDR score from the sharpened maps.....	102
Figure 4.2 Comparison of the ROC curves for FDR score.....	104
Figure 4.3 FDR score can report high scores for some residues even if the model is incorrectly built in that region.....	105
Figure 4.4 Comparison of FDR score of the models at different local resolution reveals lower scores as the local resolution decreases	108
Figure 4.5 Comparison of FDR scores.....	109

Figure 4.6 FDR score calculated for different numbers of particles.....	112
Figure 4.7 Comparison of the FDR score and Atom Inclusion.....	113
Figure 4.8 Comparison of the FDR score and Atom Inclusion.....	114
Figure 4.9 Comparison of the FDR score and Atom Inclusion for PDB 6nbb	115

Acknowledgements

I would like to thank my supervisors, Professor K. Cowtan, Professor F. Antson from the University of York and Professor P. Zhang from the electron Bio-Imaging Centre at Diamond Light Source. With their help and guidance, I was able to solve the problems that came along the way of this project. My thanks extend to other members of my Thesis Advisory Panel, Professor R. Hubbard and Dr J. Blaza, for their critical comments and helping to set the next milestones in my research.

I would like to express my gratitude to my collaborators at Diamond Light Source, Dr Y. Chaban and Dr A.-P. Joseph for their support, staying open minded and always encouraging me to try new approaches to solving the problems. I am grateful for the opportunity to work with the bright scientists from Cowtan's group in York Structural Biology Laboratory and Zhang's group in the Division of Structural Biology at the University of Oxford. I would like to thank my colleagues from eBIC for useful discussions about cryo-electron microscopy, data processing pipeline and sharing the datasets with me. I also express my thanks to the members of the CCPEM group for introducing me to the cryo-EM model building and validation procedures.

I am grateful to the Data Analysis group at Diamond Light Source for their support and for providing access to the computing resources for this project. I also thank the High-performance Computing Team at the University of York for their help in using the HPC cluster.

I would like to thank my mother for her support, my wife Anna who always believed in me for her patience and love. Finally, I thank all my friends who supported me during this wonderful adventure.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author, with the exception of the published or collaborative work listed below added as Appendix A and Appendix B.

- Olek, M. *et al.* IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice. *Structure* **30**, 522-531 (2022)
 - Contributions to this paper: M.O., Y.C., K.C., and P.Z. conceived and designed research. M.O. developed the IceBreaker software tool and processed the data. D.W. helped with the integration of IceBreaker into Relion. M.O., Y.C., K.C., and P.Z. analyzed data and wrote the manuscript.

- Olek, M. & Joseph A.-P. Cryo-EM Map–Based Model Validation Using the False Discovery Rate Approach, *Frontiers in Molecular Biosciences* **8**, 652530 (2021)
 - Contributions to this paper: MO developed the tool, performed analysis on different examples, and drafted the manuscript. AJ and MO conceived the idea. AJ helped with analysis, implementation of the tool in CCP-EM, and manuscript writing.

This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

1 Introduction

1.1 Motivation

Proteins are molecules that can have a range of functions in the human body and other living organisms. There are tens of thousands of proteins which can have different roles[1], [2]. They can act as antibodies to protect the body from viruses and bacteria, enzymes which carry on chemical reactions in cells, or cell signalling proteins which transmit the messages between cells or as structural components, etc.[3] Antibodies can recognise an antigen (e.g. bacteria, viruses, allergen, fungi) that enters a body and can bind to it and mark it or the infected cell for the phagocytes to neutralise it[4], [5]. Enzymes are proteins specialised in controlling the chemical reactions in living organisms. They can speed up or slow down processes such as digestion, nerve function, breathing or getting rid of toxins. Enzymes have an active site which binds the substrate and facilitates reactions to convert it into the product[6]. For example, lipase breaks fats into fatty acids or amylase, which converts starch into glucose[7]. Enzymes are also used in various industrial applications such as biofuel production, food processing, and biological detergents to remove stains[8]. The function of the enzyme is defined by the chemistry and geometry of its active site. The active site is built of specific amino acids which creates a chemical environment to facilitate the reactions with substrates. It also determines the strength of substrate bonding. The shape of the active site should complement the shape of the substrate for effective binding and catalysis. Additionally, the spatial arrangement and orientation of the active side can reduce the activation energy for the reaction[9], [10]. Structural proteins are different from functional proteins, thanks to their three-dimensional shape, playing a key role in maintaining the shape and structure of cells and tissues. Usually, they have a characteristic part of a sequence that repeats and forms a higher-order structure. One of the examples of structural proteins and their functions is collagen, which is formed as a three-strand helix and is found in the body's connective tissue[11]. Keratin monomers are assembled into bundles, which form intermediate filaments in nails, hairs, hooves or horns. Actin polymerises into long, stiff fibres, which form a cytoskeleton to support the shape and size of a cell[12]. Elastin allows tissues like lungs, blood vessels or skin to come back to their original shape after stretching [11]. The building blocks of proteins are called amino acids. There are a wide range of amino acids with specialised functions. The 20 amino acids encoded in the genetic code used by all living organisms to synthesise proteins are called canonical

amino acids. Each of them consists of the central carbon (C-alpha) atom linked to the carboxyl group (COOH) and amino group (NH₂). Amino acids have the side chain- a group of atoms attached to the C-alpha atom, unique for each amino acid, except for proline, whose side chain is connected back to the amino group, creating a five-membered pyrrolidine ring and glycine, which does not have a side chain[13]. How the amino acids are sequenced and organised in space determines the protein function and how it interacts with the environment. This can be described and interpreted at different levels[14]. The primary structure of proteins simply describes how the amino acids are ordered in a linear sequence. Amino acids are linked by the covalent bond created by the dehydration synthesis reaction between the carboxyl group of one amino acid and the amino group of the next one, also called the peptide bond. As a byproduct of this reaction, a water molecule is released[15]. A sequence of amino acids linked together by the peptide bonds is called a polypeptide chain. Each polypeptide chain has a free amino acid group at one end, called N-terminus and a free carboxyl group at the other end (C-terminus). This determines the directionality of the polypeptide chain and is important for protein synthesis. The interactions between hydrogens in the protein backbone result in the specific orientation of dipoles. Another factor that contributes to stabilising the three-dimensional arrangement of amino acids is side-chain-backbone and side-chain-side-chain interactions. The side chains have different chemical properties: for example, they can have positive or negative charges, and they can be non-polar or polar but non-charged. Because of this, the side chains can bond with one another by various interactions, such as ionic bonds between charged side chains, hydrogen bonds between the polar ones, or weak van der Waals interactions between hydrophobic side chains. These interactions, their type, and their location in the sequence define how the protein chain bends and folds[16]. In some cases, due to the steric effect resulting from the spatial collision of atoms, their interactions can be limited or impossible, which also affects the folding pattern of a protein. As a result of all listed conditions, the secondary structure of the protein is defined. The secondary structure contributes to the overall 3D shape of the protein in the form of alpha-helices and beta-sheets[17]. The alpha helix is shaped as a right-handed coil. Each turn consists of 3.6 amino acids, with the rise of 5.4 Å per turn. The stability of this structure is preserved by the hydrogen bonds between the carbonyl oxygen of one amino acid and amine hydrogen of another amino acid, located four residues earlier in the sequence. The side chains of the residues forming an alpha helix are pointing outward of the helix which makes it possible for the side chain to interact with other parts of the molecule or other molecules[18]. Beta strands are sections of 3-10

amino acids in an almost linear conformation. Thanks to this, the adjacent beta strands can form hydrogen bonds between them, which leads to the formation of beta sheets. Beta sheets can be formed in a parallel (adjacent beta strands run in the same direction from the N-terminus to the C-terminus) or anti-parallel way. The stability of the beta sheets is achieved by the extensive network of the hydrogen bonds between the carbon oxygen of an amino acid in one beta strand and the amine hydrogen of amino acid in the parallel beta strand. Beta sheets are involved in forming active and bonding sites of the protein and play a key role in the formation of higher-order protein structures and intermolecular interactions[19]. Additionally, beta sheets contribute to the stability and integrity of keratin structures. The alpha helices and beta sheets are connected with shorter (typically less than 10 amino acids), less regular structures of turns, which represent sharp changes in the directions of the polypeptide chain and loops, which are more flexible and can adopt different conformations[20], [21]. The tertiary structure describes how the polypeptide chain is arranged in space. This three-dimensional structure determines the function of protein, and its ability to interact with other molecules. If a protein contains more than one polypeptide chain (e.g., haemoglobin) or more than one subunit, then the quaternary structure describes their relation in space[22]. The process of a protein adopting its functional 3-D shape is called protein folding[23].

The three-dimensional shape of the protein can be described using the Ramachandran angles[24]. The angle between nitrogen and C-alpha atoms is described by the torsion angle Phi ϕ , and the angle between C-alpha and carbon atoms Psi ψ . The use of only two angles is sufficient due to some structural restraints. The peptide bonds between nitrogen and carbonyl group are planar which limits rotations about this bond, The possible values of ϕ and ψ angles are further limited by the potential clashes between atoms[25]. The Ramachandran plot is an important tool used to visualise how the angles in the amino acid fit into the energetically allowed and possible conformational regions. It helps to identify clashes between amino acids as Ramachandran outliers are placed in disallowed regions. They are usually the result of poor data quality or model building errors in protein structure determination[26].

Understanding the protein folding phenomenon has been a vital task for structural biology since the 1950s. Through the years, different techniques have been used to determine protein structures. X-ray crystallography helped solve many of them. It requires crystallising the protein and then, with the use of X-rays, obtaining the electron-density map, which describes how the atoms are organised in the three-dimensional space[27]. Cryogenic electron microscopy (cryo-EM), began with the first reported implementation

in 1981 by A.W. McDowell and J. Dubochet when pure water was vitrified to obtain ice with glass-like state without crystallisation. Vitreous ice preserves the natural state of proteins[28]. Cryo-EM method flourished after the resolution revolution in the 2010s, when the improvements in electron detector technology and data processing algorithms opened the doors to routinely obtaining high-resolution structures[29]. The cryo-EM resolution record was broken again in 2022 with the 1.22Å apoferritin map. In 2023, almost 75% of over 7000 deposited cryo-EM maps had resolutions of 4Å or better[30]. Cryo-EM can provide information about the structure and functions of large proteins and protein systems. The understanding of cellular processes makes it possible to identify proper therapeutic interventions for many diseases. Automated tools that would allow quick and reliable processing of the cryo-EM data to obtain a high-resolution structure would be extremely useful for modern structure-based drug design.

After a protein candidate is determined to play a role in a metabolic or signalling process associated with a certain disease, it can become a drug target. Some of the desired features of a good drug target would be efficacy and safety that, apart from therapeutical effects, it would not cause side effects, druggability, which describes the possibility to modify the drug with small molecules or antibodies based on size, shape, and ability to form stable complexes with ligands, and finally the availability to be bio-marked to help monitor the therapy[31]. Knowledge of the three-dimensional structure of the biological target makes it easier to design a molecule complementary to the target's binding site and achieve the desired outcome. It can be to activate or inhibit an enzyme, open or close an ion channel, or activate or deactivate a receptor. If the structure is known, ligands that can potentially bind with the target can be selected from the database of 3D structures or be newly designed[32].

Another important factor considered in modern drug design is ligand affinity, which describes the strength of binding of the ligand to the receptor at any drug concentration, so the ligands with high affinity would require a lower dose. The selection or development of the new drug is an iterative process and often requires multiple optimisation steps[33]. Automating this process would require a reliable technique that could produce high-resolution protein-ligand structures in a short time. Modern cryo-EM, with its future development directions, seems to be a perfect candidate for this task.

One of the challenges for modern cryo-EM is the need for robust automated tools at different stages of the pipeline. Despite development of new methods and protocols, the sample preparation process often lacks reproducibility. It is hard to routinely ensure that parameters like optimal ice thickness, particle distribution, and high coverage of angular

views would be sufficient to obtain high-quality 3D reconstruction[34]. There are procedures which allow assessing the quality of the prepared samples before data collection to identify damaged areas of the grid or to measure the ice thickness by tilting the stage[35] or with the aperture limited scattering method (ALS)[36], but they are not fully automated, introduce additional steps in the data collection pipeline or cannot be performed on some of the microscopes because of the technological limitations (e.g. methods which require a microscope with energy filter). The development of better electron detectors and software tools for automated high-throughput data collection, resulting in up to 500 movies per hour and hundreds of thousands of particles from each data collection session, creates the need for additional tools for data processing to prioritise particle quality over quantity[37].

The final step of the data analysis pipeline is the atomic model validation. With the range of tools for automated model building, robust and reliable validation tools are required to check the parameters like local and global fit to map, protein backbone geometry, local secondary structure correctness and sequence tracing. Combining multiple tools for the final model validation would provide a comprehensive evaluation and spot issues which could be overseen while using only a single tool.

The cryo-EM field is constantly improved with the new algorithms developed to target the potential issues at different data processing pipeline stages to help the users make the most of their data, or identify the problems or suggest the correct experiment optimization route.

1.2 Overview

In this thesis, I will focus on different stages of cryo-EM data processing, from how the sample preparation can affect data collection and image processing up to atomic model building and validation to identify the steps which can benefit from the automated tools for data processing. We observed that the ice and protein distribution can significantly differ between the samples, as the sample preparation procedures still lack reproducibility. This can affect the data processing as the particles from different ice thicknesses have different signal-to-noise ratios. Thin ice provides a higher signal-to-noise ratio as it reduces the electron scattering more than thicker ice. Additionally, individual particles are less likely to overlap in the thinner ice regions. Unfortunately, some ice conditions may introduce preferred orientations resulting in missing views in the final map. As the cryo-EM density map is reconstructed from 2D projections of the

specimen, the most complete set of angular views is essential to obtain a high-resolution final map.

We developed a software that estimates the ice thickness and distribution based on the pixel intensity recorded by the detector, which can be run at any stage of data processing and eliminates the need for an extra step of ice measurement by tilting the stage. We propose a procedure, which can be used to calibrate the estimated ice thickness to the values obtained by using the energy filter during data collection.

A validation tool based on the False Discovery Rate approach was developed. It can be used to identify the parts of the atomic models which are misplaced or fitted into the background noise. We compare the performance of this method with other validation software.

These contributions of the thesis are summarised below:

1. **Development of a software tool for processing the cryo-EM data with non-uniform ice**

This thesis presents the IceBreaker software, which can be used to estimate the ice gradient in the cryo-EM micrographs, improve the contrast by automated removal of the gradient and annotate the particles based on a parameter referring to the ice conditions from where the particles were picked. The particles from similar ice conditions can be grouped and processed together to identify the optimal ice conditions for the data collection for a given specimen.

2. **Implementation of the model validation tool based on the False Discovery Rate approach**

This thesis presents software for model validation based on the False Discovery Rate approach which allows users to rank the residues in the atomic models based on how they fit into the cryo-electron density map. Residues with low scores can be automatically removed from the model. The tool is implemented in the CCPEM software suite.

1.3 Thesis structure

- **Chapter 2** provides an overview of basic principles of cryogenic-electron microscopy, electron sources, image formation and detectors used for data collection. The methods and techniques used in cryo-electron microscopy for sample preparation, data collection, map reconstruction, atomic model building and validation are presented.

- **Chapter 3** presents further developments on the IceBreaker software tool for cryo-EM data processing with a non-uniform background. It allows image segmentation based on local background features connected with estimated ice thickness. Particles picked from micrographs can be associated with the ice condition and grouped later based on this parameter. Ice thickness can also be calibrated to an actual measured value.
- **Chapter 4** presents additional analysis of how parameters like sharpening factors, or number of particles used for refinement affects the software tool for atomic model validation based on the False Discovery Rate algorithm. Each residue in the atomic model is ranked based on how it fits into the density or background noise. The tool was compared with the Atom Inclusion score commonly used for cryo-EM data.
- **Chapter 5** summarises the research presented in this thesis and provides an outline of possible future work.

2 Background

Knowledge of the three-dimensional structure of a biological macromolecule is the key to understanding its function and how the structural alterations can change it. Structural biology focuses on how atoms are organised in space, the interactions between them and how they form three-dimensional structures that make a molecule.

Over the years, a range of different imaging techniques and instruments were developed and used to view and analyse objects invisible to the human eye. Techniques like X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and cryogenic electron microscopy were developed to push the resolution limits achieved with light microscopy. Each method has advantages and disadvantages based on the radiation source, wavelength, characteristics, and size of the specimen[38]. Every method has a theoretical limit of the minimal distance between two points that can be distinguished, called the resolution. Resolution is related to the wavelength of the light used for imaging. The shorter length makes it possible to achieve higher resolution and investigate finer details[39]. In practice, the achievable resolution is usually smaller than the theoretical one, as it is also affected by imperfections of the imaging system, interaction of the waves with the sample, image processing procedures and other factors specific to each method. Optical microscopy uses visible light with wavelengths in the range of 390-760nm, which results in a theoretical resolution of 200nm, which limits the users to analyse the sample at most at the cellular level. Typically, the view can be magnified 1000-2000 times, and the human eyes can be used as the detector[40]. The visible light does not damage the sample, so samples require minimal preparation, which is as simple as applying the droplet of solution with the sample on the slide and placing a coverslip on top[41]. Additionally, this method allows users to observe specimen movement in real-time. Light microscopy is an easy-to-use and affordable technique as the optical lens system does not require sophisticated maintenance compared to electron microscopes. Because of the limitations in resolution and magnification, the specimen cannot be analysed at the atomic level.

X-ray crystallography measures the angles and intensities of beams diffracted in a crystal. X-rays have small wavelengths from 10 pico- to 10 nanometres, with the wavelength around 1 Å used in synchrotron facilities. The short wavelengths, which correspond to inter-atomic distance, make it possible to analyse the sample at the atomic level. Among the main limitations of this method is that the protein sample must be purified and then crystallised[42]. The protein crystal is an ordered 3D array of uniformly spaced molecules

of the same protein. The areas around the proteins are filled with water which can make up to 30-70% of the crystal. The crystallisation process parameters might differ for different samples and often involve trial and error procedures to find the right protein concentration, temperature and pH conditions. Additionally, membrane proteins would require additional stabilisation during the purification process. This can be done by adding a detergent to the sample that would create a protective layer around hydrophobic areas and conserve it in a near-native state, but this can later affect the crystal formation by introducing micelle areas between the proteins in the crystal array[43]. The crystal is exposed to the X-ray beam to produce a diffraction pattern, which is a 2D representation of how the X-rays interfered with the crystal's proteins. The diffracted X-rays also interfere with each other in a constructive or destructive way, resulting in changes in recorded intensity. The diffraction pattern carries information about the intensity of the diffracted X-rays as a function of the diffraction angle. Another challenge is the 'phase problem' arising from the fact that only the intensity of diffraction spots can be measured. The phase information can be reproduced using some computational methods. Initially, for the high-resolution datasets ($<1.2 \text{ \AA}$), a set of theoretical set of initial phases can be used based on the relationship between the spot intensities[44]. Molecular Replacement can be used to reproduce the phases based on a similar, already-known structure, but this can introduce bias[45]. Another approach includes the introduction of a heavy atom into the crystal structure and comparing the diffraction pattern with the original one. The differences in intensities help to identify the heavy atom position and calculate the initial phases. The information about amplitude and the phases is used to calculate the structure factors which represent the Fourier transform of the electron density in the crystal. This information can be used to reproduce the 3D map of the position of electrons in the structure and to fit an atomic model according to the protein sequence. The X-rays used for imaging damage the protein of interest by breaking the atomic bonds but also deteriorate the crystal itself, lowering its diffraction abilities over exposure time[46].

Another method used to analyse the structure of proteins is Nuclear Magnetic Resonance is used to analyse resonance frequencies of nuclei in a magnetic field, which provides structural information at the atomic level. The method is considered as not damaging the sample as long as the specimen is not sensitive to radiofrequency radiation. The radiation power, pulse intensities and sequences should be optimised to avoid magnetic field drift and artefacts in the spectrum. The sample is prepared as a solution of specimen in a deuterated solvent, which helps to maintain a constant magnetic field, and then placed in special high-quality NMR tubes[47]. The best results are achieved for small molecules,

up to 30 kDa, as with the larger molecules, the number of resonance signals increases and starts to overlap, which makes it challenging to identify single peaks. The quality of a reconstruction depends on the sensitivity of a spectrometer[48], [49].

Cryo-EM uses electrons accelerated by the difference in electric potential for imaging. With their dual nature, the electrons can act both as particles and as waves. The wavelength depends on the accelerating voltage. In cryo-EM, the commonly used voltages and corresponding wavelengths are 100 keV for four pico-meters wavelength and 300 keV for two pico-meters[50]. This range of wavelengths should theoretically allow for a resolution well below 1Å. However, in practical applications, several other factors, such as sample quality and radiation damage, limit the final resolution of a reconstructed 3D map. The sample preparation includes applying the solution with protein on a metal mesh grid. After the excessive amount of solution is blotted from the grid, the sample is rapidly frozen in liquid ethane to maintain the near-native biological state of the specimen and avoid the formation of crystalline ice. Ideally, a single layer of the protein particles with many different angular orientations would be suspended in thin vitreous ice. High angular coverage allows users to obtain as many 2D projections of the molecule as possible and perform 3D map reconstruction. The vitreous ice provides a uniform background and improves contrast. Unfortunately, the process of sample preparation is still highly random and non-reproducible. Users do not have much control over the final particle distribution and ice thickness[51]. Recent developments in cryo-EM sample preparation procedures and devices are presented in the next sections of this thesis. The data is collected as a series of 2D images of grid sections with a sample on it. From these images, 2D views of the protein are extracted and combined to obtain a 3D specimen reconstruction. With the new data processing methods developed in recent years, especially with the rise of Artificial Intelligence, a lot of stages of the processing pipeline are still not fully automated and often depend on the user's decision i.e. about discarding low-quality particle views or selecting reference models for 2D and 3D analysis[52]. Table 2.1 shows a comparison of different methods used in structural biology.

Table 2.1 Comparison of different methods for imaging biological samples.

Method	Light microscopy	X-ray crystallography	Nuclear Magnetic Resonance	Cryogenic Electron Microscopy
Radiation source	Visible light	X-ray	Radiofrequency waves	Electron beam
Resolution range	200 nm	0.5-2 Å	0.5-1 Å	1.9-3 Å
Sample	Living sample	Purified and crystallized protein	Protein in deuterated solvent placed in a specialised NMR tube	Protein in the solution, applied to the metal grid and rapidly frozen
Collected data	Magnified image visible to the human eye or a camera	X-ray diffraction pattern	NMR spectra representing absorption of radiation by atomic nuclei	2D micrographs of the sample showing projections of the specimen

This section of the thesis introduces the details of cryogenic-electron microscopy, including the sources of the electron beam, the process of image formation and the types of detectors used for data collection. The stages of Single-Particle Analysis (SPA) are described, starting with sample preparation and data collection. Then, the major steps in the data processing pipeline are described, including methods which allow users to process 2D images to obtain a three-dimensional map that can be used to fit the atomic model.

2.1 Cryogenic Electron Microscopy

Cryo-EM became one of the major techniques, alongside X-ray crystallography and Nuclear Magnetic Resonance spectroscopy, for determining protein structures. Accelerated electrons in the microscope column pass through the sample and interact with it to create the final image on the detector.

2.1.1 Electron sources

Typically, in electron microscopy, the source of the electrons is called an electron gun. The electrons are released from the solid surface in the vacuum as the energy is applied to overcome the work function. The work function determines the minimal energy required to release the electron from the surface in a vacuum, and it depends on the materials used. Preferably, materials with lower work function should be used as they require less energy to release the electrons[53], [54]. The electron beam should be spatially and temporally coherent. Temporal coherence means all emitted electrons have the same energy and, therefore, the same speed and wavelength. This way, all of them can be focused by the lens system on the same imaging plane along the optical axis of the microscope column. Spatial coherence means that all emitted electrons come from the same direction and go through the sample at the same angle to produce a sharp image[55]. Two main types of electron guns used in cryo-EM are thermionic and field-emission gun (FEG). The thermionic gun usually consists of a hairpin-shaped tungsten that emits electrons when heated up to 2800K with current applied to it, with the required vacuum of 10^{-3} Pa. It is the least expensive and easy to replace compared to other types of electron guns[56]. Another electron source used for thermionic emission is a crystal of lanthanum hexaboride (LaB₆). The sharp tip of the crystal helps to improve the spatial coherence compared to the Tungsten wire. It can operate at a lower temperature of 1700K. Low coherence is one of the disadvantages of thermionic guns, as the electrons can be emitted in different directions from the tip of the gun. To alleviate that issue, an additional element called the Wehnelt cylinder is located below the gun and acts as an anode, bringing the emitted electrons into the convergence point. Another type of electron source commonly used in microscopy is Field Emission Gun (FEG). Instead of heating up the gun material to high temperatures, electrons are pulled out from the tip by the electric field located below the gun. The very sharp tip of the FEG helps to achieve better coherence. FEG can also be assisted by heating the gun material. The approach that uses only an electric field is called cold FEG, operating in as low temperatures as 300K but requires a high vacuum of the order of 10^{-8} Pa[57]. A major disadvantage of this solution is the costs of operation

and complicated replacement procedure resulting in microscope downtime. The operating parameters of different electron guns are summarised in Table 2.2.

Table 2.2 Comparison of the parameters of different electron guns.

	Tungsten filament	LaB6 crystal	Field Emission Gun
Operating temperature [K]	2700	1700	300
Beam crossover size [μm]	50	10	0.01
Beam stability [%/h]	1	1	5
Required vacuum [Pa]	10^{-2}	10^{-4}	10^{-8}
Lifetime [h]	100	500	>1000

There are also other parameters of the electron beam that must be set for the data collection, which affect the quality of the reconstructed 3D map. Flux is the number of electrons per unit area in a unit of time. The dose describes how much energy is deposited in the specimen measured in units of energy per unit volume. The total dose is calculated as a product of flux and exposure time. If the flux of $5 \text{ e}/\text{A}^2/\text{s}$ is applied for 5 seconds, the total dose would be $25\text{e}/\text{A}^2$. Higher flux used over a shorter time can provide better contrast but also leads to increased radiation damage compared to lower flux over a longer exposure time, even if the total dose is the same. Typically, in cryo-EM data collection total doses between $15\text{-}60 \text{ e}^-/\text{A}^2$ are used. The total dose can be fractioned during the data collection by splitting it into multiple frames. The radiation damages high-resolution features first as the dose accumulates over collected frames. The early frames of the collected movie have low contrast but contain high-resolution information. As the contrast improves in the next frames, the high-frequency information is lost[58]. To improve the signal-to-noise ratio and compensate for increasing radiation damage, the frames are dose-weighted. This approach gives weights to different frequency bands of the signal. The weight of the high-frequency information, which is high in the early frames, decreases in the later frames as the high-frequency features are damaged. The algorithms for dose-weighting are routinely used in the cryo-EM data processing pipelines at the stage when the frames are averaged and combined[59]. Some studies also reported that the first few of the collected frames have significantly lowered high-resolution information. As the exact reasons of this phenomenon are not yet fully explored, the high-frequency information from those frames is also down-weighted[60].

2.1.2 Image formation

Cryogenic-electron microscopy is a field of transmission electron microscopy. This means that the image is created as a result of the interference between the unscattered and scattered electrons that go through the sample. Additionally, the whole area of interest is illuminated at the same time by a spread electron beam, as opposed to scanning transmission electron microscopy, where the focused electron beam is rastered over the sample. The image is then recorded by a detector with a sensor made of a 2D array of pixels.

The image in cryo-EM is created as the electron beam travels through the sample and interacts with it. In principle, there are three possible outcomes of these interactions as represented in Figure 2.1. The electron beam can be unchanged, scattered elastically or inelastically. The electrons with unchanged direction and energy make up to 80% of total events and would serve as a reference in contrast formation. For the thin biological samples imaged with a 300 keV electron beam, the probability of inelastic and elastic scattering events is 3:1, which means for each scattering event useful for image formation there are three damaging events[61].

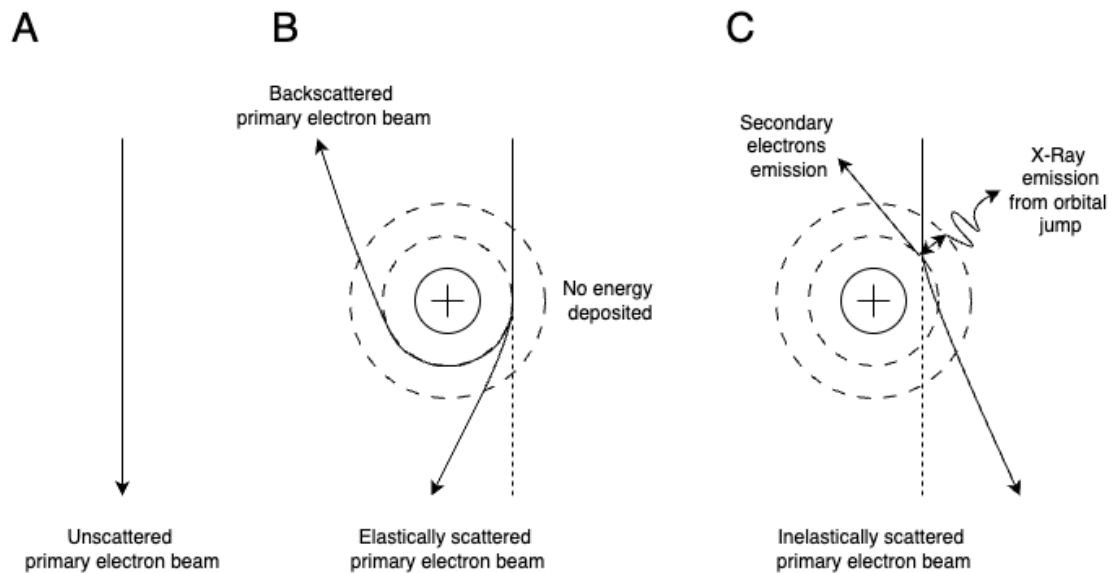


Figure 2.1 Electron beam interactions with the atoms in the sample A) unscattered beam, B) elastically scattering with no energy deposition into the sample, C) inelastically scattering, the energy deposited into the sample can result in the release of secondary electrons or orbital jumps with X-ray emission from the specimen atoms, based on[62]

The amplitude contrast is a result of the inelastic scattering events. As the electrons hit the sample, they lose some of their energy, which can be emitted in different forms of radiation or deposited to the sample, resulting in radiation damage. With lower energy,

inelastically scattered electrons would be focused by the lens system on a different plane than the original beam. Additionally, some of the electrons scattered with larger angles can be moved further from the optical axis of the microscope and be lost from the lens system or stopped by the aperture. The inelastically scattered electrons contribute to the formation of amplitude contrast but, in general, are considered damaging events. They can be removed during the data collection process with energy filters[51].

Elastically scattered electrons do not lose any of their energy. As a result of the scattering events, their path is changed, which results in a shift of their phases compared to the original beam, and they contribute to the phase contrast. The phase contrast is a result of electrons acting as a plane wave. As the wave propagates with specific amplitudes and phases, atoms of the specimen and solvent in the sample act as scattering centres. The elastically scattered waves would have different phases compared to the original unscattered beam. As a consequence, the interference between the signals can be constructive or destructive, which can increase or decrease the contrast of the final image or even cancel out the information at certain frequencies. With the objective lens system located below the sample in the microscope column, the scattered waves can be focused on the detector[63]. All of the electrons scattered at specific angles converge at the same point on what is called the backfocal plane. Effectively, the back-focal plane contains the Fourier Transform of the sample image, informing how much of the given Fourier components is present in the sample. As the electrons travel further down the microscope column, they create a real-space image (on what is called the image plane) as a result of the interference between the scattered and unscattered beam[64], [65]. Figure 2.2. presents the pathways of scattered electron beams focused by the objective lens to create a magnified image and how they interact at the backfocal plane to produce a diffraction pattern.

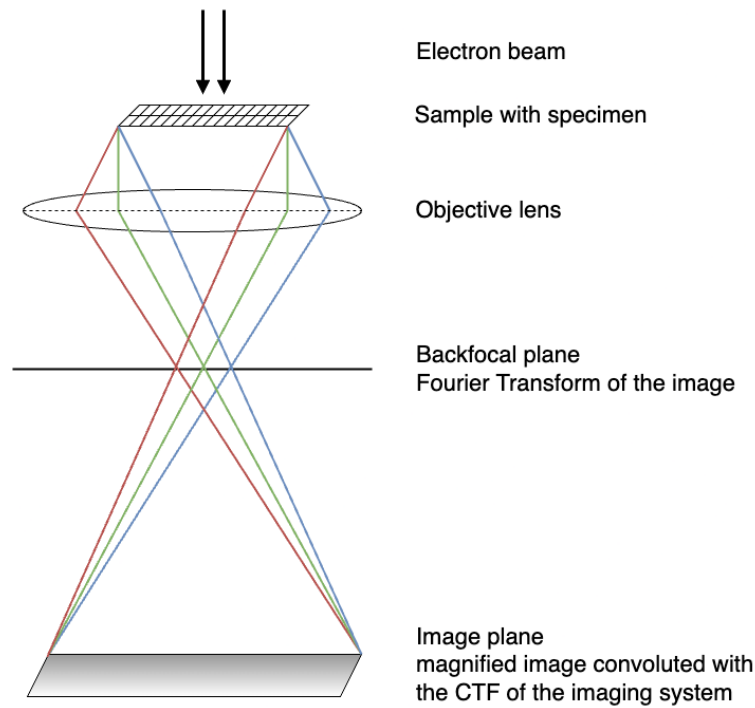


Figure 2.2 Electron beam scattered through the sample. Beams scattered with the same angle overlap at the same spots on the backfocal plane producing a diffraction pattern (Fourier transform of the image). As the electrons travel down the microscope column they recreate the real-space image on the image plane which can be magnified, based on [63]

The final image is then magnified and recorded by the detector. This is not the ideal representation of the information in the sample as it is affected by multiple imperfections of the imaging system, such as lens aberrations and defocus. The combination of all of these factors can be described by what is called a Contrast Transfer Function (CTF). It describes the changes in contrast as a function of spatial frequency. The changes in contrast result from how much of specific waves scattered at given angles contribute to the final image collected in the detector. The higher scattering angles correspond to the high-frequency information. The CTF values oscillate between -1 and 1 in a cosine-like manner, starting from 0. The places where the CTF have a value of 0 are called zero-crossings, and that frequency information is lost from the image[66], [67]. The defocus parameter is used to modulate the CTF during the data collection to get the most complete frequency coverage. The defocus value is a measure of how far from the sample the electron beam is focused. Typically the values range between 0.5-3.5 microns. Lower defocus values result in the first zero crossing at higher spatial frequency, which means that some low-resolution features, such as the general shape and positions of the proteins or virus shells, would be lost due to lack of contrast. Higher defocus results with first zero crossing at a lower frequency, providing stronger contrast of low-frequency features but introducing faster signal oscillations in high-frequency regions. As a result, the contrast

value and sign at neighbouring high frequencies start to cancel itself instead of contributing to the final image. This leads to signal delocalisation and loss of high-frequency details of the specimen. In summary, the defocus modulation of CTF leads to balancing the contrast and resolution in the recorded dataset, shifting the first zero-crossing position towards lower or higher frequencies. High defocus emphasises the low-frequency features at the expense of damping high-frequency information. Low defocus values make it easier to record high-frequency information but lose the low-frequency features[68]. The effect of different defocus values on the recorded images is presented in Figure 2.3.

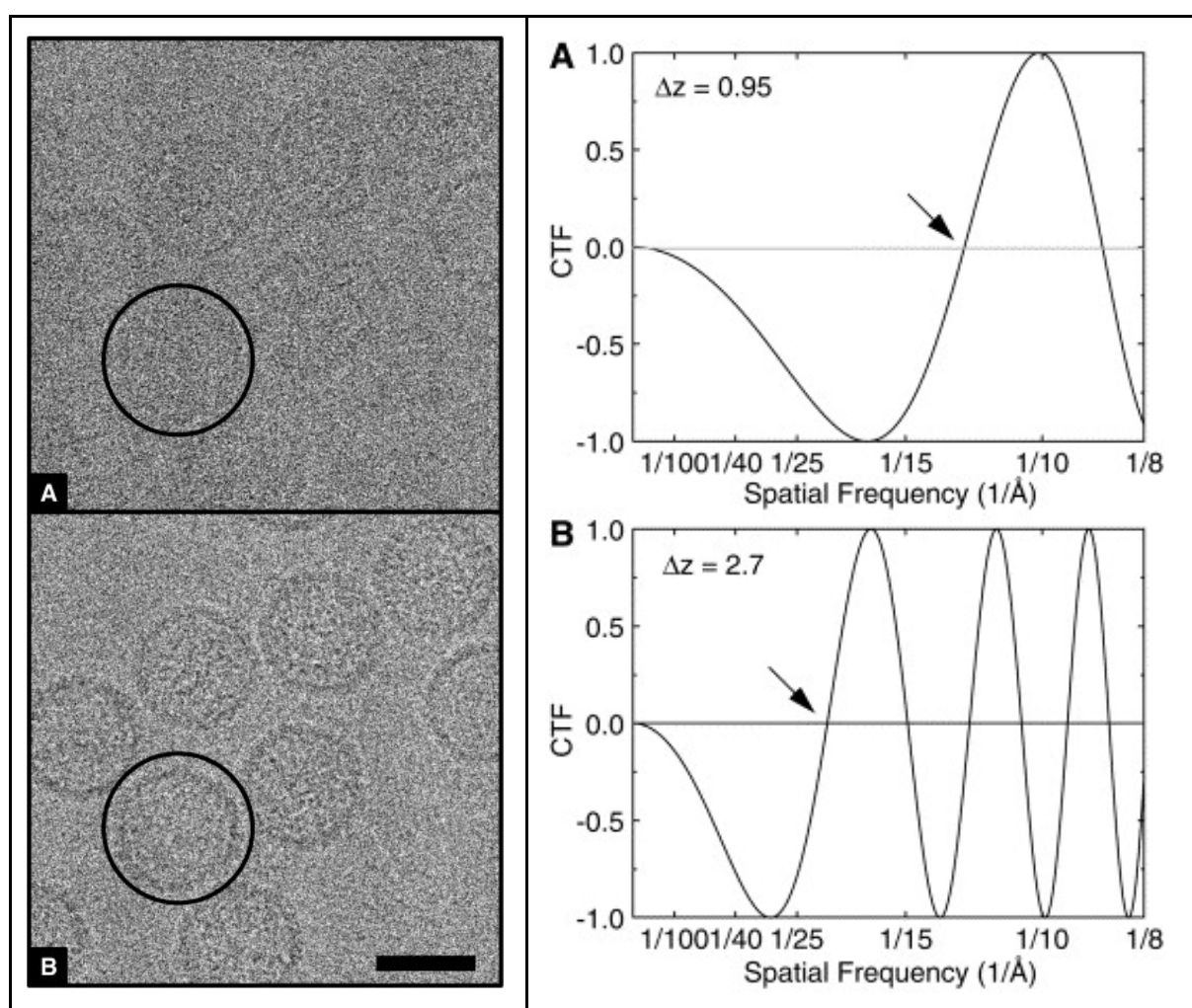


Figure 2.3 Effect of different defocus values on the collected images, and the CTF plots A) image recorded with low defocus (0.95nm) results in information loss in the low-resolution area with the first zero-crossing of the CTF is moved to higher spatial frequencies B) image with high defocus (2.7nm) reveals more low-resolution features of the specimen, as the first zero-crossing of the CTF is shifted to lower spatial frequency. Reproduced from [68] Copyright © 2000 Elsevier Science Ltd. All rights reserved. Permission to reuse obtained via RightsLink order 5856460032962

Additionally, the CTF is also affected by the envelope function, as the signal is dampened in the high-frequency ranges due to the electron beam's lack of spatial and temporal coherence. The high-frequency information is also lost to the blur introduced by the beam-induced motion of the specimen and radiation damage[69, p. 8].

Spherical aberration, commonly denoted as C_s , is one of the inherent properties of the objective lens, which results in stronger focusing of electrons passing further from the optical axis. This can be partially reduced by using the spherical aberration corrector, an additional lens system that introduces the aberration in opposite directions or by the objective aperture, which blocks the electrons with high scattering angles or with computational methods[70]. Another type of aberration is chromatic aberration, the electrons with different energies are focused at different positions on the optical axis. This is caused by the focal length of a specific lens depending on the electron energy. Electrons with higher energy will be focused behind the imaging plane and lead to amplitude contrast changes and blurring. The effect of the chromatic aberration can be reduced by using the energy filters to remove electrons with different energy[71].

Coma is an aberration caused by misalignment of the objective lens and the optical axis. It introduces directional, asymmetric blurring in the images and can be removed by carefully shifting and tilting the electron beam in the microscope or later during the image processing stage[72].

One of the other important parameters for the cryo-EM data collection is astigmatism. It generates directional changes in defocus in the image. As a result, the Contrast Transfer Function visualised in the Fourier space is represented as ellipses instead of circles, as it would be with no astigmatism. The astigmatism is determined by three parameters: defocus in the first direction Δf , second direction Δf and the angle β between the long axis of the ellipse and the X axis[73] as shown in Figure 2.4 A) Astigmatism can be corrected during the microscope alignment and calibration before the data collection. A common technique to identify the issues with astigmatism is to assess the circularity of the Thon rings in the Fourier transform of an image. The Thon rings represent how the CTF of the microscope modulate the signal. The interchanging bright and dark rings represent frequencies where the contrast is either maximised or the contrast is lost. As some of the features in the Thon rings patterns also depend on other factors such as defocus, all of them should be symmetrical and circular. In the presence of astigmatism, the Thon rings become elliptical which limits the final resolution, affecting especially the high-frequency features[74]. Figure 2.4 B) presents the Thon rings with the estimated Contrast Transfer Function.

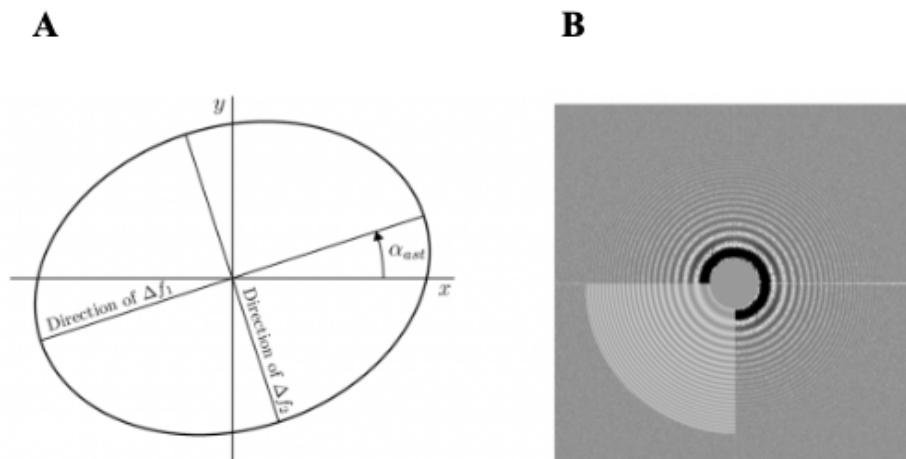


Figure 2.4 A) Parameters used for astigmatism correction, B) Thon rings from a cryo-EM micrograph overlaid with estimated Contrast Transfer Function (lower-left quartile), reproduced from [75] Copyright © 2015 Elsevier Inc. All rights reserved. Reuse permission obtained from RightsLink order 5856541467989

2.1.3 Electron detectors

Finally, the resulting image is recorded by the detector. Over the years different types of detectors were used, starting with photographic film, charged-coupled device cameras and direct electron detectors based on the complementary metal-oxide semiconductors (CMOS) technology. The commonly used parameters which can be used to evaluate the performance of the detectors are data collection speed and pixel size, but also Modulation Transfer Function (MTF) and Detective Quantum Efficiency (DQE). The Modulation Transfer Function describes the response of the detector to the input signal as a function in the frequency domain. This information can be used to check how contrast is preserved in different frequency ranges. It has a value of 1 at zero spatial frequency and decreases to 0 at the diffraction limit, but it is not monotonic. Due to this behaviour, high-resolution details of the specimen might not be able to resolve as the contrast is lost. Some factors that affect the MTF are the type of electron detector used (electron backscattering events occurring in detectors with thicker sensor layers), overall quality of the lenses and microscope alignment and electron beam coherence. Well-preserved MTF would make it possible to achieve the final resolution of $\frac{2}{3}$ Nyquist. With better MTF, the signal-to-noise ratio in the images is also higher, which makes it possible to achieve high-resolution reconstruction with a smaller number of particles or make it easier to identify heterogeneity in the sample. Some MTF information can be preserved by collecting the data with an energy filter, limiting the spread of electron energies. To recover information

lost due to the MTF decline, the signal can be weighted in specific frequency bands, or some sensor-specific MTF compensation can be applied based on the references provided for commonly used sensors at a given accelerating voltage[76].

The DQE is defined as a ratio of the detector's squared input signal-to-noise ratio and squared output signal-to-noise ratio. As a perfect detector would have a DQE of 1, this value is always lower as the signal from the electrons is affected by different noise associated with each pixel. The DQE is also represented as a spatial frequency function that allows the evaluation of detector performance in different resolution bands up to Nyquist. A detector with higher DQE can preserve the SNR better. For radiation-sensitive biological samples, it is more important to have the DQE as high as possible, even at the cost of lower MTF, as the MTF can be corrected during the processing pipeline with an additional filtration step. To maintain high DQE and minimise radiation damage, the images are collected with low dose and high framerate data collection speed[77].

Photographic film was commonly used in the early years of cryo-EM before digital detectors were introduced. The canister with a set of films is located at the bottom of the microscope column. After all films are exposed to record the data, the canister is taken out to develop the images from the films. Then, the developed images have to be digitised with a scanner for further processing. As the photographic film requires manual handling for processing, it brings the risk of damaging it or introducing artefacts to the images, but also introducing contaminations into the microscope as the film container is replaced. Additionally, there is no live feedback, and the data can be accessed and analysed only after the data collection is finished, which makes it impossible to easily optimize data collection parameters. The typical photographic film used in electron microscopy is a square piece of 8x10 cm, but thanks to the very localised interaction between the electrons and the film material, it can produce a high-resolution image up to 10,000 by 12,000 pixels after digitisation. The scanner step has to be set according to the microscope magnification to result in half of the Nyquist frequency. Photographic film has better DQE at high frequencies compared to other types of detectors but performs worse in the low-frequency range, which requires higher defocus during data collection[78].

One of the major advantages of digital detectors is that the images can be analysed almost immediately after the data is collected, which is a cornerstone for automated data collection and analysis pipelines. The CCD detectors have a scintillator layer made of phosphor, which produces photons in the places where electrons are scattered. Photons are then transferred via fibre optics to the CCD camera layer, where the image is recorded. The scattered electrons can highlight larger areas or even get backscattered and produce

photons multiple times which would lead to blurry images and artifacts. The CCD-based detectors have better DQE in the low frequency compared to the film but perform worse in the high-frequency bands. To mitigate this, a higher magnification is used to analyse the high-resolution details of the specimen in low-frequency regions. This approach, however, limits the number of particles in the field of view[79].

The development of direct electron detectors (DED or DDD) significantly contributed to the ‘resolution revolution’. The electrons are recorded directly on the silicon surface of the detector instead of being converted to photons. This brought improvement compared to the CCD cameras in terms of signal-to-noise ratio and signal localization. Additionally, with the $35\mu\text{m}$ thickness of the silicon layer, the scattering events are further reduced. The DQE of the direct electron detectors is higher in all frequency ranges compared to CCD devices. The CMOS technology used in this type of detector offers higher readout speed, which also changed the principles of data collection. With the high-speed sensors, the data can be recorded in the ‘movie mode’, which means that a series of micrographs is collected during a single exposure. Thanks to this, the beam-induced motion of the specimen can be corrected in the data processing pipelines. Direct detectors are also more sensitive than CCD systems, so the images can be recorded with lower electron dose resulting in more data before the sample is damaged by the radiation. Figure 2.5 shows the comparison Modulation Transfer Function and Detective Quantum Efficiency of three direct electron detectors: Gatan K2 summit, FEI Falcon II, and Direct Electron DE-20, used for 300 keV cryo-EM data collection[80].

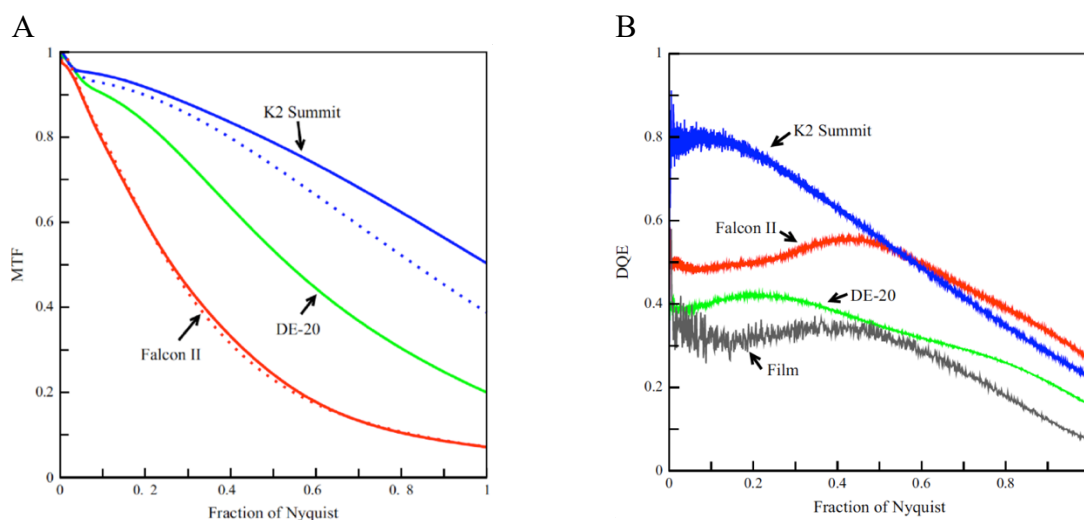


Figure 2.5 Characteristics of the detectors commonly used in cryo-EM A) Modulation Transfer Functions, B) Detective Quantum Efficiency. Reproduced from [80] Copyright © 2014 The Authors. Published by Elsevier B.V. This work is licensed under a Creative Commons CC-BY license.

Modern direct electron detectors can operate in different modes during the data collection depending on the requirements to prioritise either the data collection speed or the quality of the final reconstruction.

In the integrating mode, the total charge generated as the electrons hit the detector over the exposure time is recorded. This means that as the electron interacts with the sensor layer, the signal can be spread across the group of pixels (Fig. 2.6.A). The exposure time is shorter, and a higher dose can be used. As a result, the overall data collection speed is faster, but the noise from the electron interactions with the detector lowers the DQE. The counting mode offers better signal localisation, as the individual electron events are detected and counted over the exposure time (Fig. 2.6.B). This is achieved thanks to the high frame rate. Recorded events are reduced to the pixels with the highest charge instead of being spread across a group of pixels. Lower input noise and limited electron scattering events on the detector surface result in higher DQE in all frequency ranges. With longer exposure time lower dose has to be used to avoid rapid radiation damage. The data collection speed is also lower compared to the integrating mode, but improvements in DQE and superior image quality usually make it possible to achieve a higher final resolution of the reconstruction[81]. The super-resolution mode is the approach which makes it possible to record data beyond the physical limit of the pixel size. As the distribution of energy deposited during the electron event is analysed over a small cluster

of pixels, it can be estimated how electrons interacted with the detector in sub-pixel areas, effectively doubling the original pixel resolution (Fig. 2.6.C)[82].

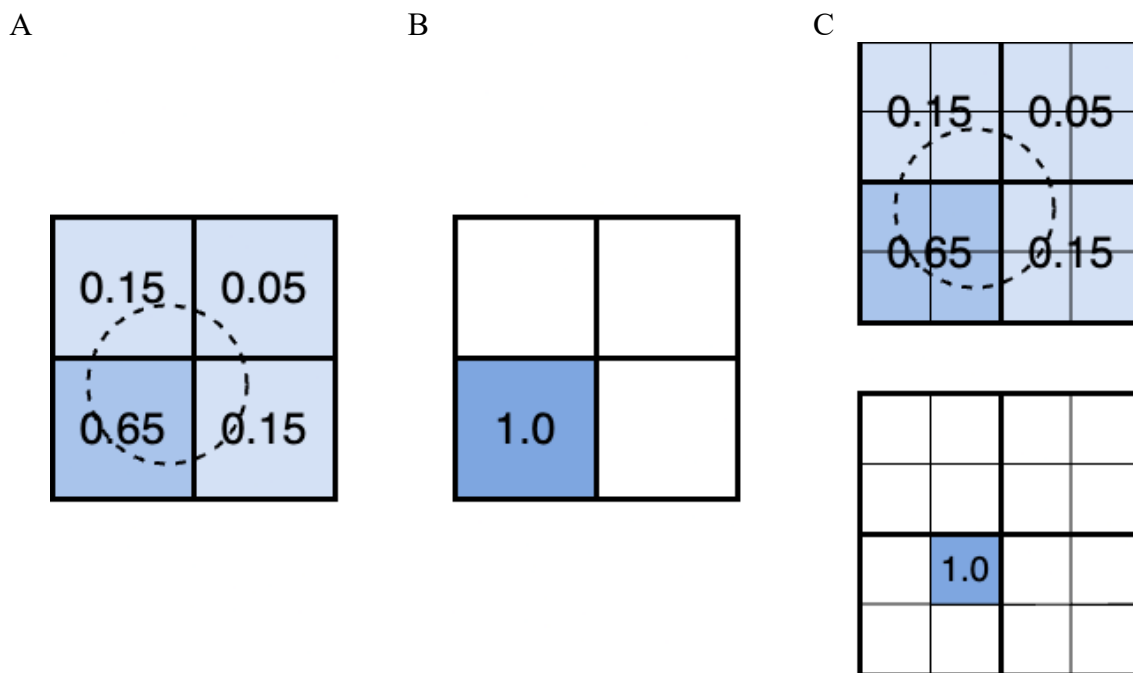


Figure 2.6 Different operating modes of electron sensors in cryo-EM A) integrating mode: recorded electrons are spread across the group of pixels, faster data collection but also a higher noise from electron interactions B) counting mode: higher frame rate leads to better signal localisation, events reduced to a single pixel with highest value C) super-resolution mode: analysis of the electron distributions in a small cluster of pixels allow estimation of the signal location at sub-pixel level, based on [82].

The advancements in the detector technology allowed for routinely obtaining a near-atomic resolution of the 3D electron density maps calculated from a series of 2D images. Recent advancements in both hardware and software tools opened the doors to the true atomic resolution with a 1.2 Å map of apoferritin [22], which made it possible to resolve individual atoms in a protein. Despite the constant improvements and efforts from engineers and scientists, the cryo-EM data collection and processing procedures still need techniques to improve the reproducibility and quality of the reconstructions. Figure 2.7 shows the main stages of the Single Particle Analysis workflow with the cryo-EM, from sample preparation to final atomic model validation.

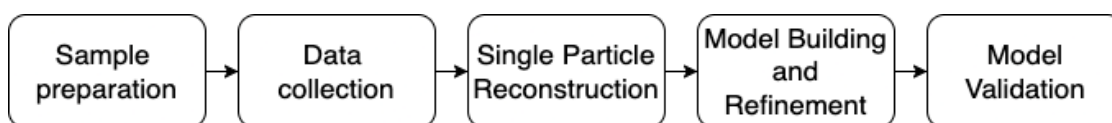


Figure 2.7 An overview general workflow of Single Particle Analysis with cryo-EM.

2.1.4 Fourier space operations in cryo-EM data processing

Since the very beginning of electron imaging the Fourier transform and operations in the Fourier space played a crucial role for the signal processing. In 1960s it was commonly used for amplification of the signal with Fourier filtering, determination of the particle's orientation by the projection matching and even 3D reconstruction[83]. This section briefly introduces basic ideas and operations used at various stages of the cryo-EM processing pipeline, with an outline of how they can be applied to specific tasks.

The Fourier transform decomposes the image into its sine and cosine components. This way the image can be represented in the frequency domain, where each point corresponds to a specific frequency. For the image processing after the Fourier transform is done, the image is represented as a series of 2D signals. The 2D Fourier transform is defined with Equation 2.1, where u and v are spatial frequencies,

$$F(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-j2\pi (ux+vy)} dx dy, \quad \text{Eq.2.1}$$

and the inverse transform with Equation 2.2:

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(u, v) e^{j2\pi (ux+vy)} du dv, \quad \text{Eq.2.2}$$

The Fourier transform can be fully inverted without losing information. An example of an cryo-EM micrograph and the amplitude spectrum from its Fourier transform is shown in Figure 2.8. The low frequencies are represented in the centre of the image and increase towards the edges.

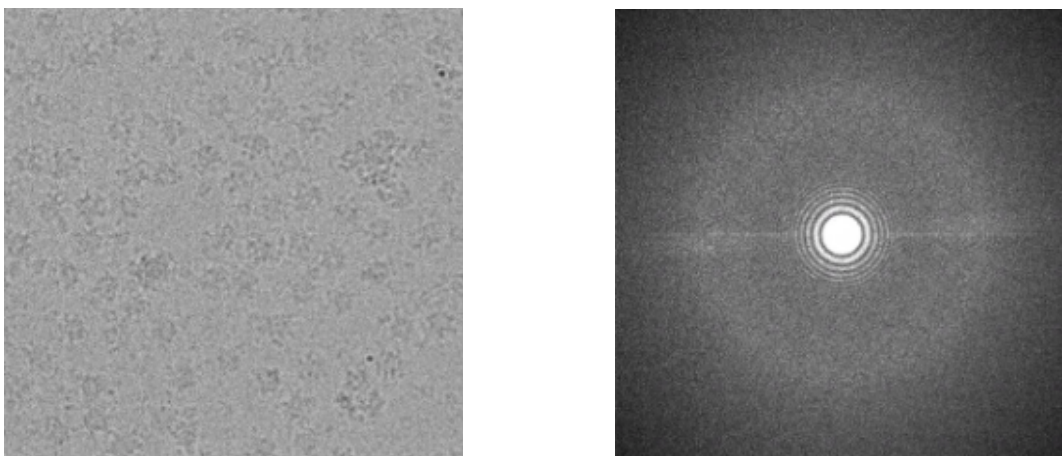


Figure 2.8 Cryo-EM micrograph and it's power spectrum

The convolution theorem states that the convolution of two signals in the real space is equal to the inverse Fourier transform of the multiplication of their Fourier transforms (Eq. 2.3)

$$f(x) \otimes g(x) = F^{-1} (F(x)G(x)), \quad \text{Eq.2.3}$$

This relationship can be useful for filtering of the images, as instead of convoluting the image with the filtering kernel, a mask in the Fourier space can be applied to mask out the low or high frequencies and after the inverse transform obtain the filtered image in real space. It is important to avoid sharp edges on the filtering mask as it can introduce Fourier ripples to the reconstructed image.

The cross-correlation can be used to identify similarities between two functions as one of them is shifted over the other. In the Fourier space, it is represented as a multiplication of one of the functions by the complex conjugate of the other (Eq. 2.4).

$$\int f(x)g(x - t)dx = F^{-1} (F(x)\overline{G(x)}) \quad \text{Eq.2.4.}$$

This property can be applied for the template matching for particle picking. As one of the functions can represent the template moving along the micrograph. The high correlation values would be reported for the locations representing the template in the image. It can also be applied for projection matching during the 2D classification or for the alignment of images.

Finally, the 3D reconstruction is based on the central slice theorem, as the Fourier transform of a 2D projection of a 3D volume is also a central slice through the 3D Fourier transform of that object. The direction of the projection can be characterised by the vector normal to the slice.

2.1.5 Sample preparation

The first step of the cryo-EM experiment is sample preparation. The quality of the sample is crucial for successful data collection and high-quality reconstruction. One of the parameters to optimise at this stage is protein concentration, typically in the range of 0.5-2mg/mL. Higher concentration can be beneficial to the formation of thin ice layers as the

particles are closely packed, but it can also lead to protein aggregation and reduce the number of particles useful for imaging. The stability of proteins is maintained by the buffer composition. Factors like ionic strength, pH, or addition of the detergents need to be optimised and typically, several different buffer conditions are tested and screened for optimal setup to identify the ideal setup[84]. Other factors that affect the sample quality are ice phase, thickness and uniformity. Crystalline ice can diffract the electrons, creating artifacts and leading to image degradation. The rapid freezing of the samples helps to achieve vitreous ice, which is a low-density amorphous ice state that preserves the near-native state of proteins and improves contrast and image quality. After freezing, the sample must be handled at low temperatures (around -200°C) to avoid ice phase changes that can build up crystals and contaminations[51]. The preferred thickness of the ice ranges between 10-100nm but also depends on the size of the particles. Too thin ice might crack or exclude some specific views of the particles. Too thick ice can introduce inelastic scattering events. Recent studies indicate that the final resolution of the reconstruction depends on the ice thickness in the sample[85]. Ice uniformity can also help in data analysis and interpretation. The non-uniform ice distribution can lead to the preferred orientation problem, where some specific views of the particles are excluded. Also, varying ice thickness can result in different focus values for individual particles, which would be a challenge in the processing pipeline. Finally, the quality of the sample is also defined by the protein distribution on the grids. A sample, which has proteins well-separated and not overlapping, evenly spread and providing many orientations is essential to obtain high-resolution reconstruction. The optimisation of the sample preparation procedures often requires a trial-and-error approach to find a proper buffer composition to ensure protein stability and achieve a sample with proper protein distribution in a thin, vitreous ice layer. This part of the thesis will describe approaches and parameters used in sample preparation routines to achieve high-resolution reconstruction.

The protein and solvent solution is applied to the support grid. The support grid parameters to consider are material, geometry arrangement of the grid squares and material and geometry of the support film. The typical support for the cryo-EM sample is a 3mm metal, usually copper or gold, mesh grid. The size of the mesh is described by the number of squares per inch with typical available values of 200, 300, or 400. The density of the grid squares affects the data collection and can be selected based on the specificity of the planned experiment. For example, larger grid holes provide a better field of view for data collection, which can also be useful for tomography as the grid bars will not obstruct the view as the grid is tilted. On the other hand, smaller grid squares may provide

better mechanical stability for the sample and allow using higher beam intensity before the sample is devitrified, as the tighter packing of grid bars helps in better heat distribution. The mesh surface is covered with a perforated foil. The foil can be a perforated metal foil that creates a regular array of holes or holey and lacey carbon, which offers larger open areas for imaging with fine, irregular mesh. The foil hole area is where the sample will be placed for imaging. (Figure 2.9). Common materials used for the support film are amorphous carbon, gold or graphene. The typical foil parameters are hole size and distance between the holes, which range between 1.2/1.3 μm for diagnostic and 0.6/1.0 μm for high-resolution cryo-EM. The notation 1.2/1.3 means that the foil hole size is 1.2 μm , and the spacing between the holes is 1.3 μm . As the smaller field of view may limit the imaging area, the smaller size of the holes helps to limit the beam-induced motion of the specimen. The selected size of the foil holes also depends on other factors, such as the size of the particle. [86].

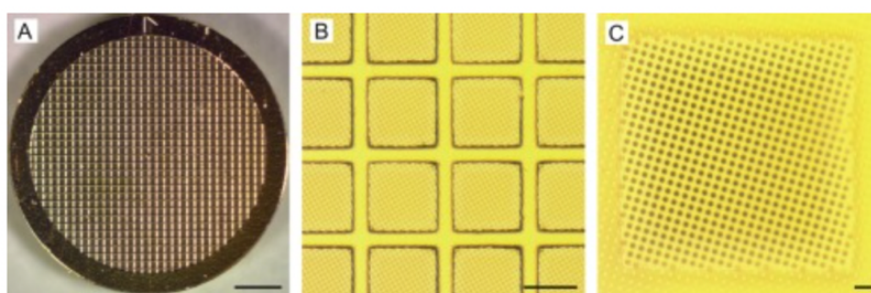


Figure 2.9 cryo-EM sample support a) 3mm metal mesh grid, b) view of the grid squares, c) close-up of the perforated gold foil on the surface of the grid showing the pattern of the holes. Reproduced from [86] Copyright © 2016 Elsevier Inc. All rights reserved. Reuse permission obtained via RightsLink order 5856210776027

The properties of the material from which the support is made also affect the quality of collected data. Two main features to consider when choosing the material for a cryo-EM grid are thermal shrinkage and charging. As the grids are exposed to the extremely low temperature of liquid ethane (77K), the material shrinks, which can introduce tensions on the surface. For example, vitreous ice has a higher thermal expansion coefficient compared to copper or amorphous carbon. During rapid freezing, the ice shrinks more than the supporting materials, leading to tensile stress in the sample. In some cases, it can lead to changing the geometry of the sample or even cracks in the supporting film. Insulating materials like amorphous carbon can be charged with electrons, and build up a positive charge in the vitreous ice, which would deflect the electron beam and result in blurry images[87].

The properties of golden supports, such as high conductivity and radiation resistance, allow reducing the specimen motion during freezing compared to carbon supports. Additionally, the lower electrical stability of amorphous carbon caused by the manufacturing process can increase the instability of the sample and result in additional blurring resulting not only from specimen motion but the displacement of the support itself[88]. The optimization of the support grid and exploration of new materials to improve the sample distribution and parameters is still a point of interest for many researchers.

One of the new materials used for the support film is monolayer graphene, which is a conductive material, only one atom thick and can be almost invisible on micrographs up to 300 keV. The reduced thickness of the graphene film (0.34nm compared to the typical 20-200nm sample thickness) reduces the background noise from secondary electron scattering. It also shows a distinctive hexagonal diffraction pattern. A graphene monolayer can be applied to the grid holes made of stable gold, which allows for keeping all of the benefits from golden support[89]. The data obtained this way usually have better quality as the particles supported in the graphene monolayer have less beam-induced movement. The conductivity of this material can help to reduce the radiation damage, dissipating the build-up charge. It also could help to reduce the radiation damage thanks to potentially achievable thinner sample. A potential downside of this approach is that the graphene layer is hydrophobic and requires plasma cleaning or other chemical treatment to reduce this. Also, the preparation of grids with graphene monolayer support could result in low coverage and limited areas for effective cryo-EM data collection. The high quality and coverage grid preparation process itself requires expensive instruments, which are not yet easily accessible by most laboratories[90].

Another direction of cryo-EM support improvements is experimenting with different geometry of the grid holes. Replacing the standard square grid holes with hexagonal holes (HexAuFoil) helps to maximize the number of the holes on the grid for more effective imaging compared to circular holes. It also allows the introduction of fiducial markers to easily identify the orientation of the grid and regions of interest for imaging at different magnification levels.[91].

Before the sample is applied, the grid should be cleaned to remove any substrates that can be present on the grid after the manufacturing process or the storage. The grids can be cleaned chemically by dipping them in chloroform, acetone and ethanol, followed by rinsing with water. This mechanical process can be replaced by more convenient glow-discharge of the selected grids in a plasma chamber to make sure that they are clean, not

too hydrophobic, and the sample will coat it [92][34]. There are many ways to apply the solvent with protein on the grid, Figure 2.10. shows a range of devices used for sample deposition A) capillary effect used in cryoChips, B) sample spraying set-up with Shake-it-off.

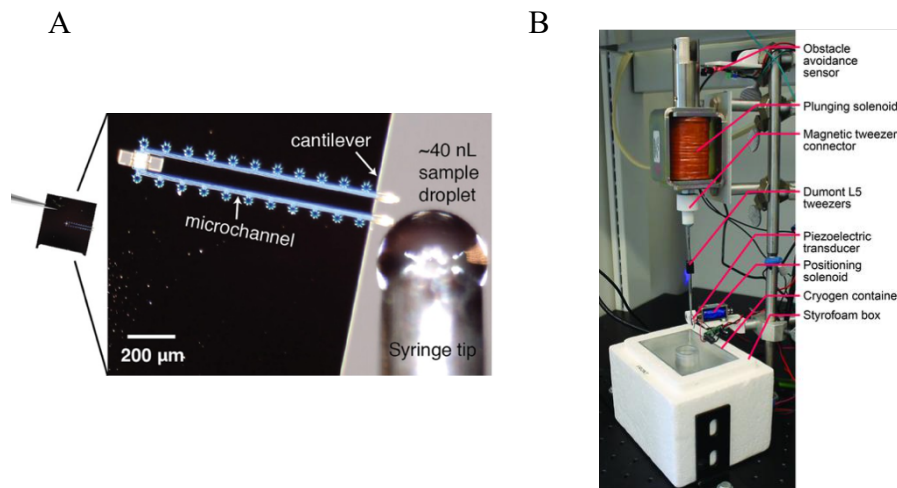


Figure 2.10 A) cryoChip sample deposition device, reproduced from [93] Copyright © 2022, Huber et al, This article is distributed under the terms of the Creative Commons Attribution License B) Shake-it-off set-up, reproduced from [94] Copyright © Rubinstein et al. 2019, This is an open-access article distributed under the terms of the Creative Commons Attribution (CC-BY) Licence

Plunge freezing can be done using a device such as FEI Vitrobot or Leica GP. A single droplet of the solution with the specimen is applied on the grid in a temperature- and humidity-controlled chamber. The excessive liquid is then removed with the blotting paper with controlled force and time. Then, the grid with the sample is plunged into the cryogen, such as liquid ethane and rapidly frozen. Liquid ethane is used at this stage as the liquid nitrogen boils at 77K (-196°C)[95], [96]. This boiling effect can have a negative impact by creating a vapor layer around the sample, slowing down the cooling rate, which can lead to the formation of crystalline ice[97].

Other approaches do not require mechanical blotting. The major advantage of the blotting-free methods is that the prepared sample volume is not wasted on the blotting paper.

Shake-it-off uses a piezo-electric sprayer to apply the solution to the grid. The piezo-electric sprayer frequency can be adjusted to control the size of the droplets in the range of 100-10 nm. However, in some cases, the specimen might be damaged by high-frequency vibrations of the sprayer. The spring-driven plunging is done when the tweezers with the grid are rapidly released from the plunging solenoid[94]. Spotiton and Chameleon allow an inkjet to dispense pico- to nano-litre volume of solution droplets on

the grid before it is plunged[98][99]. The sample is deposited on self-wicking nanowire grids. The nanowires' capillary action helps to spread the applied liquid before the sample is plunge-frozen to achieve more uniform ice distribution across the grid holes[100][101]. The latest developments in the sample preparation procedures include the nanofluidic sample support system called cryoChips[93]. It uses the capillary effect to apply the solution to the grid via the nanochannels. The fact that the same nanochannels are used to produce different samples could improve the reproducibility of the process. Grids prepared this way can be plunge frozen manually or with any robotic station. Another procedure of sample deployment to the grid is cryoWriter, also based on the microcapillary effect. Instead of blotting, to remove the excess of the solution, the sample is thinned with the laser diode, which allows the liquid to evaporate[102]. Ideally, after the plunge freezing, the specimen would be suspended in a thin layer of vitreous ice and evenly distributed in the foil holes. Figure 2.11 shows examples of different grid types used for sample deposition: A) self-wicking nanowires and B) hexagonal grid holes. Table 2.3. shows a comparison of the different sample preparation methods.

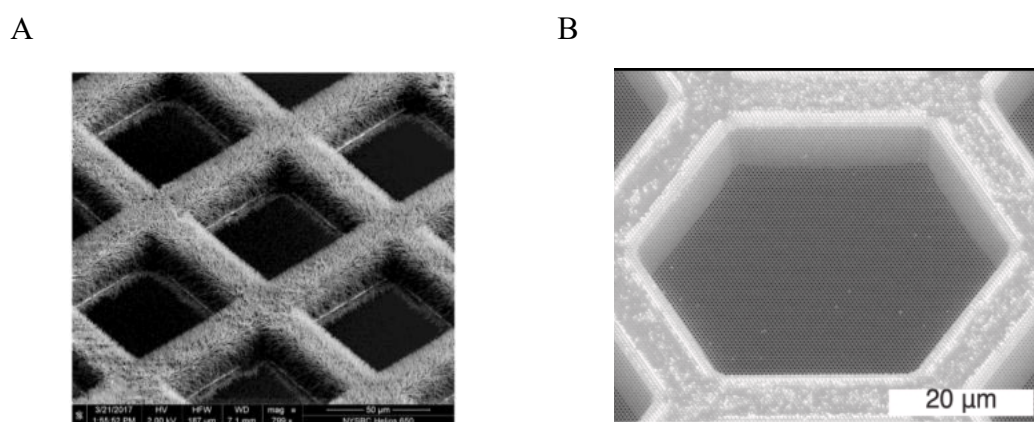


Figure 2.11 A) Self-wicking nanowire grid used with Spotiton/Chameleon foil, reproduced from [100] Copyright © 2018 Elsevier Inc. All rights reserved. reuse permission obtained from RightsLink order 5856520321372, B) example of HexAuFoil reproduced from[91] Copyright © 2021 MRC Laboratory of Molecular Biology This is an open access article under the CC BY license

Table 2.3 Different techniques used for the cryo-EM sample preparation.

Sample deposition technique	Devices	Method	Sample volume per grid
Blotting	Manual plunger, Vitrobot	Pipetting, liquid wicked through paper filter	3-5 μ l
Ultrasonic spray	Back-it-up	High-frequency droplet generation with through-grid wicking	200nl-1 μ l
	Shake-it-off	High-frequency droplet generation with self-wicking grids	50nl
Inkjet	Spotiton/Chameleon	Droplets generated with piezo-electric device, deposited on self-wicking grids	2-16nl
Capillary effect	cryoWriter	Sample deposited using capillary effect with dewpoint control	0.1nl
	cryoChips	Closed nano-channel architecture, sample thickness controlled with the channels geometry	4pl

Before collecting data using a high-end microscope, the samples should be screened to check their overall quality. Some of the issues, such as damaged or bent grids, crystalline ice contaminations, empty or cracked foil holes and problems with overall particle distribution on the grid, can be identified before data collection, and the sample should be optimised by changing the parameters like specimen concentration in the solution or plunge freezing setup and parameters. Regardless of the methods used, the major issue for most of the sample preparation procedures is the randomness of protein particle distribution, ice quality and the general lack of reproducibility[103][104].

2.1.5.1 Specimen behaviour and ice thickness evaluation in the cryo-EM samples

The ideal cryo-EM sample would have a single layer of particles supported by a thin layer of flat vitreous ice. The thinnest possible sample (that still supports the specimen without air-water interface damage) can help to reduce the chances for secondary scattering events thanks to the shorter path the beam travels through the sample and radiation damage. On the other hand, ice that is too thin might not be able to support the specimen. As a result, particles can be pushed to the edge of the foil hole and aggregated, disassembled, or denatured on the air-water interface. Too thick ice can result in more than a single layer of particles in the z-axis direction, which can lead to particle overlap and make it problematic to calculate defocus per particle. A wide range of angular orientations in the sample would help to ensure the representation of the most unique views and improve the final reconstruction. Also, no crystalline ice or other contaminants should be in the field of view as they might be detected as particles (especially by automated pickers) and affect the resulting cryo-EM map if not detected and removed from the dataset. These conditions should ensure effective data processing and allow users to obtain a high resolution of the final map. Unfortunately, due to the inevitable variability in sample preparation, all of the mentioned sample parameters may vary. Instead of the flat, thin layer, vitreous ice can create a convex or concave meniscus that will also affect the particle distribution (Figure 2.12). The non-uniform ice distribution can lead to major problems at the data processing stage as it can affect the performance of particle picking or limit the possible final resolution as the particles in different ice thickness areas have different signal-to-noise ratios. In general, areas with the thinnest ice, up to 50nm, have less inelastic scattering of the electrons, which allows for clearer imaging. Thicker ice (over 150nm) contributes to more inelastic scattering events, which can introduce background noise.

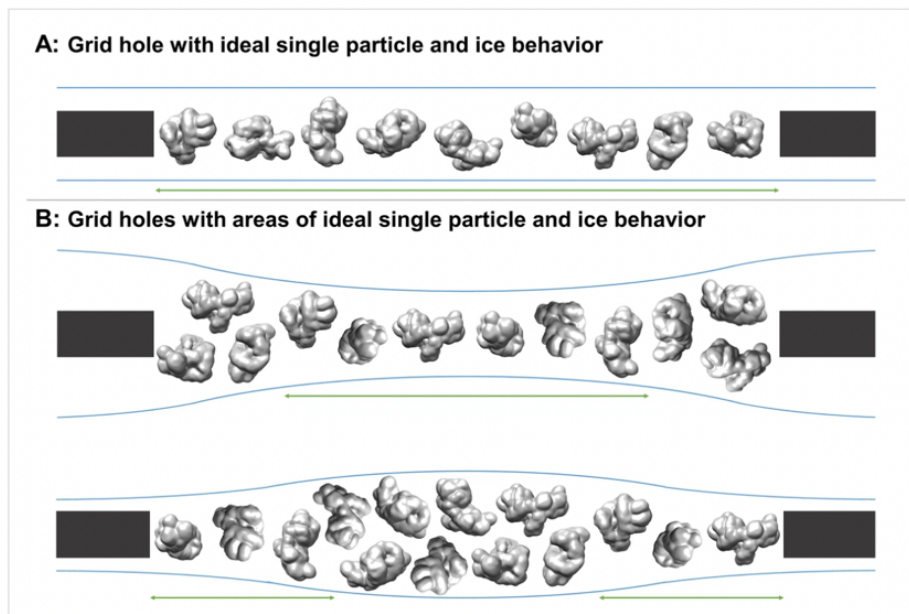


Figure 2.12 Overview of the specimen behaviour in different ice thickness conditions, A) ideal ice, B) with non-uniform ice distribution. Reproduced from [105] The work is made available under the *Creative Commons CC0 public domain dedication*.

The ice thickness is an important factor that can determine the quality and resolution of the final cryo-EM map. Most of the automated data collection frameworks offer functions to easily identify bent and deformed grids. For example, software like EPU can generate a virtual regular array of circles based on foil-hole size and spacing, which can be overlaid with the view from the microscope. If the pattern does not align with the actual image, it can mean that the grid is mechanically damaged, bent or twisted, which can happen if the sample was handled with metal tweezers without proper care[106]. Also empty or cracked holes which are not suitable for high-quality data collection can be identified based on image analysis algorithms. These solutions have been available for well over ten years now and implemented in Legion[107], UCSFImage[108] and in the EPU from ThermoFisher Scientific. More advanced methods to identify specific issues with the sample quality, ice thickness and distribution were developed over the years.

Energy filter can be used to estimate ice thickness based on the analysis of the inelastic scattering events. The images taken with and without the filter are compared to derive the ice thickness based on the mean free path of electrons in ice. This technique can be incorporated into the processing pipeline but requires a microscope with an energy filter. Aperture Limited Scattering (ALS) measures the intensity of electrons scattered outside the objective aperture positioned behind the specimen[109]. The thicker ice results in greater scattering and thinner in lower. The intensities of the remaining electrons are used to estimate the ice thickness. The ALS can be most useful at lower magnification levels

for evaluating large areas of the grids with minimal radiation damage to the sample. This way, the best areas for high-resolution data collection can be identified. With optical interferometry imaging, the ice thickness in a range from 0 to 70 nm can be measured. The optical measurements are based on thin film interferometry principles, which can predict ice thickness up to the first constructive interference point. This technique can be integrated into the sample preparation workflow, for example, with the optical camera incorporated into VitroJet system[110]. A more recent approach, MeasureIce, simulates thickness-image intensity look-up tables based on scattering physics. This tool allows for on-the-fly measurement of ice thickness without extensive calibration, making it accessible and convenient for researchers. It has been shown to correlate well with other measurement techniques and can significantly aid in selecting optimal acquisition areas on cryo-EM grids[111]. Other approaches for measuring the ice thickness include the ice channel method, which involves burning the hole through the ice with a condensed electron beam at a specific stage tilt, then tilting the stage to a defined second orientation, and the length of the channel is measured to check the ice thickness[35]. The tomography is the most time-consuming and includes collecting a dataset at different stage tilts, which leads to the reconstruction of the 3D map of the ice layer[105].

2.1.6 Data collection

Cryo-EM data collection and analysis can be done at different magnification level. The first step is to evaluate the overall grid quality and to identify some potential issues such as empty or damaged areas of the grid, large ice contaminations and to check if the grid is not bent before proceeding to select the areas for high-quality data collection. This can be done by acquiring the ‘atlas’ of the grid, a low magnification montage representing the whole grid. It is done in low-dose mode as a series of images representing different areas of the grid. The images are overlapping which allows stitching them together into one map that allows the users to assess the quality of the sample and decide if a specific grid can be used for the high-quality data collection. If, at this stage, the user observes any issues with the quality of the grid, they should consider optimising the sample preparation procedure. The ‘atlas’ of the grid can be collected in an automated way with most of the currently used microscopes. Once grid squares for the data collection are selected, the electron beam is centred over a single grid hole, still at low magnification. One of the major concerns during the cryo-EM data collection is radiation damage caused mostly by the energy released during inelastic scattering events as the beam of electrons goes through the sample. The sample can be damaged in multiple ways. Firstly, the covalent

bonds can be broken. Secondly, the excited and freed electrons freed this way move on to break more bonds in a cascade reaction. Additionally, the radiation can lead to hydrogen gas build-up within the sample, which can distort or dislocate the protein particles and create local bubbles, especially at the protein-ice interface, hence the name ‘bubbling’ effect[112][113][114]. To minimise the effect of the radiation damage, the sample is cooled below 100 K as the studies show that it offers cryo-protection, which limits the motion caused by the ionising beam. Unfortunately, temperatures below 50 K may change the structure of the vitreous ice. Radiation damage can also be reduced by proper optimisation of the exposure and accelerating voltages used for data collection[115], [116].

After setting the data collection parameters, a series of several images are recorded, which will be later summed into a single image[117]. Figure 2.13 A) shows the view of a grid, B) a grid-square image, C) foil-holes D) a cryo-EM micrograph.

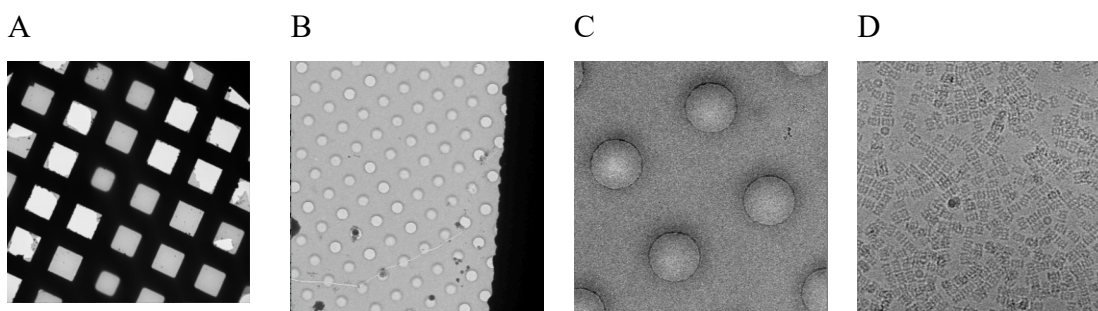


Figure 2.13 Different magnification levels for cryo-EM data collection A) view of the grid, B) grid-square view, C) foil-holes, D) an example of a micrograph, based on EMPIAR-10143 dataset

The other data acquisition strategies allow tuning the defocus for each recorded frame with the objective lens aperture voltage modulation[118]. With the use of a Volta phase plate, the data can be collected in focus, which theoretically can help to improve the signal-to-noise ratio. The Volta Phase Plate introduces a phase shift between 45° - 135° , which was shown to improve the contrast of the image without the need to apply defocus[119]. The phase shift value is hard to control as it changes over time as the carbon film of the VPP accumulates the charge. Shifts of 90° and higher lead to contrast inversion and increased image blurring of the image as more low-resolution information is included. Additionally, the introduced phase shifts can vary across different areas of the phase plate, which might affect the accuracy of the CTF estimation. As the usage of a Volta phase plate can create problems with focusing the microscope, the Laser Phase Plate uses high-intensity laser beams with continuous (standing) waves to manipulate the

phase of electrons without introducing any physical elements in the way of the electron beam, which can also get charged over time and affect the final phase shift[120].

2.1.7 Single Particle Reconstruction

The ultimate goal of the single particle reconstruction pipeline is to obtain a possibly high-resolution three-dimensional Coulomb potential map representing a 3D structure of a protein from the 2D images recorded at the microscope. To do so, several steps are required. First of all, the imperfections arising from the data collection procedures should be accounted for. The recorded images are corrected to account for the motion and radiation damage caused by the electron beam. Then, the Contrast Transfer Function is estimated for each image to compensate for the distortions introduced by the imaging system. The next step is to pick the particle coordinates that can be across multiple images. Due to the very low signal-to-noise ratio in the cryo-EM image, the projections of the particles are averaged from multiple similar views. This can be done manually, based on the template matching algorithms or with recently developed machine learning algorithms. The similar views of the particles are grouped together into 2D classes. The 2D classes can be used to generate an initial 3D model which plays the key part as a reference in future refinements. The 3D model of a protein is refined through iterative processes. This involves further classification of the particles based on their similarity to the current 3D model, followed by recalculating and updating the 3D reconstruction. This step may be repeated multiple times to improve the resolution and accuracy of the final model. If the collected data is high-quality and the processing steps are performed correctly, a high-resolution final 3D map is obtained, which can be used for structural interpretation and atomic model fitting.

2.1.7.1 Beam-induced motion correction

The biological samples used for cryo-EM are sensitive to radiation damage[121]. To reduce this damage, the final micrograph is averaged from several low-dose images recorded in the multi-second exposure time. The increased exposure time results in another factor which limits the quality of the images and the final resolution. The electron beam used for imaging causes the beam-induced motion, that introduces the 3D deformation of the whole sample but also introduces local motion for each particle[122][123]. As the whole-frame motion caused by the drift of the stage is estimated and corrected early in the data processing pipeline, information about individual particle trajectories is used for the high-resolution 3D refinement. The whole

frame movement is a combined result of the stage drift which for carbon support can be even up to a few nanometres and rotation with a few degrees tilt. As the ice layer moves, so do the particles. Moreover, as the data is collected as a series of single frames, the movement is more prominent in the first collected frames, while the last frames are more affected by the radiation damage as the electron dose builds up over exposure time. To compensate for these effects and reduce the blur, the frames are averaged. Commonly used MotionCorr2 software divides each frame into smaller patches (typically 5 in x and 5 in y direction). The motion within each patch is determined iteratively for each sub-frame of the recorded movie stack with the 2D polynomial functions to describe the position shift in time[124]. The motion for each patch is then corrected and frames are summed into the whole motion-corrected micrograph, optionally also including the radiation damage. This allows estimating the local shifts in specific regions, but also the full-frame motion trajectory. The new implementation of MotionCorr2 includes a smart selection of the centres of patches for local motion correction in places where motion features are detected[125].

One of the approaches to individual particle motion correction incorporates the calculation of the correlation between the Fourier transform of each individual frame and the transform of the summed frames. After optimisation of the objective function, the local correlation of single particle trajectories on subsequent frames can be also calculated. After smoothening these trajectories, the translation of the individual particles in the movie can be identified and used to compensate for the beam-induced motion. Additionally, this analysis allows users to identify Fourier components fading due to radiation damage and try to recover them as well[126].

The “Bayesian polishing” procedure implemented in the Relion cryo-EM data processing software uses a statistical approach to accurately trace the motion of the particles as the sample is exposed to the electron beam. Instead of calculating a running average of the movement of each particle over the frames in a collected movie, this method maximises the likelihood of particle trajectories from the likelihood from data and the prior likelihood imposed to favour smooth trajectories. The three parameters that are used to describe the statistics of the motion are the expected amount of motion (σ_D), spatial correlation of motion (σ_B) and average acceleration of the particle (σ_A). These parameters, after initialisation, are optimised iteratively by evaluating the alignment of the micrographs after each iteration. Additionally, it provides a model of radiation damage based on the dose and spatial frequency that can be used to apply proper B-factors

to mitigate that damage. The most effective use of this approach in the processing pipeline is to use the initial 3D refinement as a reference map to calculate the statistics of motion. Then, the optimised motion parameters can be applied to the original unaligned data to improve the final map. The proper motion correction for the particles can lead to routinely obtaining better results with a smaller number of particles and recovering high-resolution features of the specimen[127].

2.1.7.2 Contrast Transfer Function estimation

The Contrast Transfer Function describes how the microscope parameters and data collection setup parameters affect the recorded data. It is a sinusoid-like function dependent on the frequency and oscillates between negative and positive values. This causes some frequencies to get positive contrast and others negative. Moreover, the information at the zero-crossings of the sine wave is completely lost, and data must be collected at different defocus levels to obtain reliable 3D reconstruction. For the higher frequencies, the amplitude is attenuated. This results in the modulation of the contrast of the final image varying with the resolution.

The contrast transfer function for low-dose images can be described with Equations 2.5-2.7[128] where A is the amplitude contrast parameter, \vec{s} is the spatial frequency, $\gamma(\vec{s})$ is the function of spatial frequency representing varying phases, and $\Delta\phi$ is a global phase shift connected to amplitude contrast.

$$CTF(\vec{s}) = -\sqrt{1 - A^2} * \sin(\gamma(\vec{s})) - A * \cos(\gamma(\vec{s})) = -\sin(\Delta\phi + \gamma(\vec{s})) \text{ Eq. 2.5}$$

Given that the defocus and spherical aberration are the two factors that affect the phase shift Equation 2.6 can be used to represent the $\gamma(\vec{s})$ for simplified calculations, where s is the modulus of \vec{s} , C_s is the spherical aberration, λ is the electron's wavelength at a given accelerating voltage (typically between 100keV and 300keV for cryo-EM), and $f(\theta)$ is defocus in the direction described with the angle θ .

$$\gamma(\vec{s}) = \gamma(s, \theta) = -\frac{\pi}{2} C_s \lambda^3 s^4 + \pi \lambda f(\theta) s^2 \text{ Eq. 2.6}$$

As previously mentioned, the defocus is described by the three parameters which are the maximum and minimum values (f_1 and f_2) and the angle θ_{ast} between the long axis of the ellipse and the x-axis (see Figure 2.4 A) and can be described by Equation 2.7.

$$f(\theta) = f_1 \cos^2(\theta - \theta_{ast}) + f_2 \sin^2(\theta - \theta_{ast}) \quad \text{Eq. 2.7}$$

The CTF can be estimated by fitting a curve to the mean of rotationally averaged estimated power spectra of the micrographs in the movie which is one of the standard procedures of defocus determination in cryo-EM[67]. The other approaches use envelope function for different frequency bands[128][75] or more advanced Wiener filter implementation[129]. Figure 2.14 shows A) an example of the Contrast Transfer Function fitted to the Thon rings from a cryo-EM micrograph and B) one-dimensional fitting of the Contrast Transfer Function.

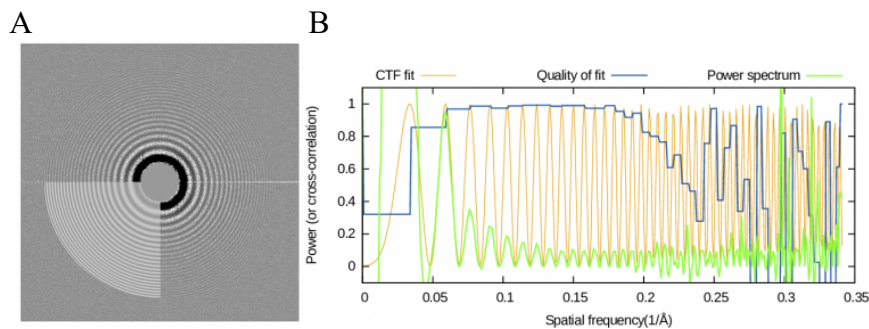


Figure 2.14 A) Contrast Transfer Function (lower-left quartile) fitted to the Thon rings, B) 1D fitting of the Contrast Transfer Function Reproduced from[75] Copyright © 2015 Elsevier Inc. All rights reserved. Reuse permission obtained from RightsLink order 5856541467989

The CTF can also be estimated from the shape of Thon rings, but for the cryo-EM data it is hard to determine their position and shape from the unfiltered image without previous knowledge about the parameters like defocus, astigmatism, and electron scattering. Inelastically scattered electrons, due to the high energy loss, are not correctly focused on the image plane because of the chromatic aberration of the microscope lens. This way, they introduce additional noise rather than improved contrast[130].

The effect of the inelastic scattering can be removed with the use of an energy filter for data collection, which reduces that noise. Unfortunately, not all microscopes are equipped with such filters.

2.1.7.3 Particle Picking

Particle picking is one of the crucial steps in cryo-EM data processing when the coordinates of each particle are obtained from the micrographs. The particle-picking strategies include manual and automated reference-free or template-based picking. The most straightforward technique is manual picking, where the users select the particles from the micrographs by themselves. As this approach gives some control over the quality of the particles, experienced users can avoid false-positive picks, try to pick a set of particles that represent sufficient angular views, avoid picking from too thick or too thin

areas and avoid false-positive picks it is not time effective as usually several thousands of particles are required to obtain a high-resolution cryo-EM map. The number of particles also depends on their quality and symmetry of the map. For example, a 2.8 Å icosahedral map of rhinovirus C was obtained from just 8973 particles[131], whereas 160 000 particles were needed to obtain a 3.4 Å map of asymmetric gamma-secretase[132]. The particle-picking task can be accelerated with automated picking tools. A hybrid method would require the user to pick around a few hundreds of particles and use them to create an initial reference for automated picking. In most cases this would have to be done for each dataset but might be useful for very uniquely shaped and sized specimens where the automated pickers would fail. The reference-free picking algorithms are commonly based on edge detection algorithms. The micrograph is scanned in multiple directions to find as many picks as possible. As the user can determine some input parameters such as minimum and maximum particle dimension, minimal distances between the particles or distance from the edge of the micrograph, still a lot of false-positive particles can be picked, which then would have to be removed at later processing stages. Template matching would require a good range of templates for particles in different orientations. This can be achieved by initial 2D classification of the dataset from template-free picking or angular sampling of the reference structure. The main issues with this approach are the time required to process large images, sensitivity to the noise in the images and the template bias. In recent years, with the rise of more advanced image processing and Artificial Intelligence methods, new automated pickers have been developed with picking models trained from historical data to optimise the number and quality of the particles.

The Laplacian-of-Gaussian auto-picker is a reference-free tool that allows picking the initial set of coordinates using the multidirectional Laplacian-of-Gaussian (LoG) filter. This algorithm is commonly used in image processing applications to locate the edges in the image. The Laplacian filter ($\nabla^2 I(x,y)$) works on the second derivative of the pixel intensities (I) of an image in x and y directions according to Equation 2.8:

$$\nabla^2 I(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \quad \text{Eq. 2.8}$$

For the discrete images, this can be estimated as a convolution kernel, which is a sum of two one-dimensional kernels in the x and y directions:

0	1	0
1	-4	1
0	1	0

The Laplacian is used to find the zero crossings of the second derivative and is very sensitive to noise. To alleviate this issue first, the Gaussian smoothing filter is applied to the image which removes the high-frequency noise and makes low-frequency features such as particle edges more prominent. This allows users to use a threshold to eliminate the detection of the zero-crossings with a weak magnitude, as they usually correspond to the noise[133]. The major disadvantage of using this method for the cryo-EM data is the fact that, as an edge-detection algorithm, it will result in a lot of false-positive picks, which could be the damaged particles, ice contaminations or the edges of the holes in the field of view. To curate the set of particles picked with LoG, the 2D classification job can be run. It should group together the similar views of the particles into distinctive classes and separate them from the false-positive picks. Then, the 2D class averages can be used for the second round of picking, now based on the template matching.

The structures that can be used as a template for particle picking can vary from basic geometrical shapes to very detailed 2D projections of 3D refined structures. cisTEM software uses a soft-edged disk, which speeds up the calculations and helps to avoid potential template bias, but is not effective if the particle is elongated in one direction[134]. The approach implemented in findEM software uses the cross-correlation coefficient which is calculated between the template and the particle candidate, but it is dependent on the local changes of intensity[135]. The approach implemented in earlier versions of Relion uses a model with the white Gaussian noise. The template matching can be done in real space using correlation methods or in the Fourier space by multiplication of the Fourier transform of the template image and complex conjugate of the micrograph. The rectangular areas of potential particles are separated into the ‘particle area’ inside the circular mask and the background noise outside the circular mask. The particles are normalised inside the circular mask to obtain the noise values with mean zero and standard deviation of one. Thanks to this, the particles are not dependent on the local changes in ice thickness, intensity levels or other experimental factors. Then, the ratio of the probabilities based on the maximum-likelihood is calculated for each particle to check if it represents one of the template images or the solvent noise. It is described with Equation 2.9 where $R_{\phi,k}(\vec{t})$ is the ratio for the particle in position \vec{t} in the image X considering the orientation ϕ , $P(X|\vec{t}, A_k^\phi)$ is the probability that the particle matches one

of the k templates and $P(X|\vec{t}, O)$ that it matches the solvent noise. The $\mu(\vec{r})$ is additive and $\sigma(\vec{r})$ the multiplicative normalisation factor used to equalise the intensity levels of the particles. \vec{r} is the coordinate system used for the whole micrograph. \vec{q} is the local coordinate system with the origin in the centre of circular mask M , therefore it can be calculated as $\vec{q} = \vec{r} - \vec{t}$. This approach is graphically presented in Figure 2.15, where panel A shows the representation of an X micrograph with some particles in the field of view, B shows the area of the box used to estimate the background noise and C the area which is expected to contain the particle and is normalised according to the background noise parameters[136].

$$R_{\phi,k}(\vec{t}) = \frac{P(X|\vec{t}, A_k^\phi)}{P(X|\vec{t}, O)} = \exp \sum_{\vec{q} \in M_i} \frac{X(\vec{q} + \vec{t})}{\sigma(\vec{r})} - \frac{\mu(\vec{r}) A_k^\phi(\vec{q})}{\sigma(\vec{r})} - \frac{1}{2} (A_k^\phi(\vec{q}))^2 \quad \text{Eq. 2.9}$$

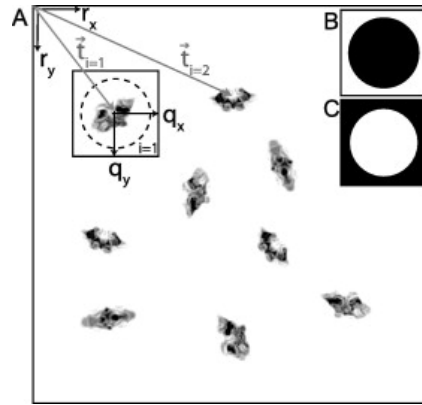


Figure 2.15 A) Representation of the cryo-EM micrograph with some particles in the field of view, B) mask used for the background noise estimation, C) a mask for the expected particle area Reproduced from [136] Copyright © 2014 The Author. Published by Elsevier Inc. This is an open access article distributed under the terms of the Creative Commons CC-BY license

The general issue with the automated particle picking procedures based on template matching is the sensitivity to matching the templates with noise called “reference bias”. This is commonly known in the cryo-EM community as the ‘Einstein from noise’ problem, where the portrait of Albert Einstein was used as a template for reference-based picking from 1000 images containing only the Gaussian noise, which resulted in a reconstructed portrait[137][138]. Among other known limitations of template matching is sensitivity to noise, especially if the templates have a higher signal-to-noise ratio than the objects or there are significant changes in background illumination, especially when run on lowpass filtered micrographs. They also lack the robustness to variability in the data, as the unique projections of the particles might be classified incorrectly or skipped.

The classification accuracy will also be lowered if there is heterogeneity in the data is not covered by the templates if particles are aggregated or occluded. Finally, as the template matching uses cross-correlation, processing large or complex data would be computationally extensive and might need to be considered when running cryo-EM processing pipelines. Considering all the limitations, in recent years, more advanced particle-picking algorithms have been developed, often using machine-learning approaches.

TOPAZ is modular software which offers tools to denoise the cryo-EM micrographs[139] and pick the particles in an automated way. In the pre-processing stage, the micrographs are normalised in order to minimise the varying imaging conditions, like different ice thicknesses and intensities. This can be done by the affinity normalisation, subtracting the mean value from the image and dividing it by its standard deviation or using the Gaussian Mixture Model. For picking, it uses Convolutional Neural Networks with Positive Unlabelled (PU) learning. For training, only a small batch of labelled particles from the micrographs is provided as true positive picks, as the other regions of the micrograph remain unlabelled. In the first step, the labelled particles are used to train the classifier, which is then applied to small regions of the micrograph to get the local predictions of the particle's presence and its coordinates. Then, the sliding window technique is used to pick the particles with the non-maximum suppression algorithm to avoid overlapping detections. Each particle position is evaluated with a log-likelihood score to reflect the machine learning model's confidence that the particle is present at the given location or if it is an artifact or a background noise. The higher log-likelihood score represents higher confidence in particle presence. This value can be thresholded only to extract the particles with a high confidence score, also considering the defined minimum distance between the particles to avoid multiple picks of the same particle[140].

Other popular approaches for particle picking use Convolutional Neural Network with sliding windows. The images are analysed by moving a fixed-width patch across them. The training of the network requires labelling both positive and negative examples (particles and non-particles). The sliding window method is sensitive to variability in the dataset, such as different particle sizes and shapes and might perform poorly in low signal-to-noise ratio areas, which may result in an extensive number of false positive picks and creates the need for additional curation step. Additionally, processing of multiple overlapping windows usually requires more computational time. To mitigate these issues crYOLO was developed, using the 'You Only Look Once' (YOLO) algorithm, which analyses the image as a whole, providing also the spatial context of the particle's

positions. In this case, for the training of the CNN, micrographs with correctly labelled true particle positions were used, as the method does not require negative examples such as labelled backgrounds or contaminants. As the image goes through the network, it is down-sampled to a smaller grid. Then each grid cell is evaluated if it contains the centre of the particle. If the confidence of this is high, the size of the bounding box and relative x and y coordinates inside the grid cell are estimated. As the network is used to analyse the whole micrograph at once, also information about the distance between particles and contaminants can be used to avoid picking from contamination regions. The loss function used for the backpropagation during training penalises the incorrect size of the boxes, incorrect particle centre coordinates, and incorrect confidence of particle presence in the grid cell. For the training, the same micrographs were passed through the network a couple of times, each time slightly altered (blurred, mirrored or with additional noise added)[141]. The general model was trained on over 60 different cryo-EM datasets. crYOLO offers an automated denoising procedure as a pre-processing step, either as a simple lowpass filter used by default or using a denoising tool JANNI[142], [143]. Additionally, if the automated picker performs poorly on a specific dataset, users can fine-tune the model, which allows users to update the general model based on a small subset of particles picked from a particular dataset[144]. Deep learning approaches, especially convolutional neural networks, can be implemented for automated particle picking with tools such as DeepPicker[145] or DeepEM[146].

Still the major disadvantage of the automated pickers based on machine learning algorithms is the lack of the large-scale ‘ground-truth’ data which can be used for the model training procedures, which in most cases is the manually labelled data. Therefore, the selection of the subsets used for the network training (from conventional manual or template-based picking) can introduce bias to the model or lower their performance on the dataset with particles of unusual shape previously not seen by the model[144].

2.1.7.4 2D Classification

The next step after the particles’ coordinates selection is the 2D classification. It is not as versatile and powerful tool as 3D classification but can be used to assess the quality of the data and obtain aligned class averages that can be used for the initial 3D model refinement. Particles with similar views are grouped into a single class. The algorithm considers the translation and rotation of the particles. The output 2D classes show the averaged image for each group of particles. At this stage, the user can evaluate the classes based on the particle distribution, pick the best ones and remove the remaining

contaminants or false positive particle picks from further processing[147]. Users can see if the 2D classes represent enough unique angular orientations of the specimen. In the recent Relion 4.0 implementation each 2D class has a quality score based on a machine learning algorithm, which considers metrics like the number of particles, and signal-to-noise ratio. Additionally, the ‘overall quality’ parameter was introduced by training the network with the historical data of what kind of classes users decided to keep[148].

2.1.7.5 3D Reconstruction

The success of 3D reconstruction depends not only on the quality and number of particles picked from the micrographs but also on the angular coverage of the projections. The 3D model is calculated based on the Fourier Central Slice Theorem stating that the Fourier transform of a 2D projection of the particle matches a Fourier transform of a specific 2D slice through the 3D density. During the 3D refinement, the relative angles of the 2D are iteratively calculated according to the initial or reference model. The more complete set of views in the collected data leads to more complete angular coverage, resulting in finer, more uniform sampling in the Fourier space and high-quality 3D reconstruction[149]. One of the biggest challenges at this stage is the preferred orientation. It happens when the particles have a biased distribution of orientations because of the interactions between the molecules or with the air-water interface. This issue is observed especially for low-symmetry macromolecules. To some extent, this issue can be mitigated by the addition of detergents to occupy the air-water interface, time-resolved vitrification or tilting the stage during the data collection.

Selected particles from 2D classes are used to generate an initial model, which will be used for 3D refinement. This model can be generated ab-initio, based solely on the experimental data. One of the methods to do this is Random Conical Tilt (RCT), which involves collecting a set of paired images where one of the images is recorded with the stage tilted. The known stage tilt angle makes it possible to determine relative orientations between the projections, which makes it a good starting point for iterative high-resolution 3D refinement[150]. Another approach used more commonly in the past is the Common Lines method. It is based on the principle that any two projections of the 3D object share common lines in their Fourier Transform. Identification of these lines leads to the determination of relative orientations between the projections. Unfortunately, with the low signal-to-noise ratio in cryo-EM images, the common lines are hard to identify or can be identified incorrectly, leading to errors in the 3D structure. Additionally, this method often cannot be used to correctly identify the handedness of the structure as it is not able

to differentiate between mirrored images[151]. Another method which can be most useful, especially when studying novel structures where there is no prior information available but also most time and computational resources consuming is iterative refinement based on the autocorrelation of the 2D class averages from a randomised initialisation[152]. Another approach would require a reference model from a known structure to be used as a starting point for 3D reconstruction, which can improve the initial rounds of refinement. The reference should be lowpass filtered to reduce the bias. If the overall shape of the specimen is known, a geometrical shape like a sphere for apoferritin or a cylinder for T20s proteasome can provide a reasonable trade-off between ab-initio and model reference reconstruction by boosting the initial processing stages and minimising the reference bias. 3D reconstruction is an iterative process of assigning the angular orientation to the 2D representations of the particles. After each cycle, the new model is used as a reference to improve the accuracy and resolution. Most of the optimisation algorithms used for 3D refinement require a reasonably good initial model to initialise the search for the global minimum. Otherwise, they are stuck in local minima. The Stochastic Gradient Descent (SGD) algorithm can be used to refine proteins from random initialisation. As the directional gradient between the objective function and the current model is calculated, the model's parameters get updated in the opposite direction of the gradient. It is used for initial refinement from low to high resolutions, usually achieving satisfactory results after several iterations. In the recent releases of Relion, SGD was replaced with the VDAM (Variable-Metric Gradient Descent Algorithm with Adaptive Moments Estimation) algorithm for 3D refinement. It is more robust to the noise, as the gradients can be averaged over iterations and achieve convergence in fewer iterations than other methods, also providing higher resolution of the initial model[153]. The 3D refinement methods in cryo-EM are based on the Fourier central section theorem, which describes the relation between a 3D object and its 2D projection passing through the origin in reciprocal space, as shown in Figure 2.16 [149]. It implies that once the 2D Fourier transforms of the projections are positioned into the 3D transform, the original 3D object can be reconstructed by calculating the inverse transform. The biggest challenge in the Single Particle Analysis is the fact, that the particles are distributed with unknown angular orientation in the micrographs. In some cases, many angular views might not be present at all due to the preferred orientation problem. The assignment of the angular orientation to the particles is done iteratively to gradually improve the final resolution of the 3D volume. The reference is angularly sampled to obtain the initial angular projections to which the particles are matched. The algorithm is optimised to

converge into local minima, which should ensure that the reconstruction matches the original model[154].

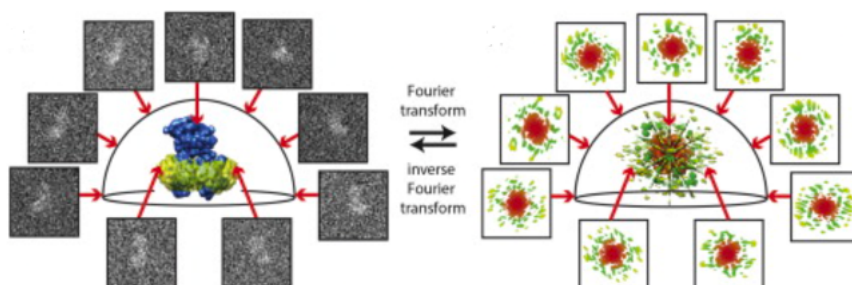


Figure 2.16 3D reconstruction from the 2D views obtained using Fourier slice theorem
Reproduced from [154] Copyright © 2015 Elsevier Inc. All rights reserved. Reuse permission obtained from RightsLink order 5856540294417

Once the 3D model is obtained, the particles can be further classified into 3D classes, which are effectively different 3D models obtained from subsets of the full set of particles. This can reveal heterogeneity in the dataset which can result in areas with lower resolution or preferred orientation, where specific views of the specimen are underrepresented, which can cause the resolution anisotropy, which is directional variability of resolution. As a result, some parts of the model might be blurry. To address this problem, one can use software tools developed to identify and group together the particles representing the specimen in different states. In Relion, users can perform multi-body refinement. This focused refinement requires a user-defined mask to determine the part of the volume which will be excluded from the reference structure. The masked region can define the flexible part of the molecule[155, p.]. As the local refinement shows improvement to the final resolution of the map, the automated identification of the regions which are poorly resolved is still a challenge. FlexEM uses segmented-based Manders' overlap coefficient (SMOC) as a metric to identify regions with low fitting scores at the different stages[156][157]. cryoDRGN uses deep neural networks to reconstruct heterogeneous datasets. The input requires particles stacks after 3D refinement. Then, the algorithm predicts the continuity of each particle and the data is decoded as 3D volumes based on the Variational Auto-Encoder (VAE) method[158].

Selected 3D classes with homogenous particles can be selected for high-resolution 3D refinement. The subset is divided into two non-overlapping halves and each of them is refined independently to avoid over-fitting. By comparing those two reconstructions, the quality and consistency of the refinement can be evaluated. The next step towards obtaining a high-resolution density map is the further sharpening of the map in a volume

defined with a binary mask, typically with a value of 1 representing the molecule density volume and 0 for the solvent region. To create a mask, it is recommended to low-pass filter the original map and set a binarization threshold to the value that eliminates all of the background noise. The final resolution is calculated only in the masked region, as the noise in the solvent can affect the resolution estimation and lower the Fourier Shell Correlation curves. Additionally, as a common practice, a soft-edge fall-off is added at the edge of the molecular density to provide a smoother transition between 1 and 0 values, typically as a sinusoid or Gaussian slope, which can help reduce artefacts in Fourier Transform. The final resolution of the cryo-EM map is routinely determined using the 'gold standard' Fourier Shell Correlation at the 0.143 threshold. Unfortunately, the mask for post-processing in most cases, has to be defined manually, especially when the specimen is not symmetrical or has irregular shapes.

2.1.7.6 Resolution of the Cryo-EM maps

The upper limit of the resolution is determined by the detector used to record the data. According to the Nyquist-Shannon theorem for signal processing, the highest possible resolution of the map, which can be obtained from the cryo-EM dataset, is twice the pixel size of the collected image. For example, if the data was collected with 1 Å pixel size, the highest resolution for the final map would be 2 Å to avoid aliasing, which introduces inaccuracies and distortions to the signal, resulting in the data loss, as due to insufficient sampling rate the high-frequency components are interpreted as low-frequency ones. In practical applications, the final resolution is also limited by other factors that affect the data collection, such as staining, defocus or beam-induced motion. The unit of resolution typically used in structural biology and cryo-EM is Ångström, which equals 0.1 nanometres. The final resolution determines the features that can be seen on the map. The length of the carbon-carbon bond is 1.5 Å thus, the maps with higher resolution allow us to see individual atoms, but even at lower resolution, some of the structural features can be seen. Figure 2.17 shows a summary of the features which can be observed at different resolution ranges. All maps are overlaid with the atomic models. The maps with resolutions lower than 10 Å show information only about the conformational changes and domain boundaries. They can still be used as an initial map for further refinement in the data processing pipeline. Maps with a resolution of 10 Å start showing more information, such as the alpha-helices, which become more detailed and precise at 4 Å. As the resolution improves to 6 Å also, beta sheets and RNA helices pitch can be seen. Maps with 4 Å resolution or better usually contain most of the information required to derive a

highly complete and correct atomic model as they show individual beta strands, the pitch of alpha helices and the side chains. The comparison of the different structural features available to recognise at different levels of the cryo-EM maps is summarised in Figure 2.17 [159].

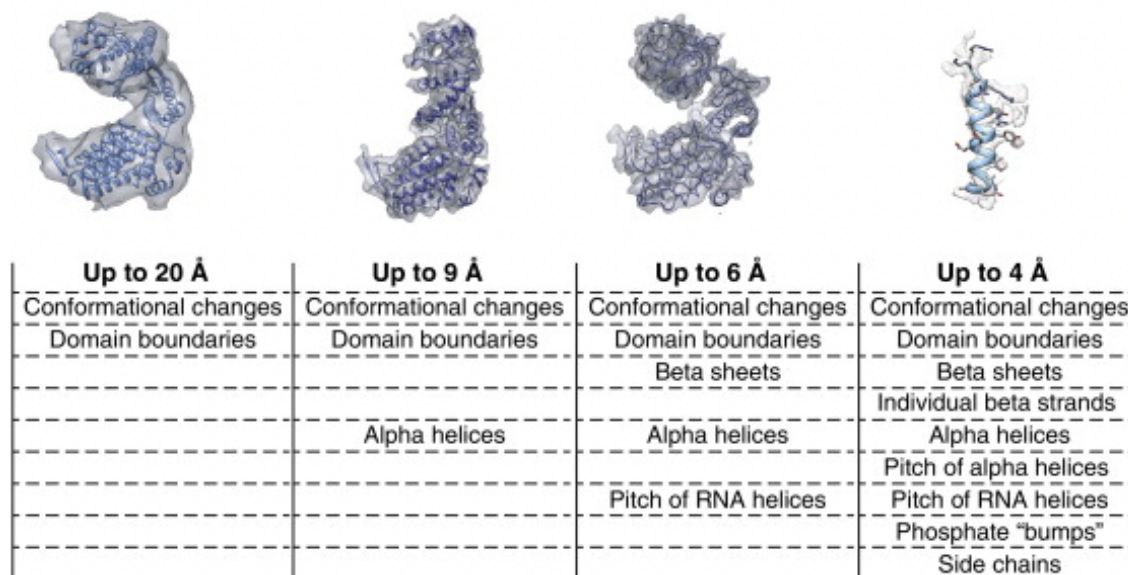


Figure 2.17 Summary of the structural features recognisable at different resolution levels of cryo-EM maps. Reproduced from [159] Copyright © 2014 Elsevier Ltd. All rights reserved. Reuse permission obtained from RightsLink order 5856540702268

The resolution of a cryo-EM map is routinely determined using the ‘gold standard’ Fourier Shell Correlation (FSC) criterion. The final reconstruction is divided into two independent subsets from which half-reconstructions are produced. Then, the FSC curve is calculated for the two shell reconstructions corresponding to different resolution levels. The final resolution is determined when the two reconstructions show a correlation of 0.143[122]. The ‘gold standard’ was introduced to counter overfitting the other methods were prone to thanks, to the usage of the two independent half-maps to avoid false correlations. Equation 2.7 describes the formula for FSC calculation where r is the radius of the shell, r_i is each individual voxel in that volume, F_1 and F_2 the reconstructed half-maps[160].

$$FSC(r) = \frac{\sum_{r_i \in r} F_1(r_i) F_2^*(r_i)}{\sqrt{\sum_{r_i \in r} |F_1(r_i)|^2 \sum_{r_i \in r} |F_2(r_i)|^2}} \quad \text{Eq. 2.7}$$

One limitation of the resolution estimation with the FSC criterion is that it is calculated across the whole density map. This can lead to an artificial under-estimation of the true

resolution due to the background noise contribution for not masked maps. Additionally, factors like the heterogeneity of the sample, damaged particles, or the preferred orientation problem can affect the global estimation of the resolution. The value of correlation cut-off has been debated for a long time in the cryo-EM community. To overcome these issues, other techniques to evaluate local changes in the map resolution were developed, which can check local and directional changes in the resolution across the map. Relion's implementation uses small soft spherical masks to estimate the FSC in different regions of the map[161], and then the results can be checked by colouring the map according to this parameter in UCSF Chimera[162] as shown in Figure 2.18. Other approaches like ResMap use the likelihood-ratio approach. The local resolution is defined by calculating the smallest value of a local sinusoid which can be detected from noise using the False Discovery Rate procedure[163].

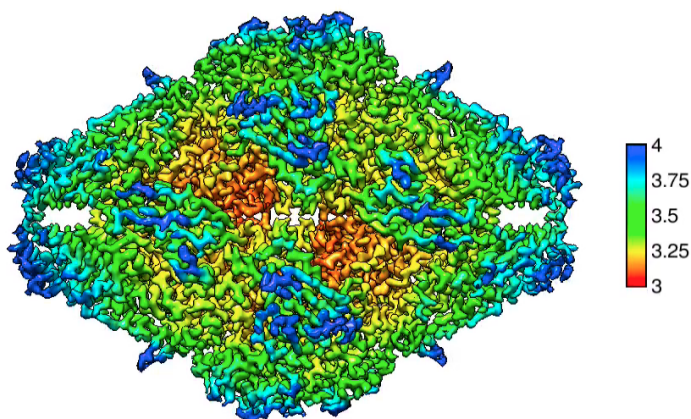


Figure 2.18 Beta-galactosidase map coloured according to the local resolution, the temperature bar shows the resolution values, from reprocessing EMPIAR-10204 dataset[164]

2.1.8 Atomic model building and validation

The final step of the cryo-EM data processing pipeline is atomic model building and validation. At this stage, the atoms are fitted into the cryo-EM density, which is represented as a mesh in the x, y, and z coordinates system. Cryo-EM density maps can be used directly for the model-building task. Before that, the map can be optimised with more advanced post-processing software. LocScale[165] is used to sharpen the map locally based on the reference structure, which can be a model of a similar protein or a partially built model. The goal of the optimisation step is to improve the overall quality of the map and reduce the noise before the atomic model is fitted into the cryo-EM density. This can prevent the automated model-building tools from placing the atoms into the background noise, especially at lower resolutions between 3 and 4 Å. Figure 2.19. shows a flowchart of the model building and validation procedure.

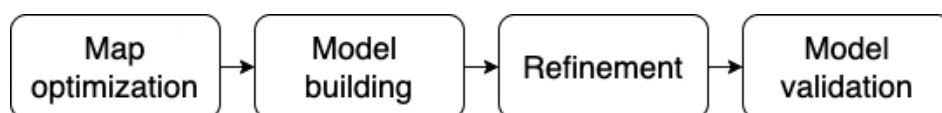


Figure 2.19 Flowchart of the map optimisation, model building and validation procedure

The software tools for automated model building in the cryo-EM maps are mostly derived from X-ray crystallography. The commonly used tools include Buccaneer[166], ARP/wARP[167] and PHENIX auto build[168]. Each of them uses different methods to resolve the structure and has strengths and shortcomings which should be considered when choosing the software to build from a particular map. The Buccaneer traces the possible C-alpha position based on the 6-dimensional likelihood-based target function based on the FFFear method[169]. The reference structure is used to generate an electron density map with features similar to those of the target map. The potential C-alpha positions are grown into chain fragments, then the overlapping fragments are joined, and clashing fragments are removed, considering the Ramachandran restraints. The next step is resolving the side chains. Buccaneer software can be run iteratively to improve model completeness after each round. In the CCPEM processing pipeline, after each round of the Buccaneer model building procedure, the refinement step with REFMAC5[170] is run and the updated model serves as input for the next round of model building. The Buccaneer is used at various target map resolutions up to 4 Å, as the software uses a 4 Å sphere to trace the C-alpha positions.

The ARP/wARP (Automated Refinement Procedure/weighted Automated Refinement Procedure) places dummy atoms in the electron density map and traces parts of the peptide chains. The partially built model is refined with REFMAC5 and another round of auto-tracing is performed. The automated peptide chain tracing and refinement are repeated until the atomic model is complete. The last step is to build the side chains.

The PHENIX model building is based on the input sequence. The target map is divided into smaller sections and the software builds models into each of them. Then, the overlapping parts of the models are merged and refined. Additionally, if the specimen is a combination of the multiple occurrences of the same asymmetric units, the model can be built only in one of the asymmetric units and then be expanded and superimposed on the target map, which makes the modelling more efficient.

With the rise of artificial intelligence in recent years, new methods and software for atomic model building have also emerged. AlphaFold is a groundbreaking tool developed by DeepMind for predicting protein structures from the protein sequence using machine

learning algorithms. It is also capable of predicting structures with DNA, RNA and ligands. Additionally, a more recent version of AlphaFold 3 can predict and model probable interactions between protein and potential drug candidates. Based on the input protein sequence, AlphaFold traverses multiple databases to obtain Multiple Sequence Alignment (MSA), which identifies all possible similar sequences across different organisms. From the MSA, crucial features of interaction between residues are extracted to identify parts of the sequence which would be in proximity when in folded state. Then, the deep neural network architecture iteratively refines the 3D position of the residues to improve the completeness of the final 3D model[171].

Relion 5.0 introduced a new feature for atomic model building incorporated into the processing pipeline. ModelAngelo is an automated tool which uses a Graph Neural Network (GNN) to combine the cryo-EM density map with a protein sequence. The cryo-EM map is segmented into voxels which then are evaluated if they contain C-alpha atoms for nucleic acids, phosphorus for nucleic acids or neither. The voxels with predicted C-alpha positions are then tested against all 20 amino acids to form a graph representation of the residues sequence. Residue positions are then optimised, combining the spatial relationships between them and information from the cryo-EM density and protein sequence to align the model with the map. In the final steps, the side chains are built, and each residue gets a confidence score, which helps to assess the overall quality of the generated model. The neural networks were trained on the pairs of cryo-EM densities and corresponding PDB models[172].

The validation step allows checking the completeness and correctness of the atomic model fitted into the cryo-EM density map. The correctness of the atomic model can be evaluated globally by overall fit-to-map metrics or locally by scoring individual residues. Commonly used techniques check the geometry restraints (Molprobit[173], CaBLAM[174]), map and model FSC (Refmac5[170], FSC-Q[175]), a local fit of the residues in density (TEMPy SMOC[157], FDR-score), evaluate the model by comparison to a reference simulated at given resolution (Map Q-Score[176]) or check the quality of protein-protein interface (PI-score)[177]. If the validation procedures show that the atomic model is still incomplete or incorrect, the map should be optimised again, and the whole workflow of model building, refinement and validation should be repeated until the final model has a high completeness score and the residues are built correctly. Table 2.4. compares different model validation techniques and which model characteristics they can evaluate.

Table 2.4 Summary of validation tools used for cryo-EM atomic model evaluation

Validation Tool	Evaluation
MolProbity	Global and local model geometry
CaBLAM	Global and local backbone geometry
PI-score	Protein-protein interface quality
REFMAC5	Global fit to cryo-EM density
TEMPy	Global and local fit to map
FDR-score	Backbone fit to the map
Map Q-score	Global and local fit to a reference simulated at a given resolution
FSC-Q	Map and model FSC

Initiatives like the EMDataResource Model Challenge allow benchmarking of commonly used model building and validation software and identifying which methods perform best in given cases[178], [179]. For model validation, it is always recommended to use at least two scores based on different methods to maximise the chances of identifying all of the incorrectly built parts of the model.

3 Development of a software tool for processing the cryo-EM data with non-uniform ice

This chapter presents further developments to the software tool IceBreaker for the estimation of the ice thickness in the cryo-EM micrographs. The main features of this software are the assessment of cryo-EM micrographs, associating particles with the local environment by evaluating relative specimen thickness and finally processing the cryo-EM micrographs based on local features, for example local contrast improvement. I will discuss the limitations of the introduced method and possible applications and present additional results, including mapping the estimated ice thickness to the measured values and processing the tomography data.

The first version of this software was published as ‘IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice’ and added as Appendix A. The Icebreaker software is implemented as part of the data collection pipeline at the Electron Bio-Imaging Centre at Diamond Light Source, Ltd.

3.1 Introduction

Noble et al. [105] have demonstrated that typical cryo-EM single particle specimens vary widely in the shape of ice meniscus and tilt landscape, resulting from micro-wrinklage of a cryo-EM grid during preparation and handling. Particles are largely attracted to an air-water interface, where they form a thin layer, and could be subjected to a “squeeze” if two opposite layers come close together, which could result in protein denaturation[180]. Bright-field cryo-TEM image records a projection of a specimen, with densities of an object of interest and media in which it is embedded integrated, resulting in variation of the contrast due to ice thickness or specimen tilt, both of which influence the length of electron paths through a sample. Thus, particle quality and angular characteristics could vary depending on the location it was sourced from. It follows that relative ice thickness is an important factor that contributes to the quality and resolution of the final cryo-EM map. Most of the automated data collection frameworks offer functions to easily identify bent and deformed grids on the atlassing stage, and also empty or cracked holes, which are not suitable for high-quality data collection at the hole targeting stage. These solutions use pre-calibrated I_0 value over vacuum to discriminate cracks and have been available for well over ten years now and are implemented in Legion[107], UCSFImage[108] and in the EPU from Thermo Fisher Scientific. As these allow the assessment of the overall

quality of the sample, different methods were developed to further estimate or measure the ice thickness in the areas selected for data collection.

One of the approaches to measure ice thickness in the cryo-EM data is to calculate tomograms from a tilt series collected from a target area. This is, however, the most time-consuming approach and adds complexity to a single particle data collection[105]. Another method, which also disturbs the data collection procedure is tilting the stage to +30 degrees, milling a small cylindrical hole through the ice with the focused electron beam, and tilting the stage to -30 degrees. The ice thickness can be calculated then from the projected length of the hole in the second image. This method is limited by the fact that the geometry of the burnt hole might not be ideally cylindrical and lead to inaccurate measurement[45].

The aperture-limited scattering (ALS) method is an alternative which does not require stage tilt. The ice thickness can be calculated from the intensities recorded with and without the sample using equation Eq.3.1 based on the Beer-Lambert law where d is the ice thickness, I is the total recorded intensity with the sample, I_0 is the intensity recorded in the absence of the sample, and λ is the free path for elastic and inelastic scattering which depends on voltage and the sample thickness. In practical applications of the ALS method the λ parameter can be limited by smaller objective aperture or by the use of energy filters.

$$d = \lambda \ln \frac{I_0}{I} \quad \text{Eq.3.1}$$

If the microscope is equipped with an energy filter, the relationship between the intensity recorded with and without the filter could be used (Eq.3.2).

$$d = \lambda \ln \frac{I}{I_{zfp}} \quad \text{Eq.3.2}$$

None of these methods, however, can identify the presence and impact of a local gradient in the micrographs since they typically return a single score. Meanwhile, modern methods of serial data acquisition place multiple acquisition areas inside a single hole, resulting in images from various places on the ice meniscus. Quantification and categorisation of such images might help further understanding of avenues for improving the efficiency of single-particle cryo-EM. The cryo-EM reconstruction process can be complicated by the variability of particles' quality and conditions. Proper classification of the particles picked

from the areas with similar ice conditions can lead to better distinction between particle states or conformations and obtain better homogeneity within the subsets. Processing a dataset with heterogeneity or preferred orientation, where the particle projections are severely imbalanced, can lead to the lack of structural details and blurring in the final 3D map and, as a result, significant local resolution variations reflected in a lower overall estimated resolution. Creating subsets of particles with similar conformations or similar angular orientations can reveal more structural details after image averaging and improve structure interpretability, revealing details such as the secondary structure or side chains of the residues. This approach can also lead to optimisation of the required computational time by choosing particle quality over quantity.

3.2 Methods

3.2.1 Average pooling

Cryo-EM micrographs are primarily affected by Gaussian and Poisson noises. The Gaussian noise arises from the support grid and amorphous ice surrounding the specimen in the sample or can be introduced during the digitisation process as the analogue signal is converted to digital. The Gaussian noise contributes to the overall greyscale of the image. The Poisson noise, on the other hand, is connected to the quantum nature of the electron detection process. The noise is proportional to the square root of signal intensity so the variance of the noise would increase with the increase of the number of detected electrons, making it hard to localise low-contrast particles[181], [182].

One of the commonly used techniques to reduce spatial dimensions of the dataset features is max pooling. It selects the maximum value from a defined region. It can be useful in scenarios focused on finding the highest activation signals, like edge detection. In cases where the Gaussian or Poisson noise is present in the signal and the signal-to-noise ratio is very low, max pooling would always select the highest value, which often corresponds to the noise instead of meaningful signal features.

For the task of analysing the background features of the cryo-EM micrographs the average pooling approach seems to be more suitable compared to max-pooling. Average pooling can smooth the image and reduce the effect of noise as it gives equal importance to all of the data points in a specific region. It is especially useful for the identification of anomalies in the images if they are spread over the image, such as the changes in ice thickness gradient. In this case, average pooling preserves more information in each

region compared to max pooling capturing the most prominent features, which could result in the over-representation of noisy features.

Average pooling is a method of image processing, which is commonly used for tasks like segmentation, or convolutional neural network training. It allows data reduction by down-sampling the features of the maps (Fig.3.1). For each defined region, the average value is calculated according to Equation 3.3, where x_i is the intensity at the given pixel, and N is the total number of pixels in that region. The regions are defined by kernel and stride. A kernel (or a pooling window) is a 2D structure determining the area over which the average will be calculated. Stride defines the step size for the moving of the pooling window over the original image; if the stride is smaller than the kernel dimensions, there will be an overlap in the average value calculation. In practice these two parameters will directly affect the results of pooling.

$$f_{avg}(x) = \frac{1}{N} \sum_{i=1}^N |x_i| \quad \text{Eq.3.3}$$

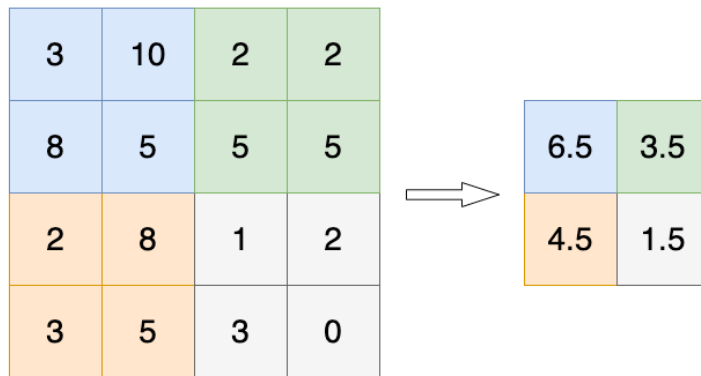


Figure 3.1 Example of 2x2 kernel applied to 4x4 matrix for average pooling, with stride 2.

For the task of analysing the background features of the cryo-EM micrographs the average pooling approach seems to be more suitable compared to max-pooling. Average pooling can smooth the image and reduce the effect of noise as it gives equal importance to all of the data points in a specific region. It is especially useful for the identification of anomalies in the images if they are spread over the image, as the changes in ice thickness gradient. In this case, average pooling preserves more information in each region compared to max pooling capturing the most prominent features, which could result in the over-representation of noisy features.

3.2.2 K-Means Clustering

K-means clustering is a method used in image processing for image segmentation. It iteratively associates the data points into independent clusters based on the minimisation of squared distances between data points and the current cluster centroid. The cluster centroids are updated each iteration to minimise the sum of squared distances (variance) between the data points and centre of the cluster until it reaches convergence when there are no changes in cluster positions after a defined number of iterations. The sum of squared distances for the defined clusters is minimised according to Eq.3.4

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 \quad \text{Eq.3.4}$$

Compared to another commonly used clustering algorithm, K-Nearest Neighbours (KNN), with K-Means user specifies the number of clusters that the dataset should be grouped into, with the centres of clusters initialised as new data points, whereas with KNN, the number of neighbouring particles to consider during classification is specified. Figure 3.2. shows the idea of clustering a 2D plane into smaller regions based on the distance from a set of points called ‘seeds’ This representation is called a Voronoi diagram. Each defined region groups together all points which are closer to the corresponding seed than to any other seeds.

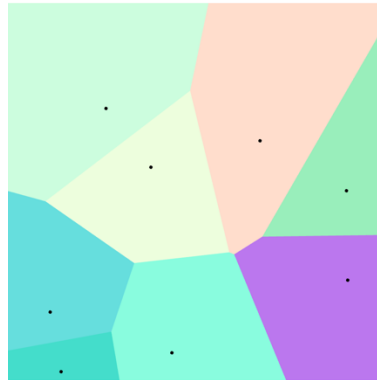


Figure 3.2 Voronoi diagram with random seed points (black), dividing the plane into 8 individual regions

3.2.3 IceBreaker algorithm description

The Icebreaker consists of two main Python scripts for image processing. The first one, *icebreaker_groups_multi.py*, estimates density gradient values caused by ice thickness variation independently for each micrograph. The second one, *icebreaker_equalize_multi.py*, can be used for local contrast improvement using histogram equalisation. The suffix ‘multi’ is used to identify the script that can process

multiple images simultaneously across the defined number of CPUs to speed up the processing. In the previous implementation, both scripts shared the same pre-processing steps. To reduce the required computational time the pre-processing steps are now adapted for each script separately. This section will focus on the script for ice thickness estimation. The contrast equalisation will be described in a separate section, as the output files are significantly different.

1. The required input for this type of job can be defined as a path to a folder containing motion-corrected mrc files. Within it, a subfolder called ‘/grouped’ is created to store the output files.
2. Based on the specified number of patches the image is average-pooled to reduce the size. The patches, representing now the average value in a specific region can act as super-pixels used for clustering. Additionally, at this stage, the data is also downsized 20 times in x and y direction.
3. The next step is image segmentation. The data is reshaped into a 1D array, which allows for segmentation based on the pixel intensity. This way, the image is clustered along the z-axis rather than based on the x and y coordinates. It allows identification of the parts of the micrograph with similar intensity even if they are not directly connected, which is especially useful when dealing with contaminants or aggregates that can appear at any place in the field of view as opposed to the hole edges which are expected at the corners and edges of the image. The resulting segments do not overlap. They can be used as independent entities and accessed directly for local processing. After segmentation is done, the array is reshaped back into the 2D image.
4. The segments defined in the previous step can be iterated over. They are used as a local mask for the image before clustering. The average value within each masked segment is calculated. This step is not computationally exhaustive as it is performed on the down-sampled image but provides an additional check of the mean values in specific regions after the K-Means algorithm selects the values for centroids. Now, each segmented region is associated with the mean value from the micrograph.
5. The output mrc file is reconstructed by combining the segments into one image, which is represented by the number of values corresponding to the specified number of clusters. The files are named the same as the corresponding original images with the suffix ‘_grouped’.

The resulting images can help in assessing the quality of specific micrographs from which they were obtained, track the changes in illumination across the dataset, and associate particle coordinates with the estimated ice-thickness levels. It is recommended to obtain these images as early in the processing pipeline as possible, as one feature is to identify micrographs representing damaged areas or containing large contaminations. This can be achieved with the *five_figures.py* script which provides box-plot analysis for each micrograph representing the overall distribution of the intensities and outliers. The outliers in regions of low intensity (darkest areas of the micrographs) are associated with high-density contaminants or edges of the hole in the field of view.

3.3 Results

3.3.1 The number of clusters and the execution time

Currently, the number of non-overlapping patches in the x and y direction is pre-defined in the code as 20 x 20 with 16 clusters. This is usually enough to characterise the overall shape of the gradient in the single particle cryo-EM micrographs since a relatively narrow range of magnifications resulting in a pixel size 1.2-0.8 Ångstrom per pixel is used in most cases and identify hole edges, contaminations and aggregates. The high number of clusters at this stage is not required to characterise the image; in future, it might be useful to set those parameters at the beginning of data collection based on parameters like specimen size or some additional information about sample conditions from screening. The currently used setup allows processing a single micrograph in around 0.8s on a workstation with Intel (R) Core (TM) i5-8250U CPU @ 1.60 GHz x 8, 8 GB RAM, which can be further improved with parallel processing. The comparison of how the number of patches and clusters affects the resulting segmented image is presented in Table 3.1. The setup with 10x10 patches with 4 clusters is the fastest to compute but does not provide enough information about local changes in intensity. On the other hand, 40x40 patches with 16 clusters can provide the most detailed information but also is the slowest of this comparison. This level of detail might not be fully used for the task of analysing the ice distribution as the point of this task is to provide a simplified representation of the micrograph. A reasonable trade-off between the computation time and the level of detail is selected to be 20x20 patches with either 8 or 16 clusters. Fig 3.3. shows how the computational time required to process a single image changes with an increasing number of patches and clusters. In general, it is recommended to keep the size of the patches used

as super-pixels at least twice the size of the specimen in each direction. Otherwise, some of the created super-pixels might represent only the density connected to the particles without the background information, which can introduce local disturbances in the segmentation. This can be especially important when collecting data, including large proteins or viruses.

Table 3.1 Effect of different numbers of patches and clusters on the image segmentation

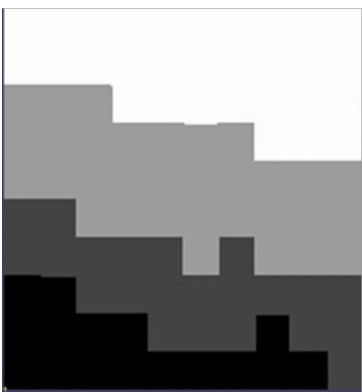


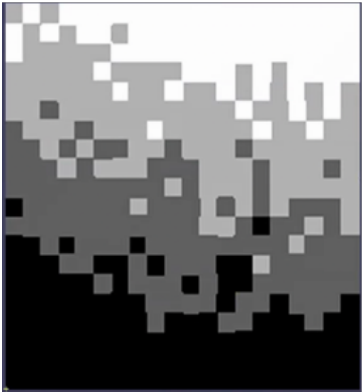
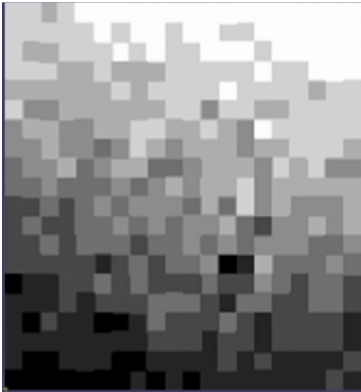
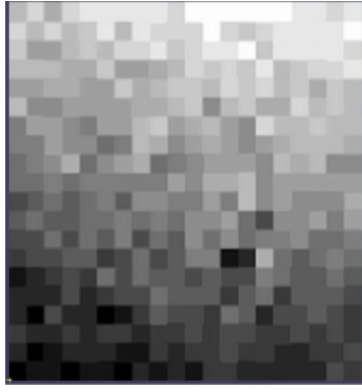
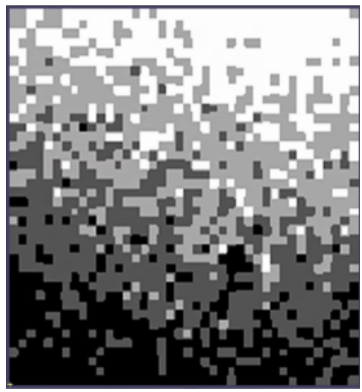
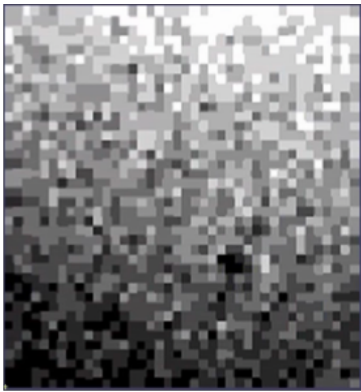
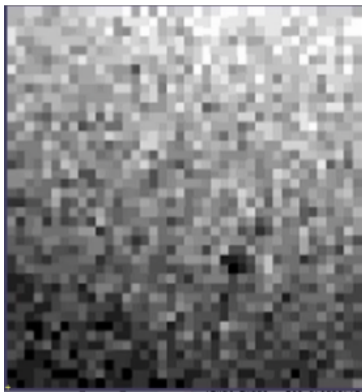
No. of patches	No. of clusters	4	8	16
10x10				
20x20				
40x40				



Figure 3.3 Segmentation parameters vs execution time

3.3.2 Assessing the quality of the micrographs

The IceBreaker grouping job is supported by two Python scripts for further data analysis. The first one allows a five-figure analysis of the output files. The minimum, maximum, median and Q1 and Q3 (first and third quartile of the distribution) values are reported for each micrograph to allow quick and easy inspection of the overall distribution of intensities in the micrographs. Mostly flat images would follow the standard distribution, whereas an ice gradient would introduce skewness, negative - coming from the dark areas of the image or positive if the image is overly illuminated in some parts. Five-figure analysis also reveals the outliers in each micrograph. At the previous stages of the processing, the images were averaged. Therefore the outliers in the low-intensity areas can be associated with aggregates, contaminants and hole edges, rather than with dead pixels or measurement errors. The output from this script is a CSV file containing a name and a figure summary for each micrograph.

This is a convenient representation for sorting, filtering and plotting. Fig 3.4. shows the recorded median values for collected micrographs. Four micrographs were selected based on their median value, two representing relatively thick ice (B, D), intermediate (C) and

thin (D). Additionally, the corresponding hole images at lower magnification were checked. These show that the hole illumination corresponds to the estimation at the micrograph level. The sample was prepared on a holey carbon support.

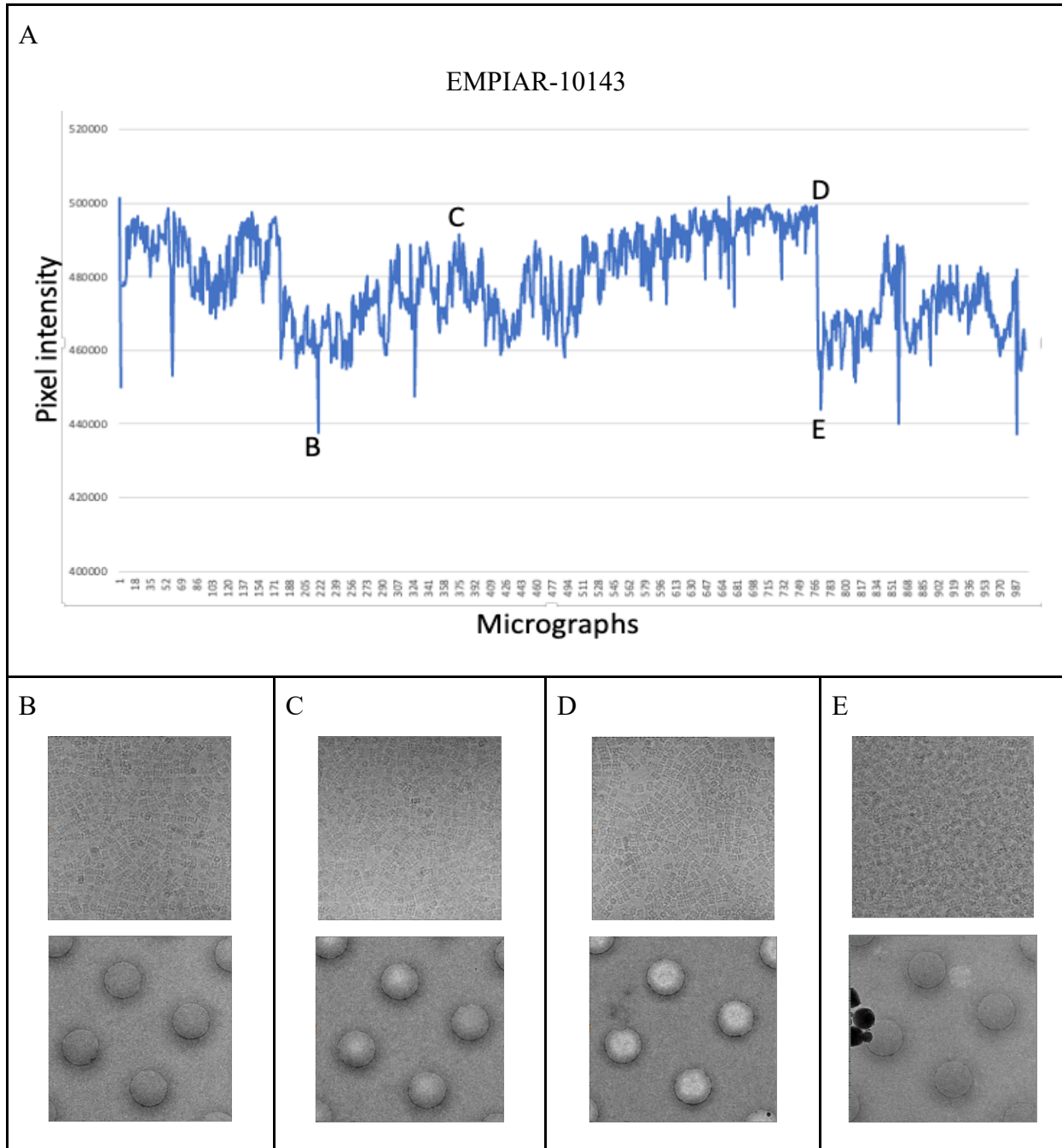


Figure 3.4 Micrograph quality assessment using median of pixel intensity. A) The changes in median values of micrograph in the whole dataset reveal a large spread in calculated values, B)-E) micrographs selected from different intensity levels with different ice conditions B) and E) thick ice, C) intermediate ice, D) thin ice, all accompanied with the view at lower magnification level showing similar ice features.

3.3.3 Estimated and measured ice-thickness

To establish the relationship between the pixel intensity values detected by IceBreaker and the actual ice thickness in the Single Particle Analysis experiments the EMPIAR-11437 dataset published with the paper ‘CryoEM micrographs of mouse apoferritin in a range of ice thicknesses on different microscope setups’[183] was used. The dataset was collected on a 300 kV Krios microscope with a K3 detector in counting mode. The deposited data was split into separate groups based on the ice thickness and grouped into ice thickness bands 0-50nm, 50-100nm, 150-200nm and 200-500nm. The ice thickness was calculated based on the average intensity of each image with and without an energy filter of 20eV according to Equation 3.2 with the electron mean free path estimated at 395nm. The dose-weighted images collected with the energy filter were used as input to the IceBreaker for the ice estimation task. Plotting the median values of the images from each ice thickness group shows mostly distinctive groups with only a few cases of overlapping as presented in Fig.3.5 A. As expected, thinner ice corresponds to the higher intensity values also there is more spread of relative intensities in the group covering 200-500nm thickness. Around 30 images (excluding overlapping ones) were selected from each ice band, and the median pixel intensity from IceBreaker was compared to the measured thickness (Fig.3.5 B). The pixel intensity was then plotted on a logarithmic scale (Fig.3.5 C), which allowed the fitting of a linear function for that sample. As this relationship was derived from the dataset collected with and without an energy filter, the accurate calibration of the pixel intensity to the actual ice thickness in future experiments would also require access to the microscope with an energy filter.

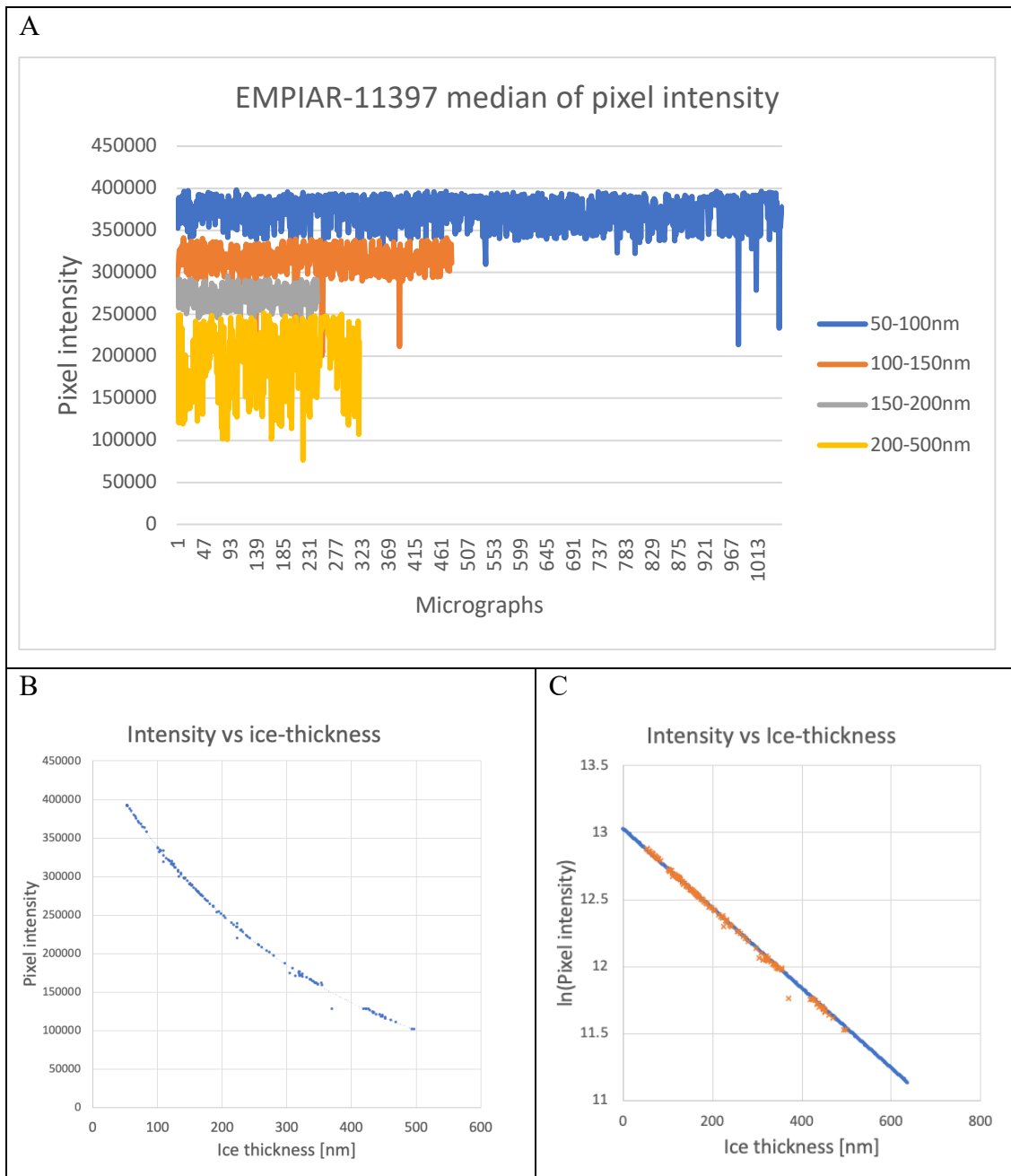


Figure 3.5 Relationship between the pixel intensity values from IceBreaker and ice thickness measured with energy filter, A) plots of median intensities in the micrographs, B) relationship between the pixel intensity and ice thickness for selected subset of micrographs, C) linear function fitted for the subset of micrographs with the ice thickness plotted against $\ln(\text{pixel intensity})$

This analysis shows that the relation between the pixel intensity and the actual ice thickness can be established by collecting small sampling data sets from various ice thickness regions, which normally are evident from an atlas, before the data collection using the ice measurement with the energy filter method, by fitting a linear function formula for the relationship between $\ln(\text{Pixel_intensity})$ and the ice thickness in nanometres. For this, one should take special care to collect the reference data from good

quality micrographs, not damaged and ideally without the edges in the field of view. The proposed framework for the calibration of ice thickness can be relatively easily estimated in the future by an automated script, which as input would take the mean free path of the electrons on a specific microscope, a set of pairs of average intensities recorded with and without filter (at least two pairs, but higher number can make the estimation more precise), and the median intensity calculated with the IceBreaker. The calculations are based on the linear regression method. The input dataset is split into 80% training and 20% testing subsets. This approach requires a higher number of testing samples, but as the linear relationship is expected, around 50 images should be sufficient. As the output, the script would produce a linear function equation to calculate the thickness for the rest of the micrographs and metrics for Mean Absolute Error and Root Mean Square Error for the evaluation of the quality of fit.

Additionally, the presented analysis confirmed that even without direct measurement and calibration to the actual ice thickness, the IceBreaker is able to sort the micrographs based on the relative ice thickness, which can be useful for the analysis of the data collected without an energy filter. In such cases, users can split the data arbitrarily into a number of selected groups based on the minimum and maximum values (after removal of the outliers) or simply target the highest recorded intensity values. This approach was also presented in the IceBreaker original paper for the processing of T20s proteasome based on minimum-maximum sorting rather than actual ice measurement, which is similar to the reported apoferritin reconstructed from different ice thickness and showed that for highly symmetrical particles, the thinnest ice produces highest resolution. Given the current limitations in sample preparation methods and challenges to obtaining uniform and desired ice thickness across the grid, estimation based on the pixel intensity can be useful. It is important to note here that an absolute ice thickness number is meaningless without understanding the specificity of a sample. Therefore, our approach has always been to prioritise categorisation over quantification, implying that a useful range of reported ice groups should be determined from a comparative data analysis by a user. The description of means provided for such analysis is outlined in the next chapter.

3.3.4 Considerations and limitations of the proposed method for ice thickness estimation related to microscope and image collection setup

The estimated ice thickness values are based on the pixel intensity in the collected micrographs. Some of the data collection and microscope setup can affect these values. To investigate this further, parameters such as defocus, dose weighting, use of energy

filters and the presence of crystalline contaminations were analysed. The defocus value determines the degree to which the microscope objective lens is out of focus. It is used to modulate the CTF function. By combining data with different defocus, one can reconstitute the most complete frequency spectrum representation of the data. During the cryo-EM experiment, the micrographs are collected with the range of defocus values typically between 0.5-3.5 microns. To investigate how the different defocus values correspond to the different ice thickness levels, the EMPIAR-11397 dataset was processed. This dataset was collected with the energy filter to measure ice. The defocus value was estimated for each micrograph with CTFFIND4 software. The distribution of these values in different ice thickness bands is presented in Figure 3.6. This analysis revealed that there is no clear relationship between the defocus setup in data collection and the ice thickness.

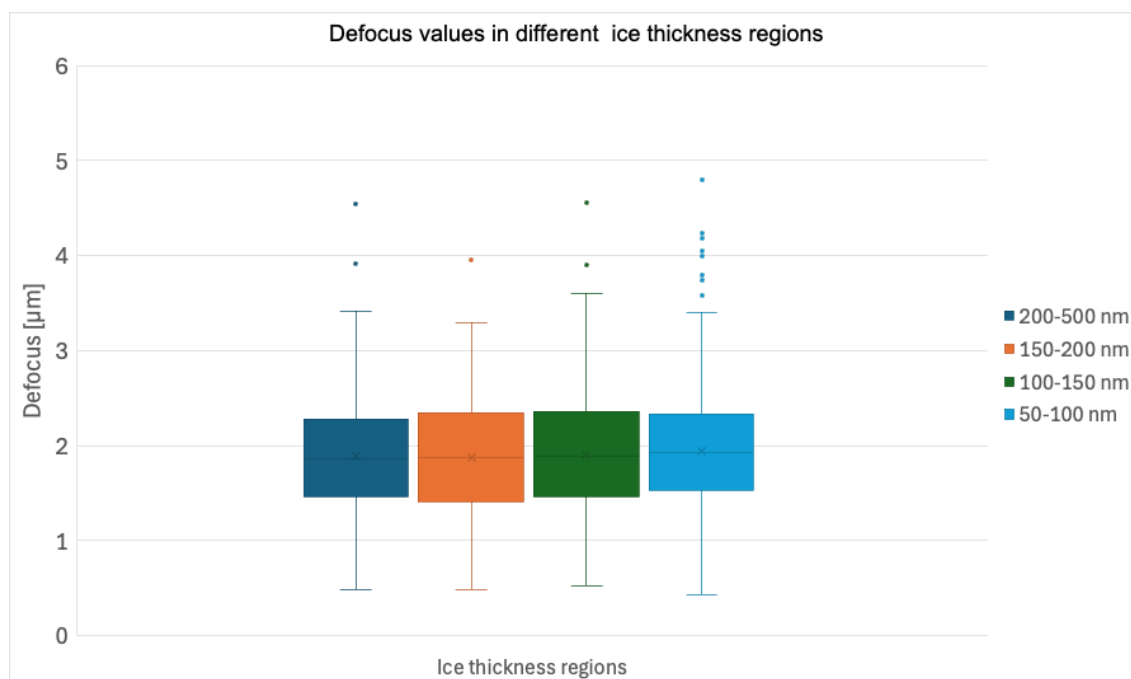


Figure 3.6 Defocus values distribution estimated with CTFFIND4 software in different ice thickness regions. For the range of the IceBreaker determined pixel values in every group (see Fig. 3.5 A).

The energy filter removes inelastically scattered electrons with energy that is different from the original beam by the specified value (typically 20 eV). Figure 3.7. shows the comparison of the pixel intensity estimated with and without energy filter in different ice thickness regions (based on EMPIAR-11397). The average intensity in the unfiltered dataset is higher than after filtration, as the filter removes some of the electrons. It still shows distinctive groups corresponding to the measured ice regions. The filtration also removes some noise from the micrograph, reducing micrographs overlapping different ice thickness regions. The recommended approach would be to use filtered micrographs

for ice estimation due to reduced noise and outliers. However, proper calibration should also make it possible to use unfiltered data.

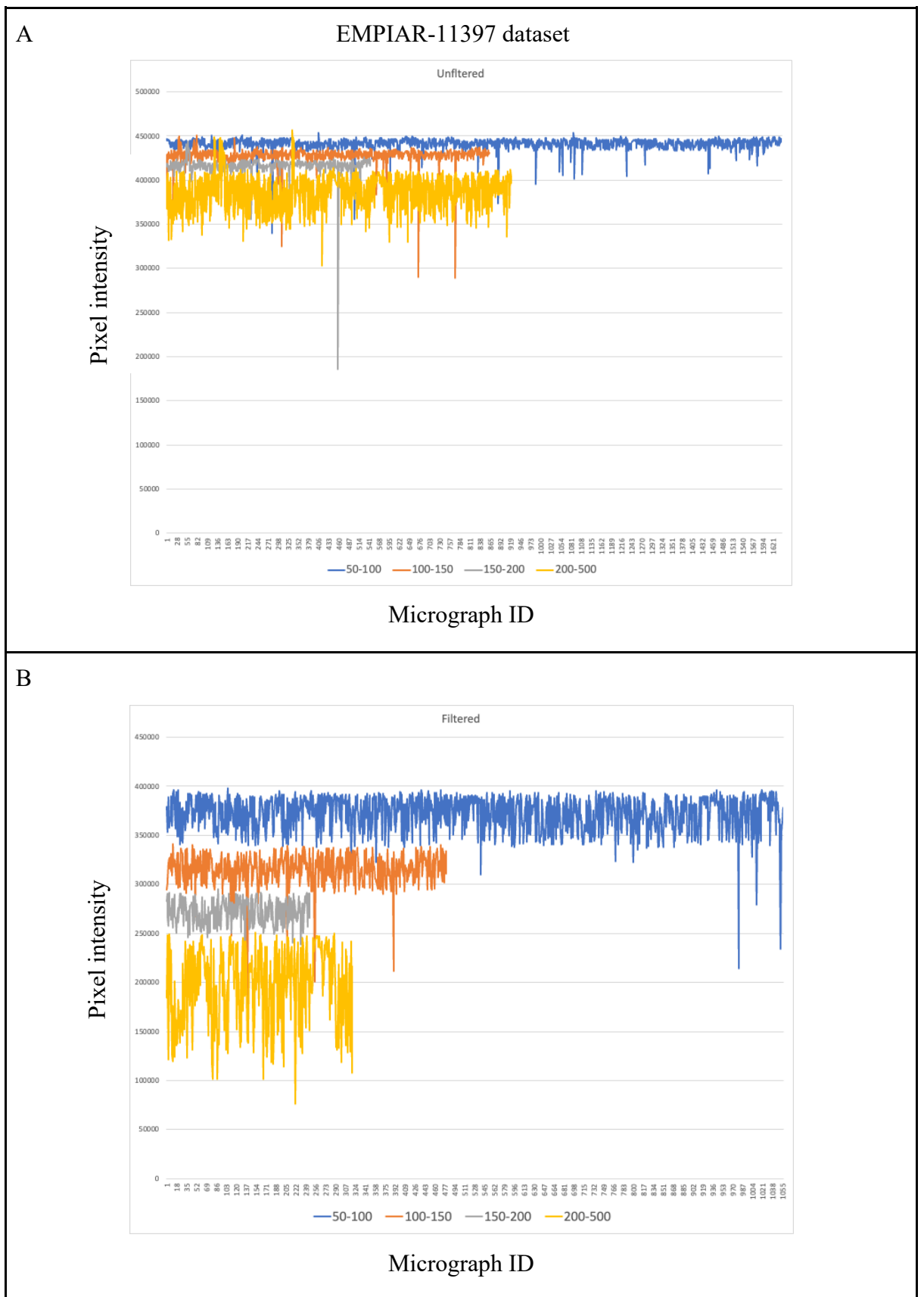
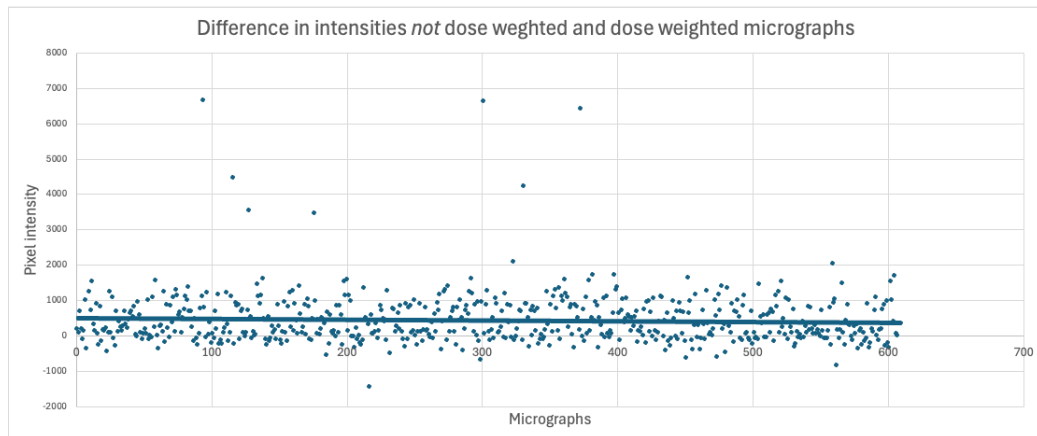


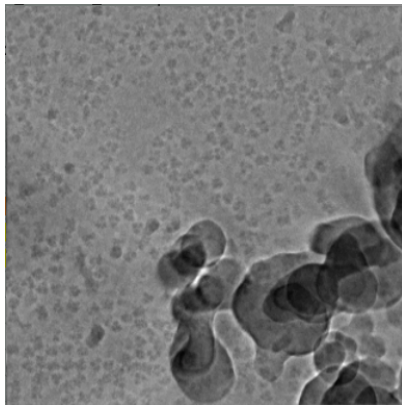
Figure 3.7 Comparison of pixel intensities in different ice thickness regions with (A) and without (B) energy filter.

The dose weighting is a procedure applied to reduce the effects of progressing radiation damage during data collection. Radiation damage in single particle cryo-EM can be quantified and propagates from higher resolution features to lower resolution over the specimen exposure to the electron beam[59]. Modern algorithms perform on-the-fly dose weighting, which suppresses high-frequency Fourier domain based on accumulated dose, while preserving low frequencies less prone to decay and required for object detection. The dose weighting is typically done with motion correction. This potentially could have an effect on the overall grey values of the resulting micrographs, thus its influence on IceBreaker performance needs to be investigated. The not-dose-weighted frames are rarely deposited to the cryo-EM repositories. To check the effect of dose-weighting on the estimated ice thickness, the EMPIAR-10132 dataset, which has such data, was analysed. Figure 3.8 shows the difference in median intensities between not dose-weighted and dose-weighted micrographs processed with the IceBreaker software. With the average estimated intensity of 150000, the difference of 1000 is less than 1%. The images with the biggest differences were additionally investigated to reveal that the highest positive difference results from significant ice contaminations in the field of view and the highest negative differences from blurry images, possibly with the collapsing ice (Figure 3.8 B, C).

A



B



C

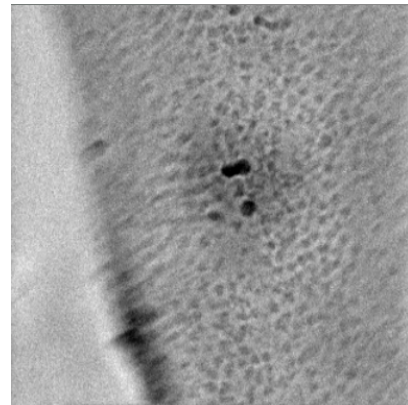
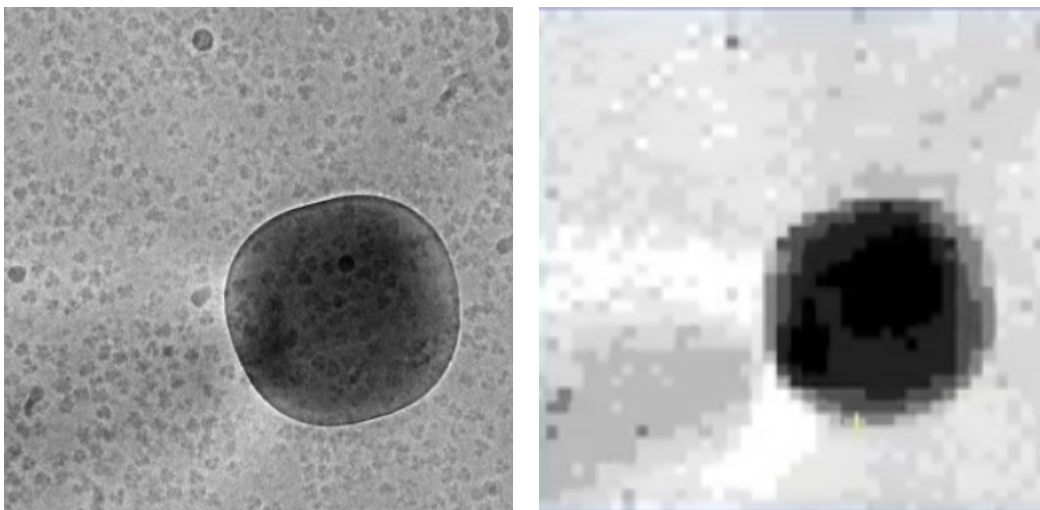


Figure 3.8 A) Differences in intensity values between dose-weighted and not dose-weighted micrographs, B) example of an image with a high difference, C) example of an image with one of the lowest difference

The crystalline contamination, if present in the micrograph, not only obstructs the field of view but also introduces additional electron scattering. Similarly to the foil-hole edges, it has a significantly different density than the micrograph background. Figure 3.9. shows images with crystalline contamination and the edge of the hole before and after segmentation. In both cases, the artefacts were clearly separated and associated with much lower intensity. Moreover, in such cases, the overall intensity distribution across the image will be heavily affected, and micrographs like these should be easy to identify and eliminate with the `five_figures.py` analysis script developed with the presented software.

A



B

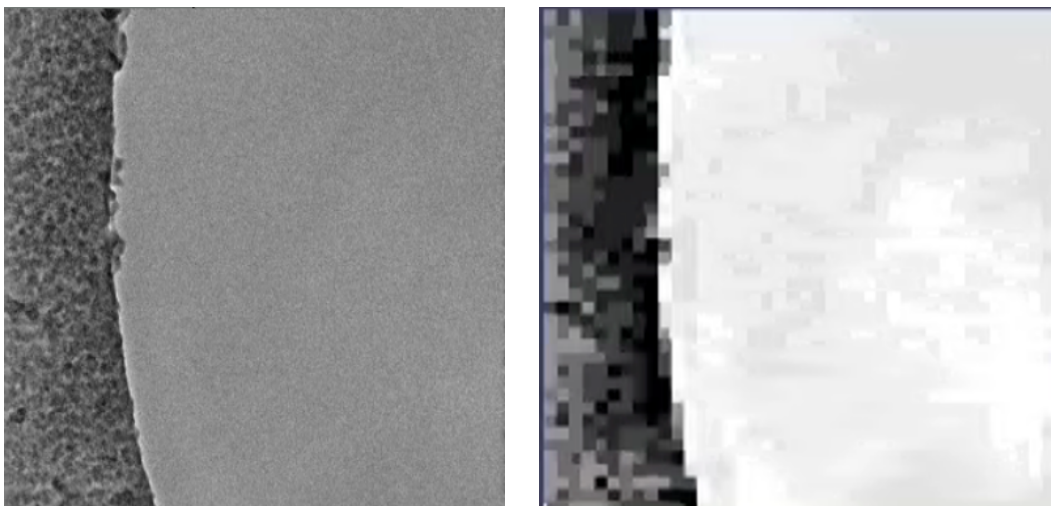


Figure 3.9 Micrographs with crystalline contamination (A) and foil-hole edge (B) in the field of view before and after segmentation.


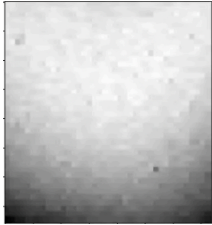
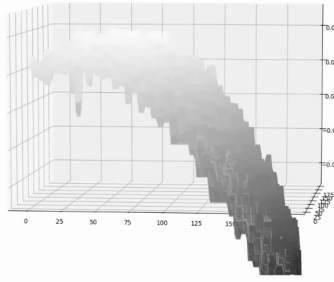

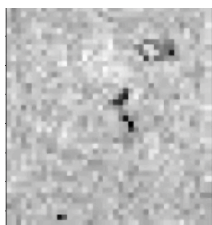
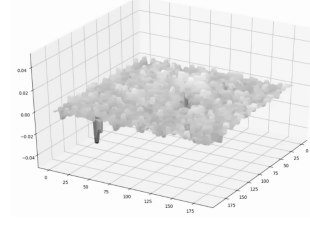

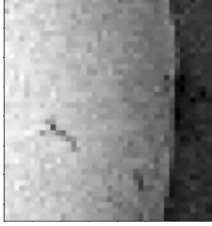
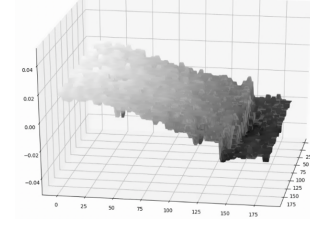

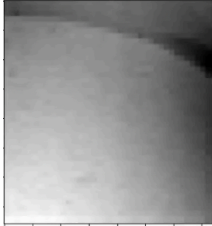
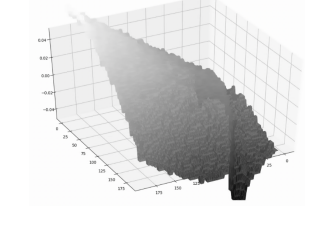
This software was not thoroughly tested for datasets collected with the Volta Phase Plate. The plate introduces a phase shift of the scattered electrons in order to amplify the contrast of particles against the background noise. It is especially effective at low spatial frequencies, enhancing phase contrast and hence improving the visibility of the particles outlines. The contrast improvement is not constant as the Volta potential evolves over time in the phase plate material[119]. Moreover, the ice thicker than 100 nm can introduce additional scattering, resulting in blurring and lower contrast which can reduce the benefits from using the phase plate. The Volta Phase Plate is also reported to dampen the high-frequency resolutions, which is another factor to consider during data collection

setup. Handling the datasets collected with phase plates with the IceBreaker, if possible at all, would require additional calibration to accommodate for the changes in intensities resulting from phase shifts.

3.3.5 Ice distribution estimated from cryogenic electron tomography dataset

In the attempt to connect the estimated ice thickness values with an actual measurement, a couple of datasets from EMPIAR database were reprocessed using IceBreaker. The idea behind this was to trace the changing intensity at the defined positions as the stage is tilted and relate it to the measured thickness. Unfortunately, all images in the deposited tilt-series were standardised to mean 0 and standard deviation equal to 1. The authors of the paper[105] confirmed that data before standardisation is not available. This indicates limitations of the proposed technique for the comparative pixel intensities analysis - it would work well with raw data but not normalised images. Instead, the IceBreaker was used to analyse the ice distribution profile based on a single image from the tilt series to compare it with the reported ice meniscus shape of the sample. Table 3.2 presents the results for samples from datasets EMPIAR-10129, EMPIAR-10137, EMPIAR-10138 and EMPIAR-10139. The high pixel intensity values correspond to the thin ice and low values to thick ice areas. Even though, with the missing data, it is not possible to estimate the relationship between the pixel intensity and measurement, the IceBreaker can successfully identify the overall shape of the sample and if it is flat, or if there is a meniscus or a slope from a single 0 degree image, while the authors of the original paper needed to collect, reconstruct and then evaluate tomogram to retrieve this information. In the future, this feature for on-the-fly ice profile analysis can be used to inform the selection of the best areas for data collection.

Table 3.2 Pixel intensity profiles estimated from tomography datasets

Sample ID	Reported ice profile	Clustered image	Estimated 3D ice profile
Sample 04 EMPIAR-10129	<p>4*† Hemagglutinin</p> 		
Sample 35 EMPIAR-10137	<p>35*† Apoferritin</p> 		
Sample 36 EMPIAR-10138	<p>36*† Apoferritin</p> 		
Sample 37 EMPIAR-10139	<p>37*† Apoferritin (1.25 mg/mL)</p> 		

3.3.6 Mapping particle coordinates to the estimated ice thickness levels

In contemporary single particle data analysis often only a fraction of initial particles contributes to the final high-resolution 3D maps. It takes weeks, and often months to define such a group of “best behaved” particles retaining high-resolution information. Authors in [183], have demonstrated that the particles in the thinnest ice should be prioritised for high-resolution work. In our IceBreaker paper[184], we have demonstrated that relationships between the ice thickness and the quality of the resulting 3D map are more intricate and are, likely, sample dependent. In reality, modern automated data acquisition methods prioritise throughput over pin-point precision, thus methods for post-acquisition data evaluation and discrimination are required to save on processing time and resources. The second IceBreaker script outlined below, can be used to associate the particles' coordinates with the local ice estimation. The required input is the set of grouped micrographs in a STAR file format obtained with the previous script and the set of particles with coordinates. In the previous implementation of this software, only the value of intensity at the reported coordinates was taken. This might result in incorrect classification if the particles fall on the edge between two clustered regions. To alleviate this problem, a new mapping system was developed based on associating the particles with the average of the values at five coordinates: the centre of the particle, and four corners of the box containing the particle. This requires an additional input argument for this task to define the size of the box but results in more robust mapping. Under the assumption that the ice gradient is continuous across the micrograph, otherwise, the hole would be damaged, the large differences between the adjacent groups are not expected within the same micrograph and misplacing the particle in the neighbouring group should not affect the later reconstruction. If the particle is located too close to the contaminant or the edge it might be annotated as a bad pick. The output file will match the input STAR file with an additional column containing the estimated ice value for each particle. This approach allows for keeping the processing pipeline intact and accessing the estimated ice values as a parameter for particle subset selection. Additionally, the new STAR file can be easily used as input to other job types. Due to the limitations in the previous version of Relion and the lack of possibility to introduce new variable names the pre-defined ‘_rlnHelicalTubeID’ variable was reused, as the IceBreaker in the current state is not designed to work on datasets containing helical tubes (a helical tube would normally represent an extended continuous object which could skew the IceBreaker statistics, and full implications of this were not yet assessed). It was successfully used within the Relion

pipeline in the subsequent jobs where particle subset selection was based on this parameter. A more convenient and more standardised solution would be to introduce a custom variable named ‘_ibEstimatedIce’, which would contain information about the estimated ice thickness values for each particle in the STAR files. One of the limitations to consider is the fact that the STAR files are appended with additional columns with data as the processing pipeline proceeds, the introduction of a variable outside of the regular Relion scope can negatively interact with already existing automated tools and would require careful implementation.

The subsets of particles with the estimated ice thickness parameter can be useful as soon as the first particle-picking job is done. At the early stage of the processing pipeline, filtering particles by estimated values can help to remove false picks. Later in the processing stage, another approach may include 2D or 3D classification based on the ice thickness parameters to identify possible preferred orientation problems, or even for generation of ab-initio models from different ice groups to see if the local ice conditions affect protein conformation or angular distribution.

3.3.7 Local contrast improvement based on the histogram equalization

The processing steps to obtain contrast equalized image are:

1. The required input for this type of job can be defined as a path to a folder containing motion-corrected mrc files. Within it, a subfolder called ‘/equalized’ is created to store the output files.
2. The image is low-pass filtered to 20Å to remove the high-frequency noise and reveal the ice gradient. This is done in Fourier space by multiplication of the transform of the image by a circular function with soft edges to avoid Fourier ripples. This is a commonly used step for automated particle picking. The high-frequency noise is removed to reveal low-frequency details such as the particles this would also give more weight to the ice gradient background since it is a low-frequency feature. The low-passed filter micrograph is used as an input for the contrast equalisation, which is performed locally in the regions determined after the clustering.
3. Based on the specified number of patches the image is average-pooled to reduce the data size. The patches, representing now the average value in a specific region, can act as super-pixels used for clustering. Additionally, at this stage, the input images are also downsized 20 times in x and y direction.
4. The segmentation step is the same as in the grouping mode.

5. The segmented image is upscaled to the original size to retrieve the data.
6. The segments defined in the previous step can be iterated over. They are used as a local mask for the low-pass filtered image. In this case, histogram equalisation is performed into each region defined by a local mask. This step is more computationally exhaustive compared to local averaging in the grouping mode, as it is performed on a full-scale micrograph.
7. The output mrc file is reconstructed by combining the segments into one image. The files are named the same as the corresponding original images with the suffix ‘_equalized’.

It is important to emphasise that if the contrast-improved images are used for the particle-picking step (the only application we have tested so far), the particles for subsequent high-resolution data analysis should be extracted from the original images. Because contrast-equalised images are scaled back to an original micrograph size, the coordinates can be directly used for particle extraction from the untreated micrographs. Figure 3.10. shows the idea of applying the local mask from segmentation (blue) to the corresponding area of the micrograph.

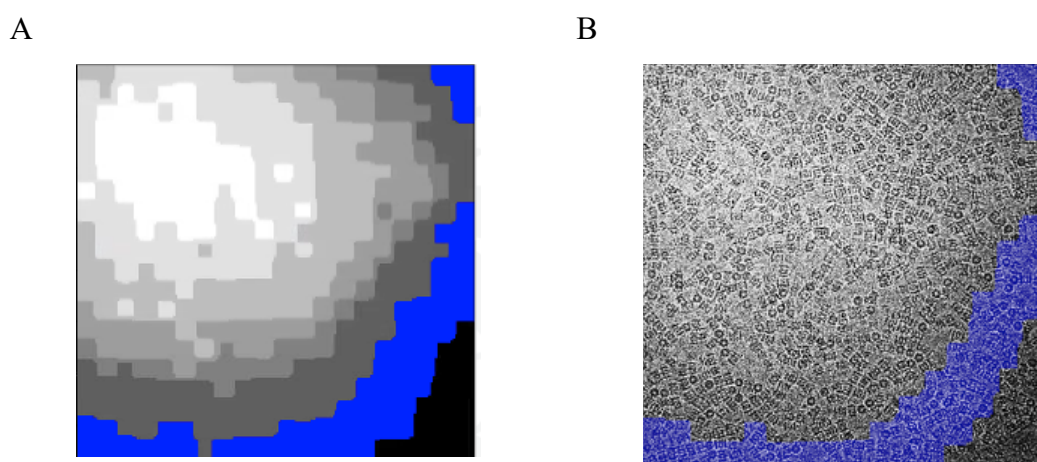


Figure 3.10 Segment from clustered image used as a local mask A) clustered image with one of the segments highlighted (blue), B) segment applied to the lowpass filtered image as a local mask

3.3.8 Adaptive histogram equalisation for cryo-EM micrographs

Histogram equalisation is commonly used in image processing applications to enhance the contrast in the image. It is achieved by spreading out the values from the most frequent intensities range to cover the whole intensity range of the image. To redistribute the pixel intensity values, the cumulative distribution function (CDF) is used. It is based on the original histogram of an image and represents the distribution of probability of pixels with

intensity lower than a given grey level. The cumulative distribution function (CDF) of the greyscales in the image can be described with Eq.3.5 where p is a pixel of the image and i is the greyscale level in range $0 \leq i \leq L$, where L is the total levels of grayscale. The resulting CDF is continuous and increasing, effectively resulting in an accumulated histogram with a bin for each greyscale.

$$cdf_x(i) = \sum_{j=0}^i p_x(x=j) \quad \text{Eq.3.5}$$

According to Inverse distribution function properties, there is a constant K which allows obtaining an image y with a new flat distribution of pixel intensities (Eq. 3.6)

$$cdf_y(i) = (i+1)K \text{ for } 0 \leq i \leq L \quad \text{Eq.3.6}$$

Which can also be described as a transform $y = T(k) = cdf_x(k)$

The linearisation of the CDF leads also to histogram equalisation, as a linear CDF would mean that the pixels with different intensities contribute equally to the new image.

Adaptive histogram equalisation is the approach which aims to improve contrast locally in different parts of the image. Local masks obtained with the IceBreaker group together the areas of the micrograph with similar background features, which makes them good candidates to define the areas for adaptive processing. The resulting image after the adaptive equalisation will heavily depend on the number of clusters used. Fig 3.11.A shows a low-pass filtered micrograph and the distribution of pixel intensities, panel B shows the image after segmentation with 40×40 patches and 8 clusters, and panel C presents a micrograph after adaptive contrast equalisation with the distribution of the particle intensities (blue) overlaid with the original distribution (yellow). The skewness of the original is reduced, resulting in a more even intensity across the image.

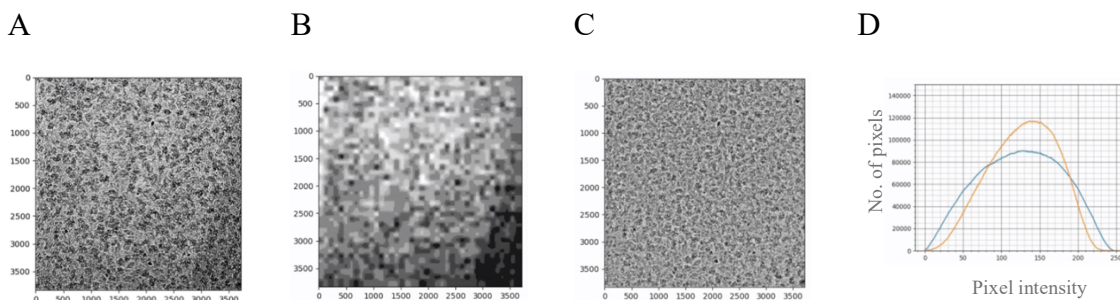


Figure 3.11 Effect of contrast equalisation, A) original lowpass filtered, B) image clustered with 40×40 patches and 8 clusters, C) image after local contrast equalisation, D) distributions of intensities in the images; new intensity distribution (blue) overlapped with original (yellow)

To check how the selected number of clusters affects the final image, an example with a coarser gradient was selected and presented in Fig.3.12.A along with pixel intensity distribution revealing a large number of pixels both in dark and bright areas. In this case, the 8 clusters are not enough to fully recover the data from the micrograph (Fig.3.12.B), moreover performing the contrast equalisation in a region with too high a difference in the intensities can introduce artefacts on the edges of the masks. To reduce this effect a larger number of 32 clusters was used. In this case (Fig.3.12.C) contrast was improved successfully, additionally isolating aggregates in the field of view as separate entities. As the results improved, the required processing time increased from 1.94s for 8 clusters to 9.35s for 32 clusters.

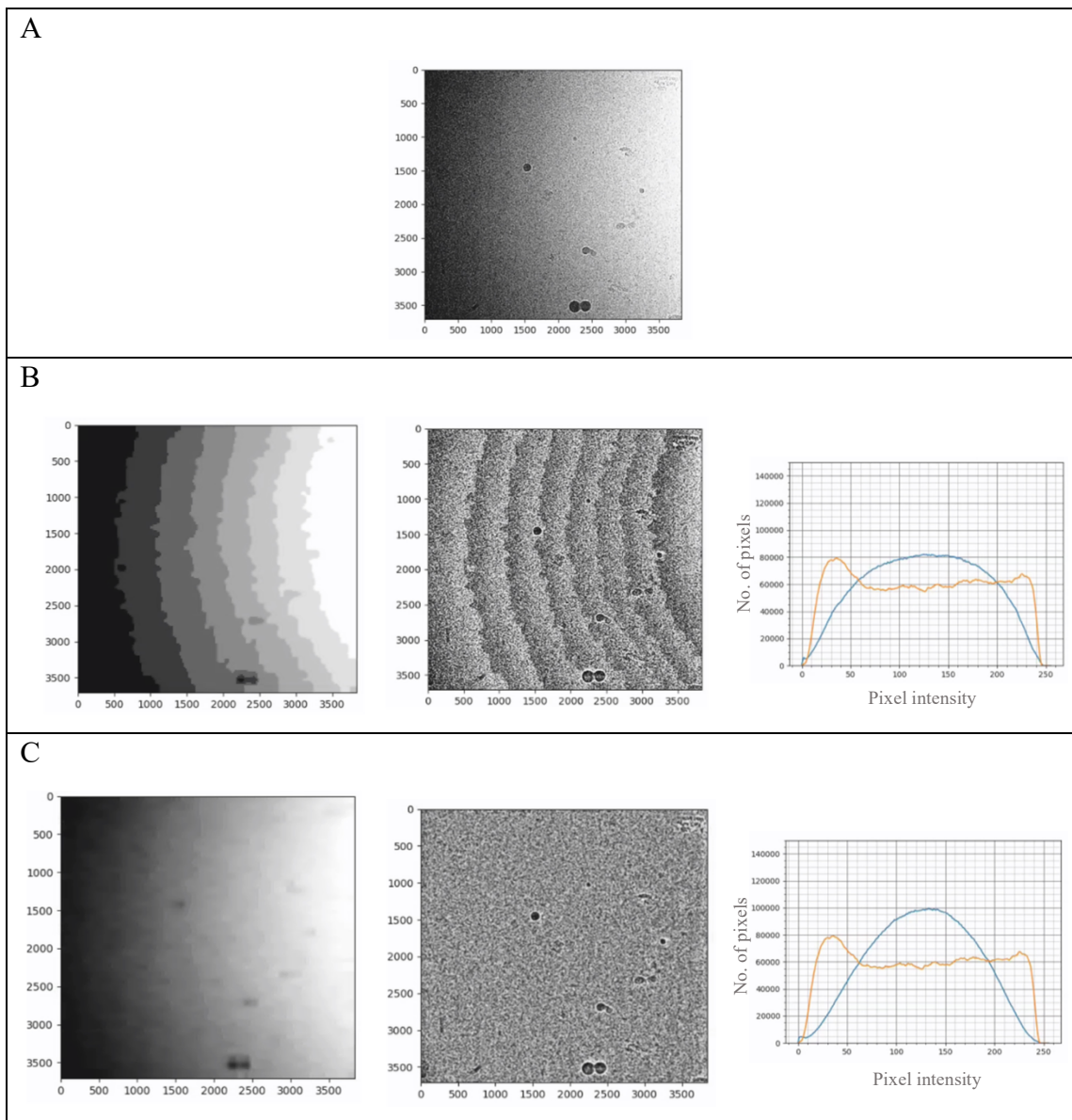


Figure 3.12 Example of image with coarse ice gradient, A) lowpass filtered original image with pixel intensity distribution, B) image segmented with 8 clusters, result from contrast equalisation that introduces artifacts on the edges of the clusters areas, C) image segmented with 32 clusters; higher number of clusters allowed to eliminate artifacts and remove the ice gradient; pixel intensity distribution after equalisation (blue) and from the original image (yellow)

The proper choice of the number of clusters used for adaptive contrast equalisation proved to be affected by the initial image conditions. It is yet to be analysed how to optimally set this parameter to balance the satisfactory results with the required computational time. One of the possibilities could be to analyse the histogram of the original low-passed image to identify what level of correction is required and based on this to choose the appropriate number of clusters. Additionally, the contrast enhancement used in the presented work is not a linear operation and its main objective is to improve the contrast between the particles and the background in order to improve micrograph interpretability and particle picking. Once the particle coordinates are obtained, they should be re-extracted from the original micrographs for high-resolution refinement.

Another approach for contrast improvement on the cryo-EM micrographs includes band-pass filtration. It removes the high-frequency noise as well as the low-frequency noise that can be connected to the background intensity changes caused by non-uniform ice thickness. The optimal band-pass filter parameters would be different for micrographs with different defocus values. To demonstrate this, two micrographs were selected, one with a relatively high underfocus of 3.2 microns, which is typically at the end of defocus range, and another one closer to focus at 1.3 microns underfocus. From each micrograph, particles were picked with automatically using Laplacian-of-Gaussian. At each defocus level raw micrograph was tested, as well as micrograph after filtration with different bandpass filter setup, and after contrast enhancement with the IceBreaker software. Table 3.3 shows how the number of picks from two micrographs with high and low defocus can change based on the bandpass filter setup and how they compare to the original micrograph and after IceBreaker flattening

Table 3.3 Number of particles picked from micrographs with different defocus with and without bandpass filter.

Defocus [microns]	Bandpass filter setup	No. of particles picked
3.22	No filter (raw micrograph)	882
	20-200 Å	1068
	20-250 Å	1006
	20-400 Å	918
	20-500 Å	913
	Enhanced with IceBreaker	1014
1.35	No filter (raw micrograph)	900
	20-200 Å	1059
	20-250 Å	1009
	20-400 Å	966
	20-500 Å	954
	Enhanced with IceBreaker	996

For currently used particle pickers, the contrast level between the particle and background is one of the main factors. Therefore, more particles can be picked from thinner areas with high contrast than from thicker ice. Modifying the threshold for picking to include thicker

areas often leads to many false positives being selected. IceBreaker solves this problem by equalising contrast on the whole micrograph. This means that a single threshold can be used for efficient particle picking across the whole range of ice gradients. Figure 3.13 shows a comparison of picks from a fragment of micrograph at high-defocus where red circles indicate ‘particles’ found after 20-200 Å band-pass filtration and blue circles ‘particles’ selected from IceBreaker treated micrographs. In both cases particles are evenly picked from all areas, despite the lower left corner was significantly darker. Close analysis of picks reveals that in both cases false positives containing ice contamination are present. However, only IceBreaker algorithm would allow ice distribution analysis for their efficient removal. The same algorithm can be employed for the removal of the particles picked from the carbon film edges of the hole since they normally have a very distant optical density profile. After applying the band-pass filtration, even if the background is more uniform, the information about the local ice conditions is lost and cannot contribute to further processing stages.

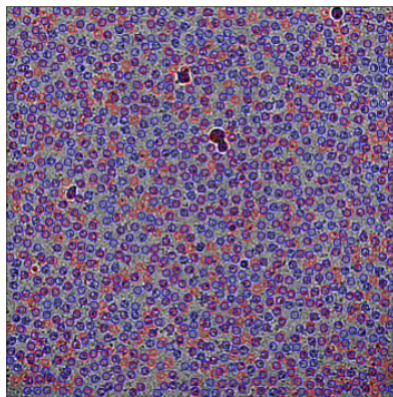


Figure 3.13 Comparison of picks from a 3.2 micron underfocus micrograph treated with band pass filter (red) and after contrast improvement with IceBreaker (blue).

3.4 Conclusions and future works

The relationship between the local pixel intensity calculated with the IceBreaker and the actual ice thickness measured on the single particle analysis data with the energy filter method was established. A description of a framework which can be used for calibration of the values is provided. It would also be useful to relate the ice thickness estimation with measurement with different methods based on the stage tilt like tomography reconstruction or hole milling. Initial results showed that the local pixel intensity can be used to represent the overall shape of the sample similar to that obtained from tomography.

The exact estimation of the ice thickness would be helpful, but even in cases where the calibration is not possible, sorting the intensities between maximum and minimum values and creating subsets can provide information about the behaviour of the sample and a particular cryo-EM specimen characteristic.

IceBreaker was developed and evaluated using the historical dataset available in the public archives. The possibilities of optimising data collection parameters and microscope setup to get the most out of the ice distribution analysis would require a considerable amount of time and ideally, access to the microscope to thoroughly test different scenarios and draw conclusions. Results presented in this part, related to dose-weighting, energy filter, defocus, and presence of crystalline ice contaminations, are limited to the selection of deposited datasets and time restrictions.

The identified issues with not optimal contrast equalisation performance were pointed out in this work and addressed by increasing the number of clusters in presence of the coarse ice gradient to improve the results. The presented example of band-pass filtration indicated inconsistent performance of this approach in different defocus regions. Extensive trials of different contrast improvement algorithms, probably with different input parameters as well, would require a substantial amount of time and would not guarantee finding solution to cover each case. The main strength of presented software is that, even after applying filters or other operations to the micrographs or particles, it still can carry the information about local ice conditions which can be used at later stages of processing.

The distribution of local pixel intensities can be also used to evaluate the quality of micrographs rather than single particles, which can lead to better data management and selection of the best images for processing. Analysis of the image at a lower magnification level can help to select the areas for data collection. Figure 3.14. shows the results of processing at lower magnification. The intensities in the holes can indicate the local conditions and shape of the ice layer. Panels A and B show images from EMPIAR-10143 where the holey carbon support was used. Panel C shows an image from EMPIAR-10138 with holey carbon Spotiton support. After segmentation, the hole areas corresponding to high and low intensity can be colour-coded, similar to the ‘fire-n-ice’ colour scheme used in Relion. The image magnification in Figure 3.14. is consistent with the one used for medium-magnification maps in tomography and could be utilised for single-particle tomography area targeting. It remains to be explored if GridSquare magnification used in pure single particle data collection routines has sufficient detail for the accurate direct ice gradient profiling with IceBreaker.

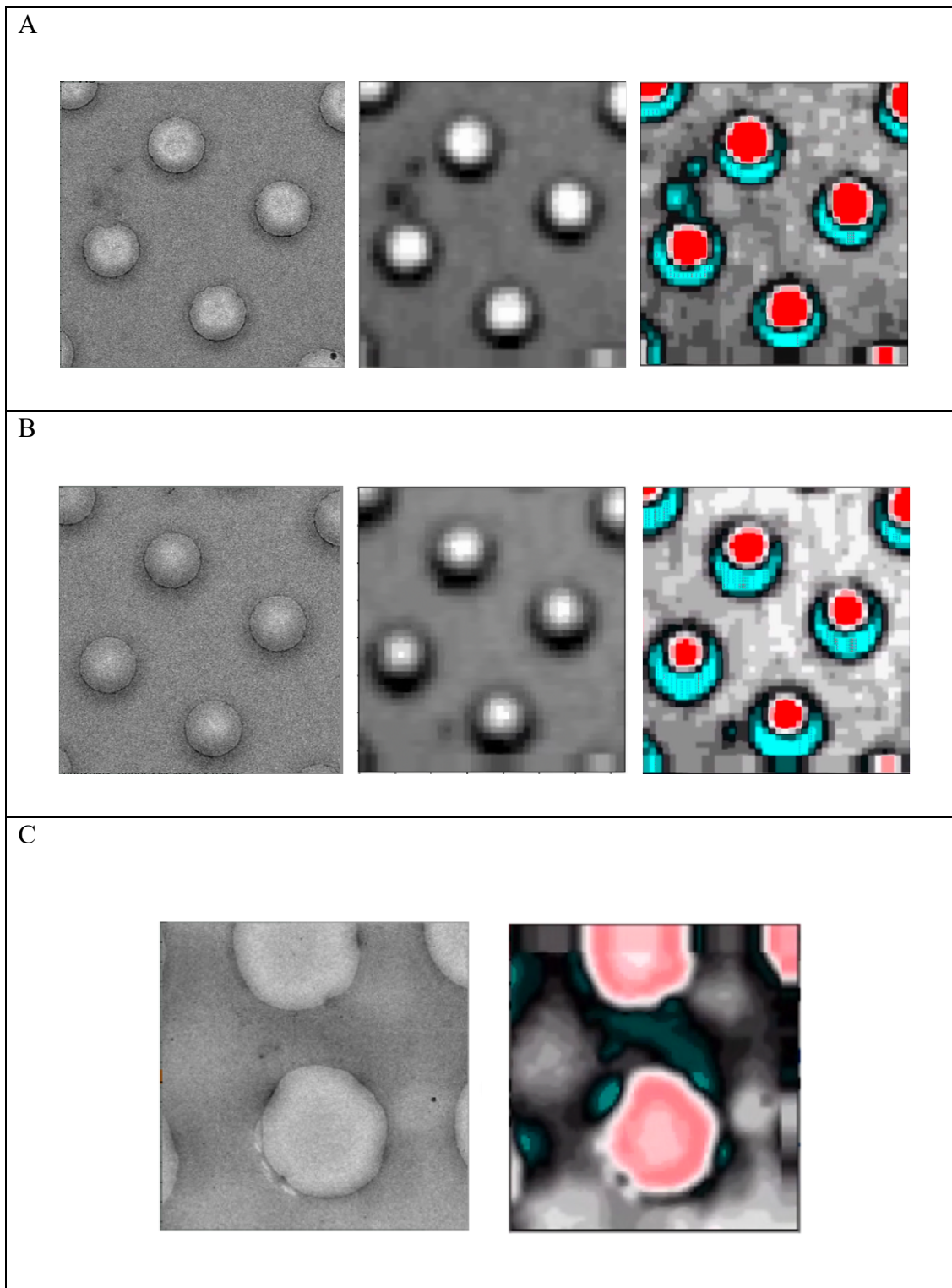


Figure 3.14 Segmentation based on the intensity for lower magnification level, images at the medium magnification used as the input for IceBreaker reveal changes in the intensity inside the hole, in the future this can help to annotate the grid holes and select regions of the grid for data collection; images coloured with Relion ‘fire-n-ice’ scheme where red corresponds to high intensities (thinned ice) and blue to lower (thick ice)

As the local contrast equalization shows promising results in the removal of the ice gradient from the background of the micrographs, the proper choice of parameters

requires further analysis to balance the benefits of the improved interpretability of the micrographs and the required processing time.

The improved, flat representation of the micrograph can be useful for maximising the particle picking with methods based on edge detection, but the fact that the resulting images are processed with non-linear transformation could affect the performance of particle pickers based on the correlation methods. Methods based on machine learning methods could be retrained on this new type of image to optimise the picking.

Finally, the definition of local masks which group together parts of the micrographs with similar background features open the doors for a variety of local processing applications.

4 Implementation of the atomic model validation tool based on the False Discovery Rate approach

This chapter presents updates to the atomic model validation tool based on the False Discovery Rate approach. Metrics for evaluation of the classifier performance are presented based on the root mean square deviation (RMSD) of the atomic positions in superimposed target and reference atomic models. The parameters such as input map sharpening, number of particles used for map refinement and local resolution are taken into consideration in evaluating the approach as they affect the overall score performance. The first version of the presented software was published in the *Frontiers of Molecular Biosciences* as “Cryo-EM Map–Based Model Validation Using the False Discovery Rate Approach”, added as Appendix B. This tool is implemented in the CCPEM software suite.

4.1 Introduction

Atomic models are crucial for interpreting biological structures at near-atomic resolutions and help to understand their functions, mechanisms and interactions. In recent years, a number of tools have been developed for building atomic models from cryo-EM density maps [185], [186], [187], [188]. This allowed users to quickly and routinely obtain highly complete structures with a recent boost of methods adapted from the field of Artificial Intelligence. Unfortunately, even robust atomic model building and refinement workflows can introduce errors, especially when building from maps with resolution ranges 3-5 Å. The local changes in cryo-EM map resolution can lead to an inaccurately traced backbone or incorrect secondary structure elements with misplaced α -helices and β -sheets [189]. Other common issues with auto-generated models are over- and underfitting of the atomic coordinates to the density or reference bias introduced if the model is built or extended from a reference model, all leading to the output model not representing the actual features of biological structure.

Several methods were developed to validate the correctness of atomic models based on different criteria to ensure the accuracy and reliability of the model. Completeness and continuity of the model are used to assess if all residues and parts of the protein are correctly connected to represent the protein sequence without discontinuities and register shift errors [190].

The geometry of the model can also be assessed without referencing it to the map, by evaluating the model according to stereochemical restraints applied to the whole model[191] or its specific elements such as water molecules or targeted ligands.

Ramachandran plots assess high resolution features such as backbone torsions while CaBLAM[174] evaluates lower resolution features involving backbone C-alpha geometry.

Another criterion of model validation is density fit metrics, where the model is validated against the map calculated from experimental data[157], [175], [176], [179], [192]. The major issue with these methods is that they are prone to errors in validation where there is a variance in local resolutions across the map.

During the deposition of the maps and models to EMDR, the depositors define a numerical threshold for the rendering of the cryo-EM map in visualisation tools, such as COOT[193] or UCSF Chimera[162]. This parameter is important for data interpretation as the cryo-EM maps are typically not standardised or normalised between any fixed range. This value is also a parameter used for Atomic Inclusion score validation. This method would threshold the map at a given value and calculate the fraction of how many atoms are inside the density represented at that threshold, resulting in an overall score between 0 and 1 for the model. Additionally, each residue is colour-coded green if it is entirely inside the map or red if it is entirely outside, based on the position of all the atoms in that residue. As a recommendation, the threshold should be chosen to represent the volume corresponding to the molecular weight of the specimen of interest, and the software reports the defined volume, but still, the user input is used by default[194].

The output of most validation tools is similar and typically includes a list of potentially problematic residues or atoms, which can later be inspected in visualisation tools like UCSF Chimera or re-refined manually in the interactive COOT interface. As an outcome of the EMDR challenge, it is recommended to use multiple scoring parameters, ideally based on different criteria, for a comprehensive model assessment.

4.1.1 Raw and sharpened cryo-EM maps

Cryo-EM maps are typically reconstructed from two independent datasets (half-maps). Half maps are two separate reconstructions of the same structure, refined from the same data collection[195]. Half-maps are calculated from two distinctive sets of particles that do not overlap to improve the resolution and signal-to-noise ratio of the final map. This approach is used to avoid false correlations when the Fourier Shell Correlation is calculated[196]. After all the data has been processed and combined, and can be used for

post-processing tasks, like sharpening or local refinement, to improve the interpretability of the final map. A common practice for map sharpening is applying the global B-factor to the map. The B-factor value is estimated based on the Guinier plot analysis[122], which describes how the power spectrum of the cryo-EM map decays with resolution, resulting in the loss of high-frequency information on the map. The rotationally averaged power spectrum of the density is calculated in the resolution shells between the maximum resolution of the map reported from the value at 0.143 FSC curve, typically up to 10 Å, as this is the resolution range, where Wilson statistics can be applied, as the scattering amplitudes of randomly positioned atoms decrease with resolution in a roughly linear relation. The amplitude decay plot for the experimental data is compared with the plot from the reference structure to see how much compensation is required. The B-factor is calculated as a difference in slope parameters between the linear fit of data from the reference model and experimental data. The correctly chosen B-factor should boost the signal in the high-frequency region, but it can also amplify high-frequency noise. B-factor can be applied to the map in the reciprocal space in the form of multiplication of the Fourier transform of the map and the transform of function $e^{-(B\text{-factor}/4d^2)}$, where d is the map resolution[197]. Usually, maps with a resolution better than 4 Å and an estimated global B-factor lower than 150 would be suitable for model building. Given that the B-factor is applied globally, local resolution changes can affect the final model's quality. Therefore, both raw and sharpened maps should be checked to avoid oversharpening. Another approach to avoid oversharpening is using methods for local sharpening, such as LocScale[165], [198].

4.1.2 False Discovery Rate approach

An atomic model validation tool is presented based on the framework introduced for thresholding of cryo-EM densities with False Discovery Rate (FDR)[199]. The 3D confidence map, which associates each voxel within the volume of interest as carrying a molecular signal or noise, is generated by multiple hypothesis testing and an FDR control framework.

In hypothesis testing, the null hypothesis (H_0) is the initial statement defining some relationship between the variables, while the alternative hypothesis (H_a) usually suggests the opposite. A p-value is a statistical measure that represents the probability of observed results if the null hypothesis is true. It quantifies how likely it is that differences between groups are random, with values ranging between 0 and 1. A p-value lower than the set significance level (typically 0.05) suggests that the observed difference is unlikely to be

caused by chance, and as a result, the null hypothesis is rejected, indicating the statistical significance of the observation. The p-value analysis can be used to reject the null hypothesis, but at the same time, it does not automatically indicate the truth of the alternative hypothesis. The null hypothesis used for the Confidence Map calculation is that the intensity at a given voxel of 3D cryo-EM map is greater than the observed background. The requirement to test each voxel results in a multiple testing problem. Each test results in its own p-value, which is adjusted to meet the False Discovery Rate framework requirements. The FDR describes the expected proportion of false rejections out of all rejections. The p-values are adjusted to control the proportion of false discoveries to account for multiple hypotheses comparison [200]. The resulting Confidence Map is the representation of adjusted p-values of each voxel. For the interpretability of the output, 1-FDR is used, which means that the map thresholded with a value of 0.99 represents all voxels with a maximum p-value of 1% [201].

4.2 Methods

4.2.1 Processing stages description:

1. The required inputs to the procedure, as implemented in the CCPEM software suite, include a cryo-EM electron density map and the atomic model in cif/mmcif or pdb format. Users can define the size of the noise cube for p-value calculation. By default, the cube that represents 5% of the full map volume is applied. Users also have the option to display the cube size and position relative to the map to ensure that the cube contains the background noise and does not overlap the protein density. The standalone Python script for model validation requires as input an already calculated Confidence Map and the model. The Confidence Map can be calculated from the task with the same name in the CCPEM suite or from the source script ([Maximilian Beckers / FDRthresholding · GitLab](#)).
2. The coordinates of the atoms in the model are mapped to the density grid from the Confidence Map. The current implementation allows users to choose the size of the volume of interest for the atoms to be validated. Available options include 1 Å radius for every atom, Van der Waals radius for each atom fetched from the GEMMI library [202], and the volume calculated from the Sigma threshold in a similar manner as it is implemented in volume visualisation tools according to equation $r = 1.5 \times 0.225 \times res$, where r is the radius for validation and res is the map resolution. By default, the radius of 1 Å is applied, but users can choose other

modes with input parameters *-mode vdw* for Van der Waals radius, the Sigma Threshold mode also requires to specify the map resolution with *-mode st -res map_resolution*. The sigma factor value, with the resolution, is used in the molecular visualisation tool UCSF Chimera to determine the width of Gaussian distribution to describe each atom. 0.225 is a default used value which corresponds to the Fourier Transform of the distribution fall to 1/e of it's max value at the wavenumber 1/resolution [203]. We use $1.5 \times \text{sigma}$ as the radius, which represents ~82% of the distribution.

3. Based on the selected radius, each atom is associated with the average value from the Confidence Map. By default, the FDR score is calculated as the average of the values at the coordinates of the backbone atoms: C-alpha, C and N. This approach is used to find the misplaced residues not supported by the map density. The carbonyl oxygen is often not supported by the map density at resolutions lower than 3 Å. For the validation of the position of the nucleic acids, the coordinates of the C1', C2', C3', C4', C5', O3', O4', O5', and P atoms are used. In the case of water molecules and ligands, all atoms are used. Additionally, users can use the option for minimal validation based only on C-alpha positions for residues and C1' for nucleic acids, which can be useful for *ab initio* models built from low-resolution cryo-EM maps. In the recent implementation, we also added an option to run the validation for all atoms in the model, as it can be useful for further investigation of the positioning of specific atoms rather than whole residues.
4. As the output, a CSV file with FDR score for each residue is provided, with the chain and residue name and ID. The output file has a suffix indicating which radius mode was used for validation. The output from the task implemented in CCPEM would also include the attribute file, which allows users to colour the residues according to their FDR scores in UCSF Chimera and easily identify areas of the model for further inspection.

4.2.2 Metrics for the evaluation of classifier performance

The performance of the classifier was evaluated against the root mean square distance (RMSD) calculated for the backbone atoms of the target model and the reference model. The true positive (TP) is a residue predicted to be correct and is actually correct (in this case, placed inside the density), the true negative (TN) is a residue predicted to be placed incorrectly and actually placed incorrectly (outside the density), false positives (FP) are residues reported to be correct, but actually are incorrect and false negatives (FN) that are

classified as incorrect, but are actually correct. Equations 4.1.-6. show metrics used to check the quality of the classifier.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad \text{Eq.4.1}$$

$$Error = \frac{FP+FN}{TP+TN+FP+FN} = 1 - Accuracy, \quad \text{Eq.4.2}$$

$$Sensitivity = \frac{TP}{TP+FN}, \quad \text{Eq.4.3}$$

$$Specificity = \frac{TN}{TN+FP}, \quad \text{Eq.4.4}$$

$$False\ negative\ rate = \frac{FN}{TP+FN} = 1 - Sensitivity, \quad \text{Eq.4.5}$$

$$False\ positive\ rate = \frac{FP}{TN+FP} = 1 - Specificity, \quad \text{Eq.4.6}$$

Furthermore, the Receiver Operating Characteristic (ROC) was plotted to select the optimal cut-off threshold to identify the incorrectly placed residues. The area under the curve (AUC) serves as a global measure of the ability of the classifier to discriminate between true and false positives. An AUC of 0.5 represents a test with no discriminating ability, no better than a random guess, and an AUC of 1.0 means perfect classification.

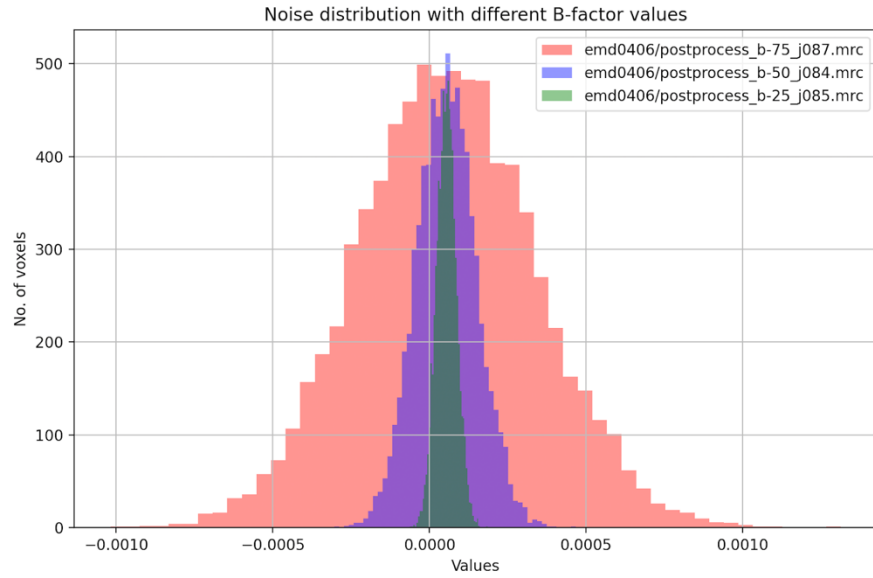
4.3 Results

4.3.1 FDR score and cryo-EM map sharpening

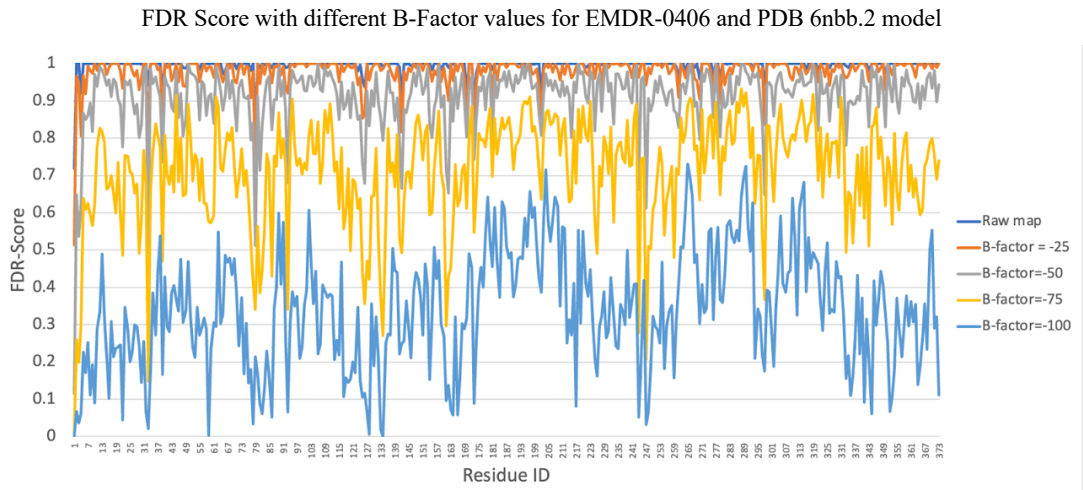
The FDR score is calculated based on the background noise distribution. The residue scores will change based on what kind of sharpening was applied to the input map. To demonstrate this, I used PDB model 6nbb.2 with EMDB-0406 Alcohol Dehydrogenase map, which was deposited with raw map, half-maps and the mask the depositors used for sharpening. The map was sharpened with the Relion 4.1 post-processing task. The auto B-factor estimation returned a value of -50. Additionally, B-factors of -25 \AA^2 , -75 \AA^2 and -100 \AA^2 were applied to the raw map, also using the Relion 4.1 framework. To further investigate the results, a cube of the background noise with the size of 20x20x20 voxels was selected, and the noise distribution from the maps sharpened with B-factors of -25

\AA^2 , -50\AA^2 and -75\AA^2 in this volume of interest is presented in Figure 4.1.A. As expected, the sharpening changes the background noise distribution, which could affect the p-values of the distributions and also the FDR score. The sharpening levels will also change the distribution of values in the molecular volume. As presented in Figure 4.1.B the FDR score gradually degrades with increased sharpening. However, it can be seen that many of the peaks are aligned. The correlation coefficient between the raw map and the one sharpened with the B-factor of $-50[122]$ was calculated and equals 0.71. To align the FDR score plots, the third quartile of the distribution was brought to 0.8. This way, a global FDR score threshold can be applied to identify the outliers in different models (Fig.4.1.C). The initial approach was to use the Z-score metric to find the outliers. The Z-score was calculated for each residue of the model by subtracting the mean value from the residue value and dividing it by the standard deviation. It is a commonly used method to identify outliers in a dataset, ideally when the distribution is close to normal. This approach was helpful when scoring the high-quality models with only a few outliers if the Z-score was lower than -2, meaning that the residue score was at least two standard deviations below mean score. It proved to be less robust when the mean is lower and standard deviation spread is larger for lower-quality models. Moreover, the different sharpening levels affect the overall distribution of the scores per model, making it more challenging to assess different models or models built from maps differently sharpened. Figure 4.1.D shows the boxplot representation of the distribution of the FDR scores obtained from high-quality PDB model 6nbb.2 built from EMDB-0406 Alcohol Dehydrogenase map. Without any sharpening, the FDR scores from the raw map density highlight only the most serious issues in the model. The median value and overall score distribution drastically change as the additional sharpening is introduced. The evaluation of the model performance in this section of the thesis was performed based on aligning the distributions on the Q3 value. The Q3 corresponds to the top 25% of the scores which is more robust compared to Q2, even for bad models.

A



B



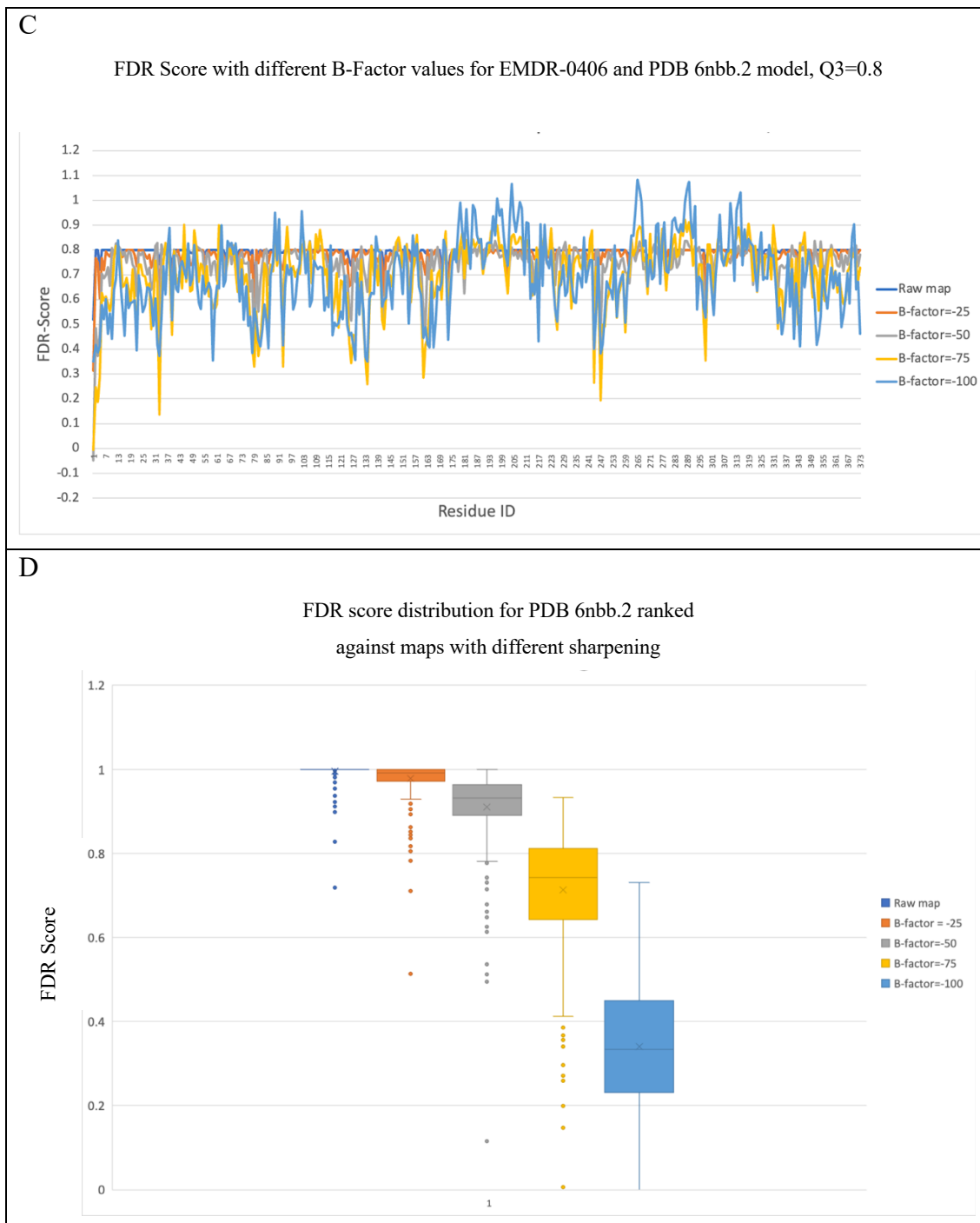


Figure 4.1 FDR score from the sharpened maps, A) noised distribution in the background box of the maps after sharpening with different B-factors, B) FDR-score degrades as the applied B-factor increases, C) moving the third quartile (Q3) of the distribution to 0.8 shows that low scoring parts of the model align, D) FDR score distribution for PDB 6nbb.2 model ranked against maps with different sharpening

4.3.2 FDR score evaluation

As the reference to evaluate the performance of the classification, the best scoring models were selected from EMDR model challenge for target maps of apoferritin EMD-20026 at 1.9 Å resolution, alcohol dehydrogenase EMD-0406 2.9 Å and T20 proteasome EMD-6287 2.8 Å. A total number of 36 test models were chosen, and from each one monomer was selected, resulting in a total number of 8861 residues. The FDR-score is evaluated based on the RMSD for the distances between the backbone atoms C-alpha, C and N atoms of the reference and test models expressed in Angstroms, calculated according to equation 4.6, where n is the number of atoms, and v,w are atomic coordinates:

$$RMSD(v, w) = \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_i - w_i\|^2}, \quad \text{Eq.4.6}$$

The conditions, with the assumption that all atoms are positioned correctly in the reference model, were defined as follows: True Positive when the FDR-score is above certain threshold, and the RMSD is lower than the cut-off distance, True Negative when the FDR-score is lower, and the RMSD is higher, False Positive when the FDR-score is high, and the RMSD is also high, and finally False Negatives, when FDR-score is low, and the RMSD is low. Table 4.1 presents the metrics values for FDR-score thresholded at 0.65 for the 1 Å RMSD evaluation. 1 Å radius used for validation results in higher AUC for each RMSD value tested. As it covers a smaller volume of interest, the FDR score can be calculated more precisely without overlapping with the background noise. Table 4.2. summarises the confusion matrix at the selected threshold with 91% of true positive predictions. Figure 4.2. shows ROC curves for two different validation radius values. The ‘1A’ mode performs better than Van der Waals atomic radius with every tested RMSD value. The smaller radius is more precise and does not introduce additional volume which in some cases can represent the noise.

Table 4.1 Performance metrics for FDR score

Accuracy	Error	Sensitivity	Specificity
0.9480	0.0520	0.9634	0.6496

Table 4.2 Confusion matrix for the FDR score with 0.65 threshold and 1 Å RMSD

		Actual values	
		Positive	Negative
Predicted values	Positive	0.91	0.02
	Negative	0.04	0.03

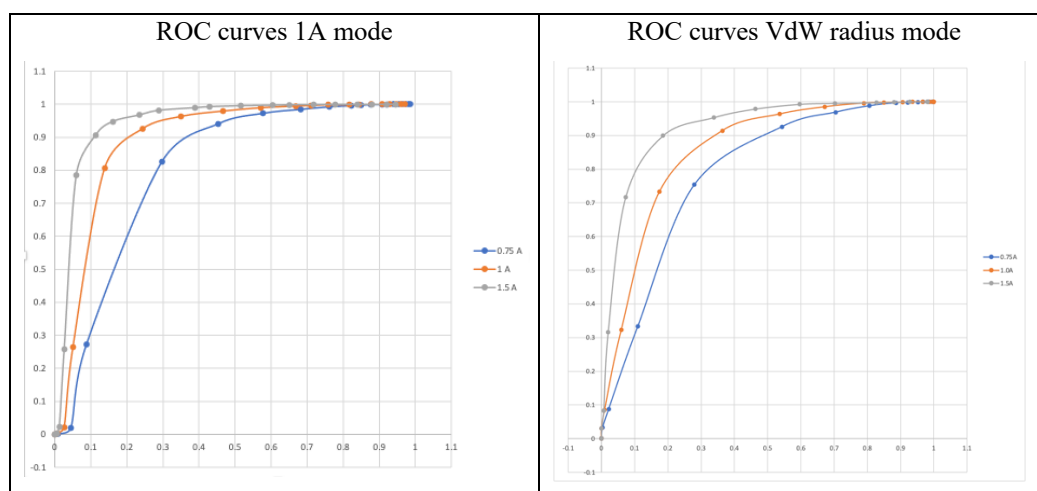


Figure 4.2 Comparison of the ROC curves for FDR score calculated with 1Å and Van der Waals radius modes for different RMSD distances from the reference model

The proposed method scores the atomic model based on how well it aligns with the molecular volume. And this is irrespective of whether the correct residue is in that position. Such limitations of the FDR-score were observed during the RMSD analysis. In cases of register shifts where the atoms are shifted along the residues, they can get high FDR scores but match incorrect positions in the map, for example, atoms shifted along the alpha helix or moved from their true position but still supported by the density. The other case is where some or all of the mistraced residues are still within the molecular volume. In this case, these residues will still hold a high FDR score. An example of this is shown in Figure 4.3., where the model of alcohol dehydrogenase 60_2 shows areas where the overall quality of the model is low, but still some of the residues score high. Nonetheless, low-scoring residues, highlighted in the output report, should bring user

attention to this part of the model for inspection. Also, the automated pruning implemented in the programme would remove not only each low-scoring residue but also one before and one after it, which should overall clear this area of the model. The pruning option which removes a low scoring residue (and additionally one residue before and after in the backbone to make more space for rebuilding) is available and implemented in the CPP-EM suit and will be added in Doppio. As a result it writes out the PDB file with pruned model and a text file with IDs and names of removed residues.

Overall, we recommend the use of auto-sharpened maps to generate FDR scores. The use of raw maps tends to highlight only the serious mis-traces in the model. Oversharpened maps are likely to generate more false positives (errors in the model) with this approach. To compare the performance with respect to different map sharpening levels, we aligned the distributions on Q3 value. Although it is more robust than baselining on Q2, especially at higher sharpening levels, it is not optimal and a better normalisation based on sharpening levels is required to make the score less sensitive to differences in sharpening levels. In the new version of the CCPEM suite (Doppio) a local Z-score is calculated for each residue within 12 Å radius. This minimises the effect of variability in local resolution and sharpening levels on the scores. Errors are identified based on the local distribution of scores around each residue.

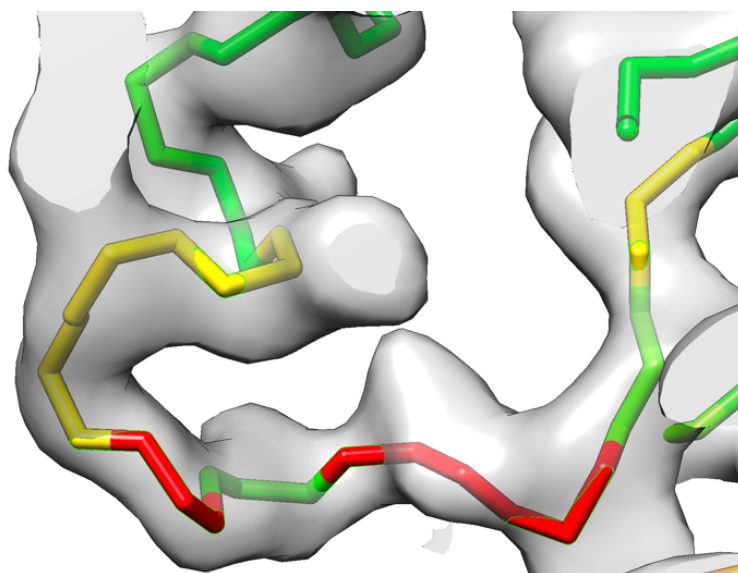


Figure 4.3 FDR score can report high scores for some residues even if the model is incorrectly built in that region, which can result in increased numbers of False Positives in the classification

4.3.3 FDR score and Local Resolution

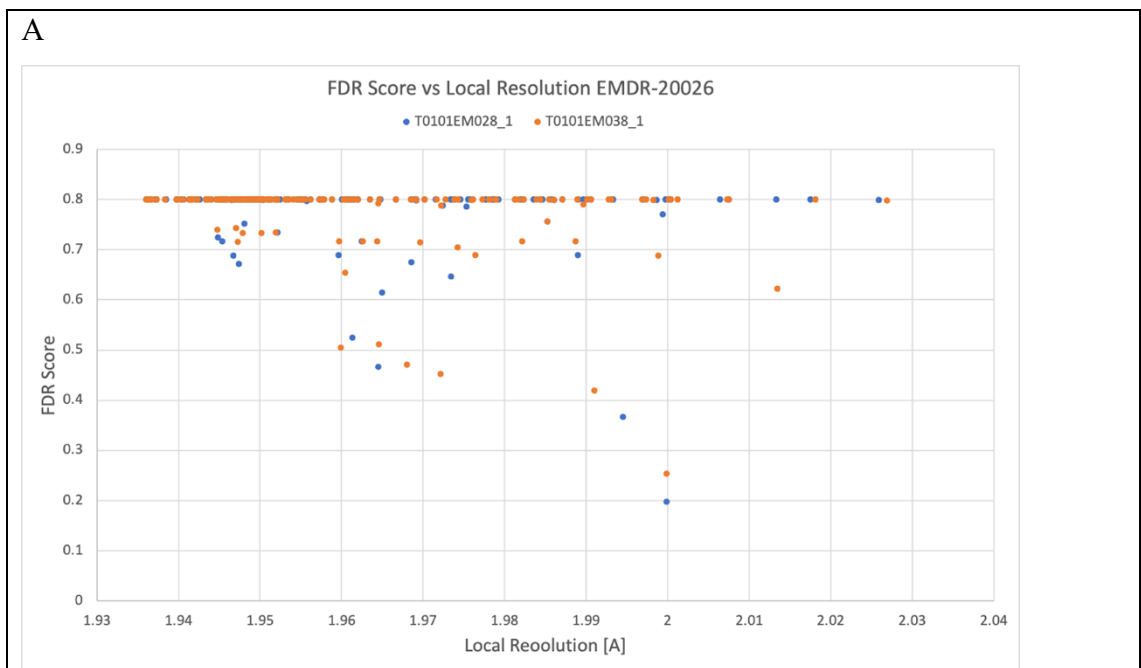
The final resolution of the cryo-EM map is often calculated from the Fourier Shell Correlation between the half-maps. This provides only a global estimate as the resolution can vary locally. This is usually governed by the inherent dynamics, heterogeneity and

other effects from image processing and reconstruction. Other factors that affect the local resolution of the cryo-EM map are specimen symmetry, application of a 3D binary mask that separates molecular signal from noise, sharpening procedure, or map anisotropy caused by the preferred orientation of the sample. Typically, due to averaging effects, the resolution can vary based on the distance from the centre of the molecule, with the lowest resolution on the map periphery.

To minimise the effect of sharpening and masking of the final map, I used the FDR score from the auto-sharpened maps calculated with '1Å' radius mode and Q3 value transposed to the value of 0.8 and compared it against the local resolution map calculated with LocRes implementation in Relion 4.1[163]. The selected dataset includes maps and models submitted for the EMDR 2016 and 2019 model challenges. In each case two models are evaluated, one high-quality model and one low-scoring according to the EMDR Challenge outcomes. The apoferritin models 28_1 and 38_1 built from map EMDR-20026 with local resolutions ranging from 1.94 Å to 2.03 Å (Fig 4.4.A), T20s Proteasome models T0002EM123_2 and T0002EM189_2 from EMDR-6287, with resolution from 2.71 Å to 3.02 Å (Fig. 4.4.B), the alcohol dehydrogenase models were T0104EM028_1 and T0104EM060_2 submitted for the target map EMDR-0406 which covers resolution ranges between 2.96 Å and 3.82 Å (Fig. 4.4.C). From the plots, it can be seen that in every case the overall FDR score gets lower with lower local resolution. In each shown resolution band, models are scored higher in high local resolutions. The low-scoring residues also appear independently from the changes in local resolutions. In conclusion, the FDR score can be used to evaluate the models across the range of resolutions, but local changes in resolution can lower the score. It also needs to be taken into consideration that parts of the maps can be genuinely disordered or poorly refined in lower local resolution areas. The presented analysis does not show a clear correlation between the local resolution of the map and the FDR score.

The information from the local resolution estimation done, for example, with LocRes in Relion or similar programs, can be used as an additional information to calculate the FDR maps by local weighting[204]. This approach would often require the half-maps required to obtain information about the local resolution. This approach was tested and the performance of the FDR score from globally sharpened map was compared to FDR score from local resolution weighted raw and sharpened map of T20s proteasome EMDR-6287 where local resolution ranges from 2.7-3.1 Å. Figure 4.5.A shows comparison of the FDR score plots, showing that even with the local resolution information the raw is useful only to identify the worst outliers. Figure 4.5 in panel B shows the EMDR-6287 map coloured

accordingly to the local resolution, panel C shows the confidence map calculated from auto-sharpened map, panel D- confidence map from raw map with local resolution information and panel E- confidence map from auto-sharpened map with local resolution information. Snapshots of the outliers are provided to further investigate the quality of fit. Glutamate 31 is marked as an outlier by all FDR maps. The raw map, even with the local resolution information, was the only one which did not correctly identify Methionine 101 as an outlier.



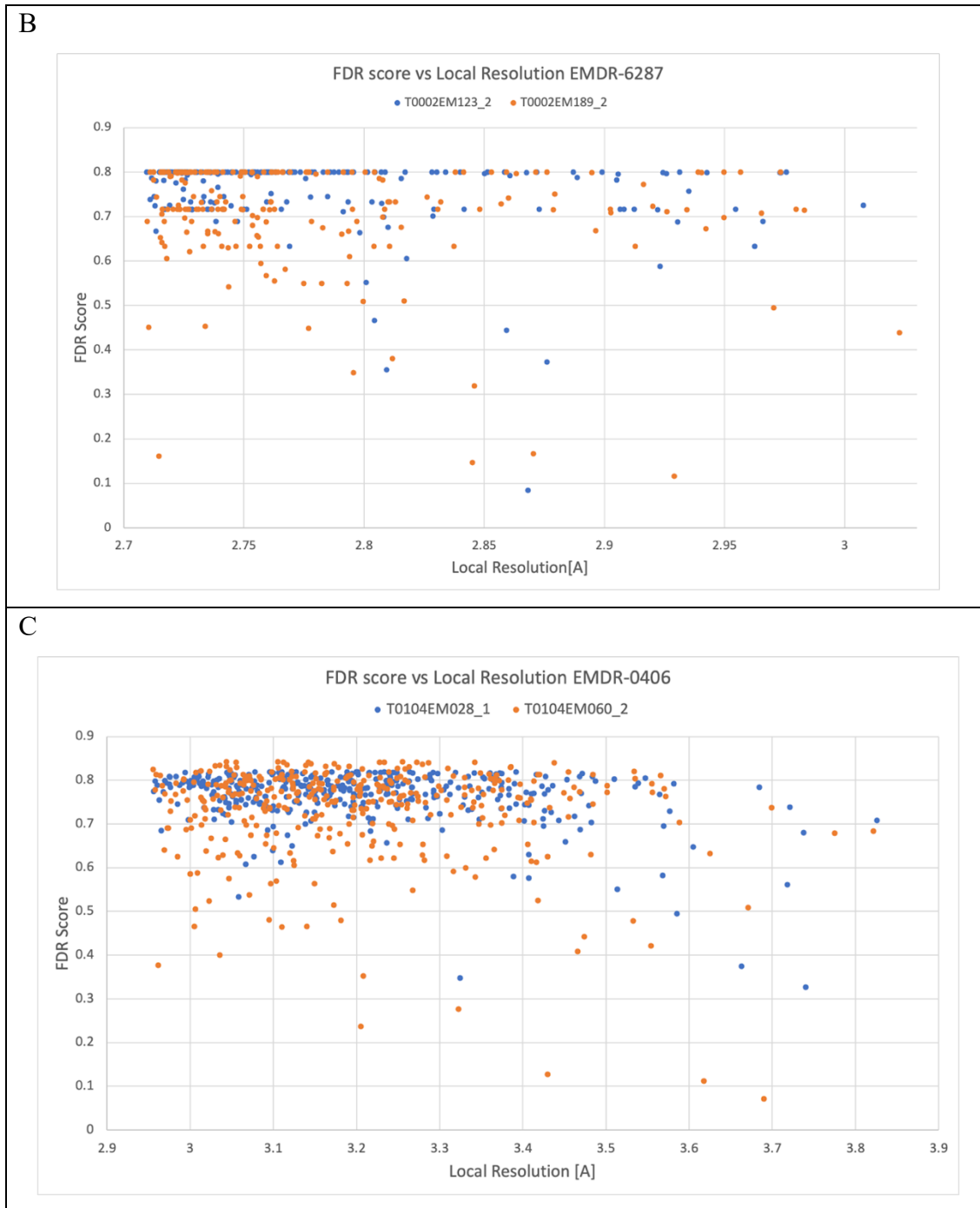


Figure 4.4 Comparison of FDR score of the models at different local resolution reveals lower scores as the local resolution decreases

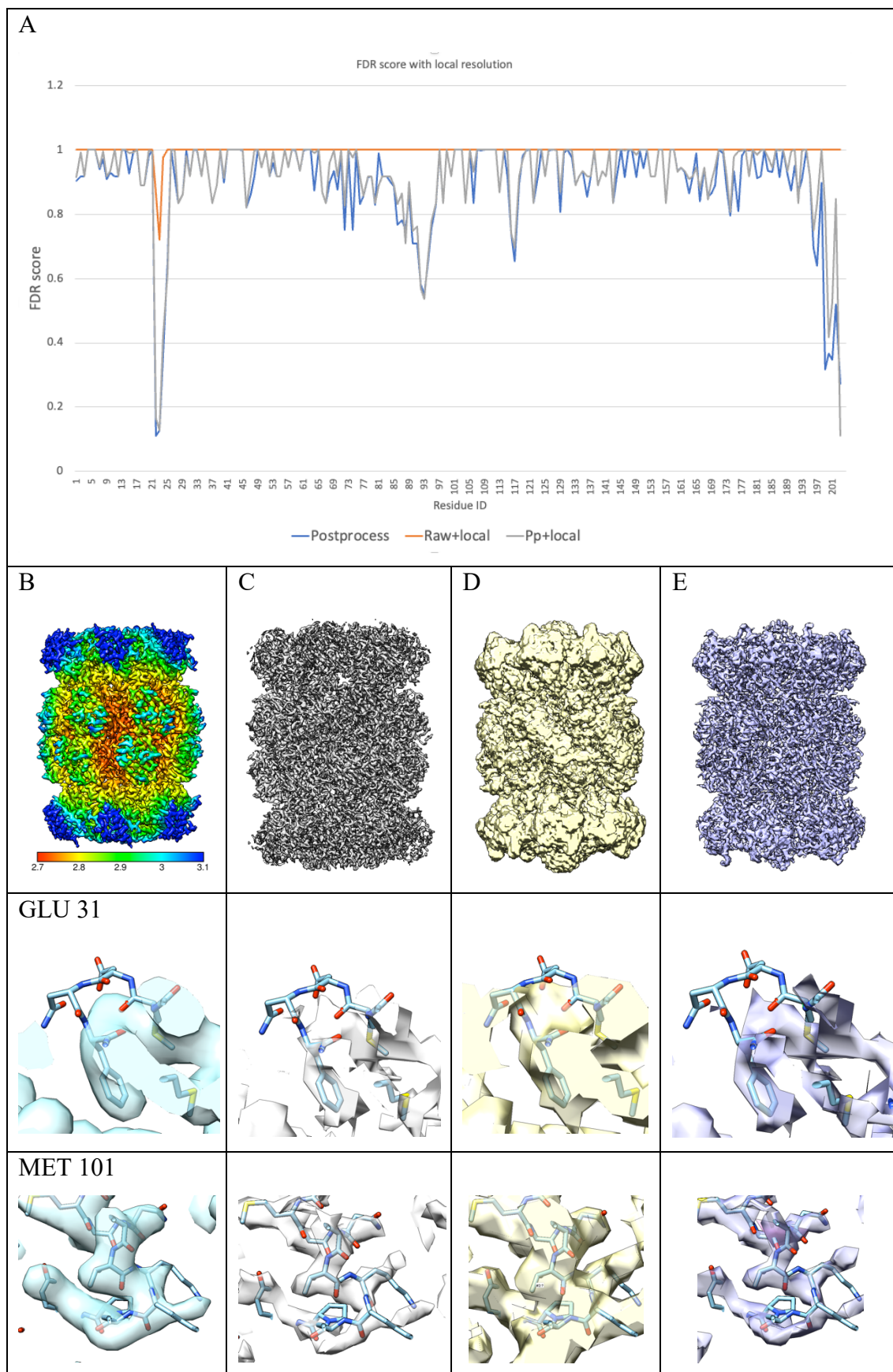


Figure 4.5 Comparison of FDR scores calculated from postprocessed map, raw map weighted with local resolution information and post processed (pp) weighted with local resolution information. For the comparison panel B shows original post-processed density coloured accordingly to the local resolution. C- Confidence map calculated from auto-sharpened map, D- Confidence map calculated from a raw map supported by local resolution information, E- Confidence map calculated from an auto-sharpened map supported by local resolution

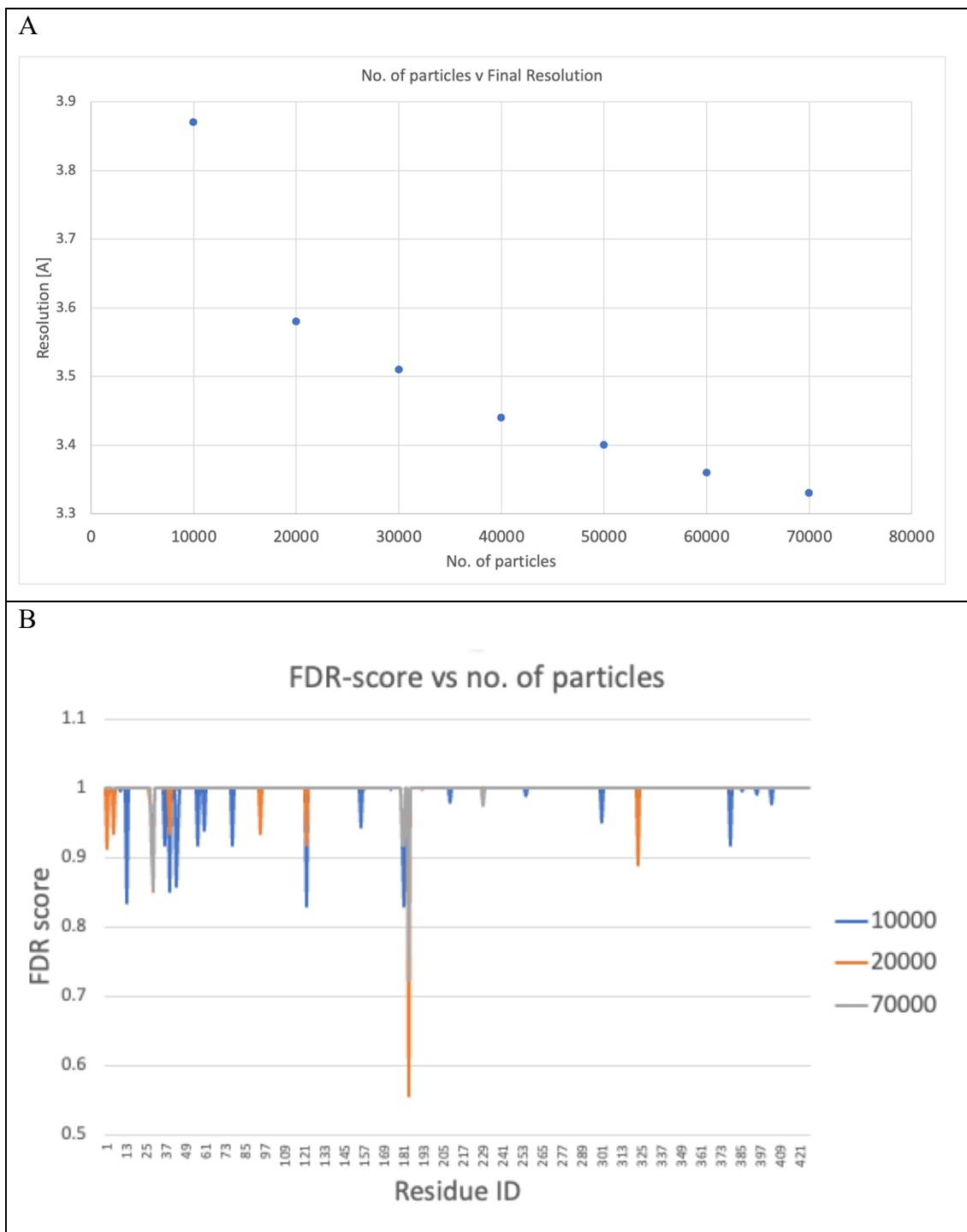
information. Additionally, positions of residues GLU31 and MET101 in the maps are shown for comparison.

4.3.4 Number of particles vs final resolution

The final resolution of the map would be affected by the number of particles used for reconstruction. In theory, as the number of particles increases, the resolution is expected to improve. More particles add to signal about the structure and allow more accurate refinement. However, the relationship between the number of particles and the final resolution is not linear. There are other factors that limit the practically achievable resolution, such as the particle quality, radiation damage, sample heterogeneity, angular view coverage (preferred orientation) or defocus setting of the microscope[122]. To investigate how the number of particles affects the performance of the FDR score, I used the T20s Proteasome EMPIAR-10025 datasets. The data was processed using the Relion 4.1 pipeline, after motion correction with MotionCorr[124], and CTF estimation with CTFIND4[75]. The first round of particle picking was done template-free with the Laplacian of Gaussians method. Initial 2D classes were used for template-based particle picking. After selecting the best classes with a total number of 73263 particles, the ab initio 3D model was generated and used for 3D classification. After that, the subsets were selected by adding 10000 particles each time and running the 3D refinement with D7 symmetry and the same 3D reference map, resulting with final maps refined from 10000 to 70000 particles. The relationship between the number of particles and the final resolution is presented in Figure 4.6.A.

The FDR score was calculated for the unsharpened maps to base the comparison solely on the number of particles and not to introduce the additional parameter of the sharpening factor. Analysis of the FDR scores revealed some residues marked as outliers in each map but also some that, with the maps with the higher number of particles, started to fit better. Figure 4.6.B shows the plots for models validated from maps refined with 10000, 20000 and 70000 particles, as they have the highest difference in the final resolutions. The residues selected for further investigation in UCSF Chimera included Tyrosine (Tyr132, Chain S), which scored low with maps from a lower number of particles and high with 70000 particles (Fig.4.6.C) and Glycine (Gly192, Chain S), which scored low with every map (Fig.4.6.D). To compare the maps, they are represented in UCSF Chimera with the thresholds that represent the same molecular volume. This comparison shows that there are some outliers in the maps, which can be found at different resolution levels. Some residues marked as outliers might be positioned correctly, but the quality of the 3D map

density is not good enough to support them. The higher resolution map has lower number of outliers.



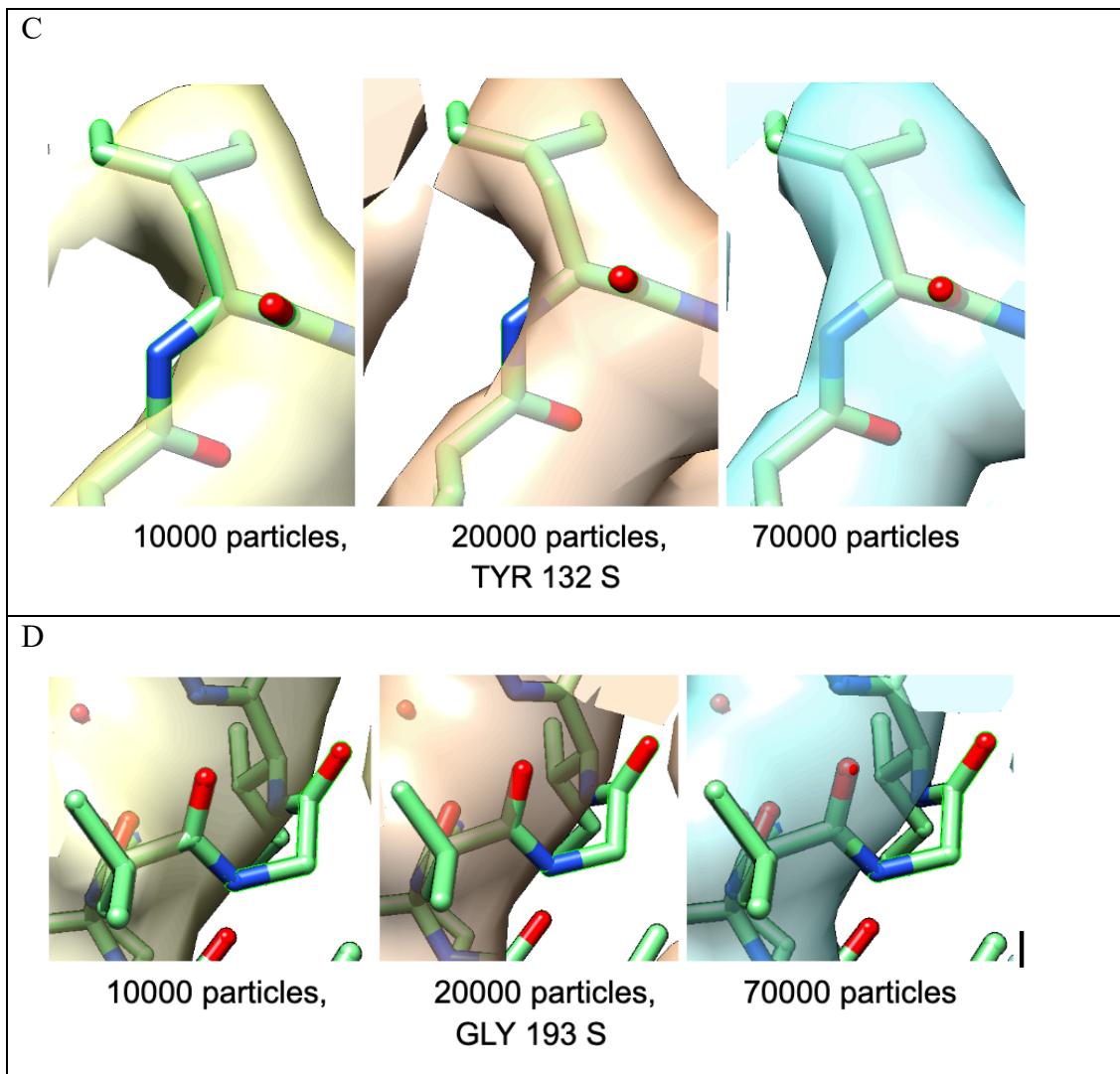


Figure 4.6 FDR score calculated for different numbers of particles used for T20s proteasome map refinement, A) the resolution increases with a higher number of particles, B) FDR scores calculated from different numbers of particles shows that with higher number of particles and higher resolution the residues score better but still some of the outliers are found in every case, C) comparison of the position of Tyrosine 132 chain S shows that with higher number of particles the density improves to fit the model backbone, D) an example of a residue (Glycine 193, chain S) which is not supported by the density at any resolution, This worst outliers can be identified at different resolution ranges.

4.3.5 FDR Score compared to Atom Inclusion

The FDR backbone score was compared with the Atom Inclusion score, which also reports whether the atoms and residues are traced within the molecular contour. The Atom Inclusion score requires the user to select a density threshold for calculating the scores. This value should ideally correspond to the molecular weight of the protein, but the threshold recommended by the author is often subjective. This makes the Atom Inclusion score sensitive to the selection of the contour level, artificially increasing the assessed model quality if the threshold is too low or lowering it if it is too high. Additionally, the globally set threshold might not properly represent the local variability of the density map. The new method based on the False Discovery Rate does not require users to specify such

parameters as it separates the molecular density from background noise. The inclusion score for residues is calculated for all atoms, while the FDR score uses only C-alpha, C and N. The Validation tools were compared based on the potentially misfitted residues they can identify. The Atom Inclusion scores are calculated for the deposited primary map at the contour level recommended during map deposition. The 3ajo model was validated against EMD-20026 apoferritin map (Fig.4.7). Two areas were selected, to investigate cases where Atom Inclusion scores were low and FDR scores high (B). In the area where the sidechains are not supported by the density the Atom Inclusion reports low score for the whole residue, where the FDR score shows the good fit of the residue's backbone. Both methods score low where the local density is not supporting the model (C).

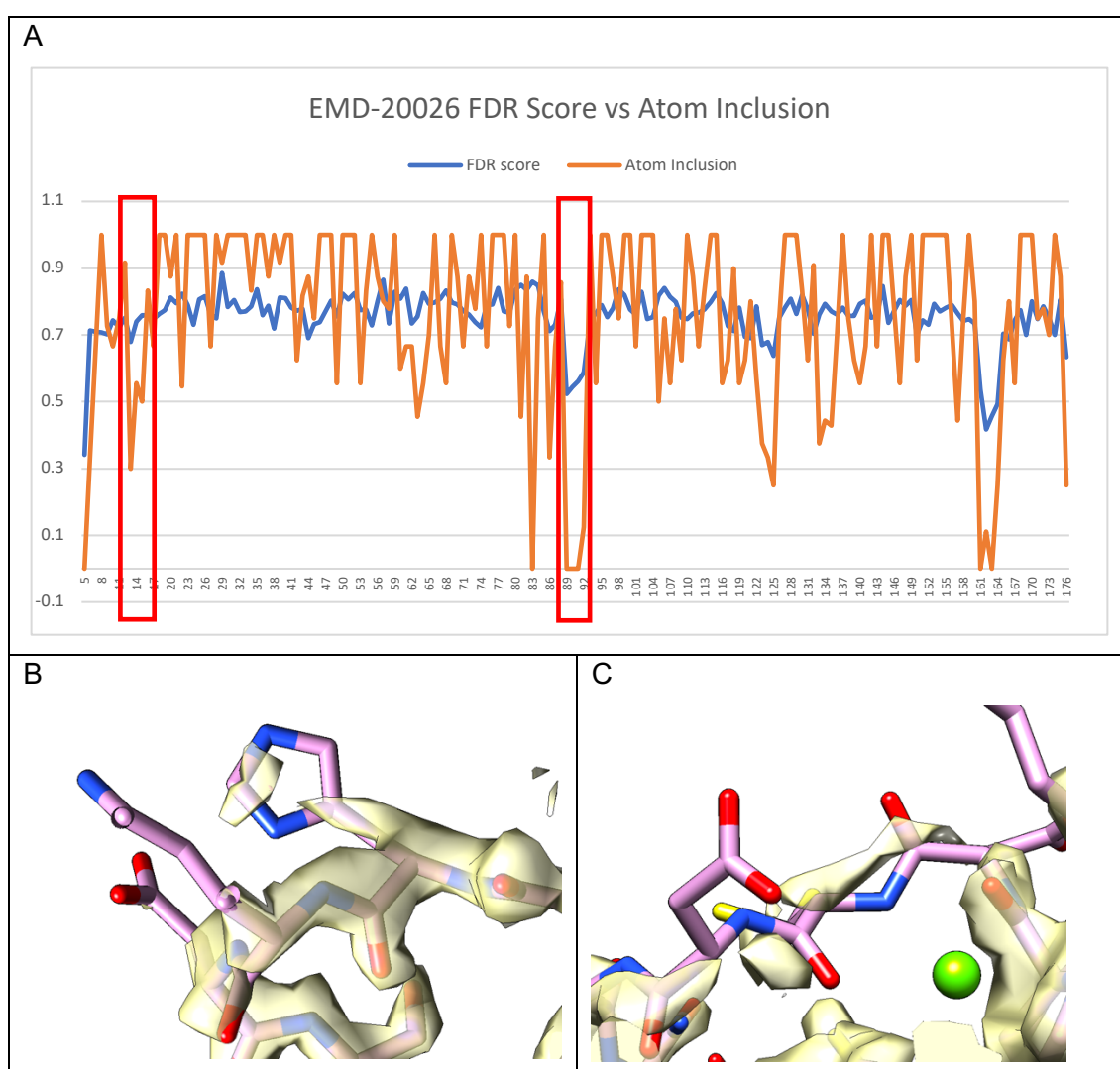


Figure 4.7 Comparison of the FDR score and Atom Inclusion for PDB 3ajo model and EMD-20026 map, A) plots of the FDR score and Atom inclusion shows regions where the scores are different, B) TYR 12 – ASP 15 with low Atom Inclusion score due to the lack of support for the side chains, C) ASP 89 – ASP 92 where both methods score low due to poorly refined density

The PDB 3j9i T20s proteasome model was scored both with FDR using auto-sharpened EMD-6287 map and Atom Inclusion with the map thresholded at 0.025 level. The

comparison presented in Fig 4.8.A reveals areas of the model which is scored differently by each method. Figure 4.8.B shows that the Atom Inclusion score gets lower if the side-chains of the residues are not included even if the backbone is well placed into the density. Fig. 4.8.C shows part of the model, where the FDR score identified the misplaced backbone, but the atom inclusion gave high overall scores to the residues as other atoms are included within it.

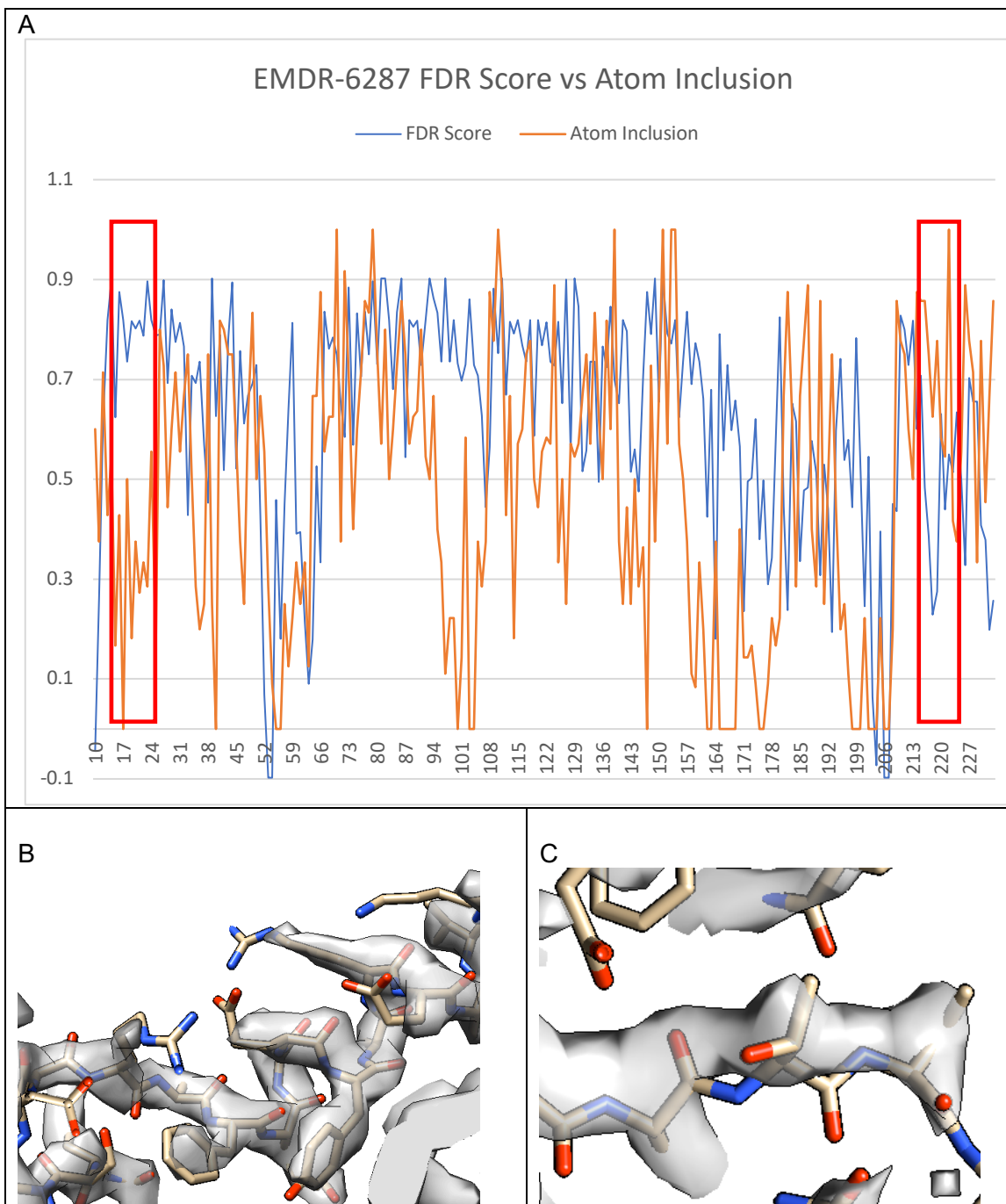


Figure 4.8 Comparison of the FDR score and Atom Inclusion for PDB 3j9i model and EMD-6287 map, A) plots of the FDR score and Atom inclusion, B) PRO 17 – ALA 27 with low Atom Inclusion score due to low visualisation threshold that excludes side-chains, C) ILE 215 – GLY 218 where FDR score identified backbone atoms out of the density, but Atom Inclusion scored the residues highly as atoms from side chains are shifted into density

Comparison of the FDR score and Atom Inclusion for alcohol dehydrogenase model PDB 6nbb and map EMDR-0406 in Fig 4.9.B shows the area scoring low with both methods where the residues are not supported by the density at the beginning of the chain A. Fig. 4.9.C. shows that the recommended threshold level 0.02 introduces gaps in the density resulting in low Atom Inclusion score, where the FDR method gives high score to the residues.

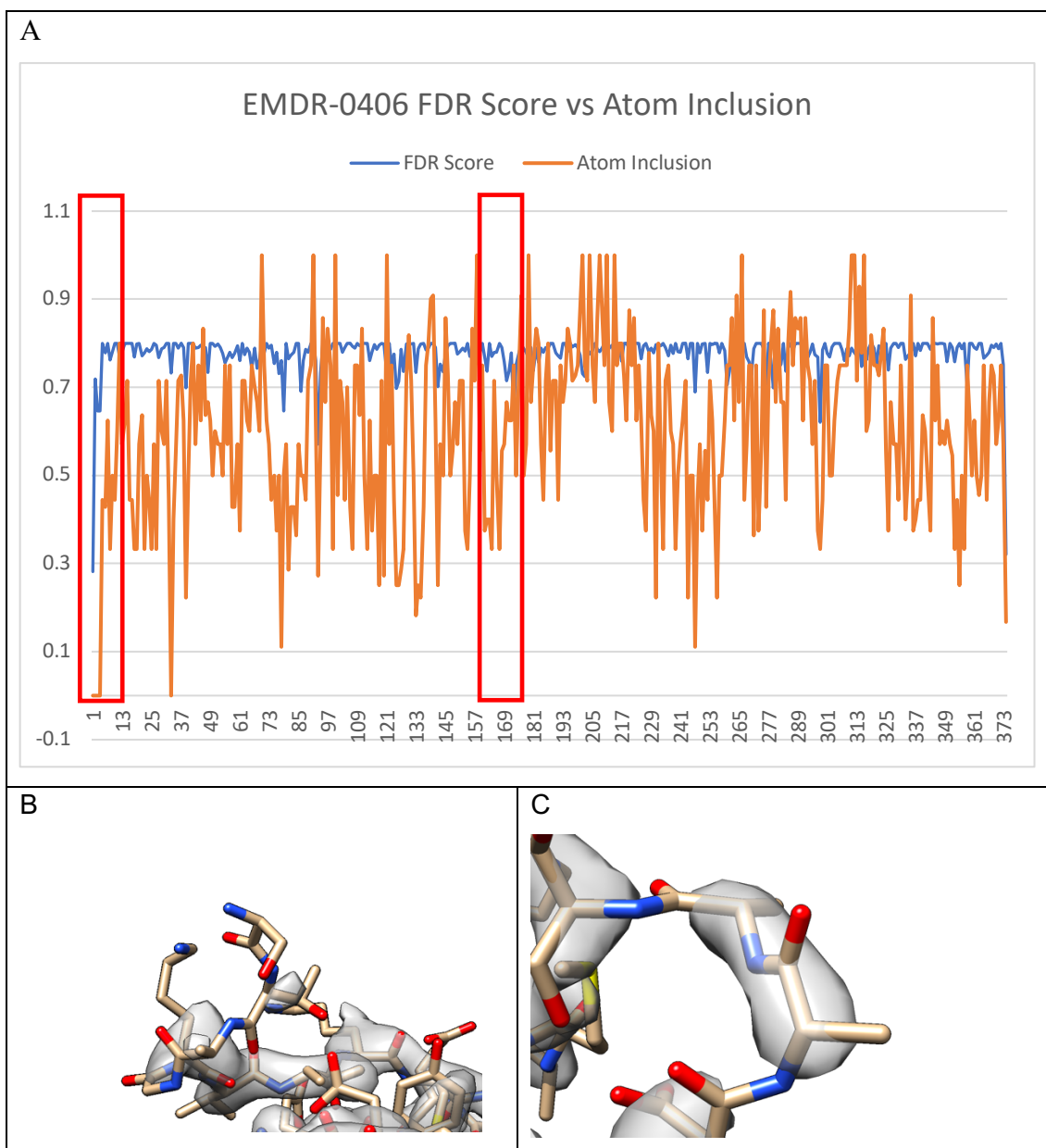


Figure 4.9 Comparison of the FDR score and Atom Inclusion for PDB 6nbb model and EMD-0406 map, A) plots of the FDR score and Atom inclusion, B) SER 1 – GLY 4 with both scores low as there is no map density to support the residues C) ALA 162 - SER 164, too high visualisation threshold creates gaps in the density resulting in low Atom Inclusion score, the FDR score is high.

4.4 Conclusions and future works

The FDR score is affected by the sharpening level of the map and the proposed approach of moving the third quartile of the distribution to 0.8 helps to align the differences in the scale of FDR scores. This doesn't account for the differences in the outliers identified at different sharpening levels. Nevertheless, it makes it possible to identify the worst outliers between the differently sharpened maps. By default, the auto-sharpened map is recommended for FDR score calculation and observed to be more useful in identifying issues with model backbone trace.

The local changes in resolution and quality of the maps can also affect the FDR scores. One way to approach this challenge could be weighting the Confidence Map according to the estimated local resolution. Examples presented in this chapter did not show considerable improvement in outlier detection compared to default Confidence Map calculation. Obtaining the local resolution map requires additional input (unfiltered half-maps) and extra computational steps of calculating the local resolutions reference and then weighting the FDR map. In order to investigate this problem fully, a reasonably large dataset of cryo-EM maps deposited with half-maps would be required. Also, the collection of maps should not only include entries with different global resolutions, but also different levels of variance in the local resolution to see to what extent the local changes in resolution can be compensated with local sharpening. This kind of project could be beneficial for the cryo-EM community and add another tool for local model quality evaluation, but should also allocate a significant amount of time for collecting the proper test dataset, scoring the models and finally, analysing the results.

The number of particles affects the overall quality and resolution of the map. Maps refined with fewer particles contain fewer outliers, but the worst outliers are also identified when a high number of particles is used.

Based on the performance of the score in identifying errors in models, the 1Å radius mode performs better than the radius defined by Van der Waals of each atom. We note that some residues can score high if they fit into the density even at an incorrect position but within the molecular volume. In the future, another layer of validation can be added by also checking the quality of a group of residues in specific areas, which can be implemented using the sliding window technique or similar.

The comparison of the FDR Score with Atom Inclusion showed that Atom Inclusion is sensitive to the threshold selection, additionally if too high threshold is selected it can introduce discontinuities in the map and a large number of identified issues with atom

positions. Finally, the Atom Inclusion score calculated per residue uses all atoms in the residue, which can still score high even if the backbone atoms are outside of the map.

5 Conclusion

5.1 Summary

As a result of this PhD project, automated software tools and procedures were developed to annotate the cryo-EM images. The introduction of a new parameter based on the local pixel intensity, which can be calibrated to the measured ice thickness, can improve the data processing. This parameter can be used at the initial stages of data processing to evaluate the overall quality of the micrographs. It can also annotate individual particles based on local ice conditions to investigate their quality and behaviour. With many new tools used for sample preparation developed in recent years, the major challenges are still the lack of reproducibility and ice distribution control. There are established procedures that can identify the sample quality and ice condition in the sample, but they require additional steps during data collection (e.g. stage tilting) or specific microscope setup (energy filter or Aperture Limited Scattering methods). The main advantage of the presented approach is that it requires minimal additional steps, and then only if the calibration is required. It can also be successfully run for historical datasets for additional analysis.

In Chapter 3, we presented additional analysis and applications of the introduced software tool, which can be used to locally improve the contrast and estimate the ice distribution on the cryo-EM micrographs. Estimating the ice thickness distribution allows users to annotate each of the picked coordinates with the average pixel intensity and identify the contamination regions or to group particles according to the local ice conditions. Users can group together particles from similar ice thickness areas and run local refinement to check the resolution and distribution of the angular orientations of the particles with cryoEF or similar tools. This way, the optimal ice conditions for a given specimen, which will provide a high signal-to-noise ratio and good angular coverage, can be identified and targeted during the data collection at the high-end microscope. The presented software can also be used to improve particle picking from a new set of micrographs with uniform contrast distribution. This approach is used to pick as many particles as possible, considering that particles recovered from thick ice regions might have lower quality. In some cases, where the particles have preferred orientations in the data set, analysis of different ice thickness regions can reveal some of the unique views and improve the initial 3D model despite the lower quality of individual particles. With these options, users can not only improve the particle picking with more picks from the regions that were

previously skipped because of the non-uniform contrast distribution but also apply the local or global pixel intensity threshold to remove the contaminations from picks.

The software was written in Python 3 and is also implemented as an "External job" type in Relion. It is freely available from GitHub or can be installed from PyPI (<https://pypi.org/project/icebreaker-em>). With this approach, we would like to make it easily accessible for users and encourage them to install it and use it at their local institution. The initial ice conditions evaluation, even at the mid-range microscope, can be used to identify the most promising ice thickness levels, for example, those which would provide good angular view coverage and a high signal-to-noise ratio. This approach could provide additional information about overall sample quality and more effectively plan the data collection at the high-end microscope.

There are existing atomic model validation software tools that use different techniques. They allow addressing a variety of potential problems with the model, from the global and local fit to the map, backbone and rotamers geometry or protein-protein interfaces. As the different programs specialise in detecting particular issues with the model, it is always recommended to use a proper validation approach or multiple metrics to get a comprehensive evaluation of the final model. As a part of this research, a new model validation tool based on the False Discovery Rate was developed with the outcomes summarised below.

Chapter 4 presents the additional evaluation of an atomic model validation tool based on the False Discovery Rate approach. The tool allows users to determine if the parts of the model are built into the actual cryo-EM density or if they are placed into the noise. Each residue is ranked, and users can investigate the potentially problematic ones in a visualisation tool like the UCSF Chimera or automatically remove the low-scored ones. We also presented a detailed comparison between this tool and other commonly used metrics such as FSC-Q, Map Q-score, PHENIX validation toolbox and TEMPy SMOC and SCCC scores. This study revealed the benefits of using multiple validation tools based on different methods and metrics, as some of the problems with models can be easily spotted with one and overlooked by the others. This software tool is now implemented and available from the CCPEM software suite for the cryo-EM community.

5.2 Limitations and future work

Progress in cryo-EM data collection and processing can be achieved either by upgrading instrumentation or by the introduction of new software tools. As the hardware upgrade is always something that makes the most important changes, the software can be created to pinpoint some specific problems at different stages of the data processing and to automate the processing procedures to routinely obtain high-resolution cryo-EM reconstructions. The main limitation encountered during the research presented in this thesis is the limited availability of unoptimised or failed datasets. Even if the quality of the final resolution cryo-EM map is low or the model is incomplete, the data is well curated, and people do not frequently deposit flawed datasets to EMPIAR or PDB repositories. Instead, they try to optimise the sample preparation procedure and try using a new dataset. The availability of a range of datasets with different quality is crucial for software development to properly understand and describe the problem and then find a correct method to solve it or automate the process. The annotated negative and false negative datasets are also a vital part of the machine learning model training workflow to train a neural network with high precision and recall. The development of software for users to improve data processing heavily depends on the feedback provided by users. The opportunity to present these projects during various seminars and poster sessions, as well as collaboration with Electron Bio-Imaging Centre and Collaborative Computational Project for Electron Microscopy during this PhD, allowed me to reach out to a large potential user base and better understand the needs of cryo-EM community.

The IceBreaker software was introduced to the data collection and processing pipeline at eBIC. We hope this will encourage users to use it for their data acquisition sessions. The software was developed and tested mostly on historical datasets, and feedback from actual users can help further optimise the parameters for data collection. This software can be run automatically, without user interaction for on-the-fly processing. We developed a robust ice-thickness estimation framework, which does not interrupt the data collection setup and helps people to make the most of their experiments. In the future, the contrast-improved micrographs can be used as input data for machine-learning based automated particle pickers, such as ‘fine-tuning’ with crYOLO. The software is parallelised to run on multiple CPUs, which works reasonably well given the higher availability of CPUs as the GPUs are used for computationally heavy tasks. The ice thickness levels are estimated from the pixel intensity values recorded by the detector. The pixel intensity values are now compared with ice thickness measured during data collection with and without an

energy filter. If the IceBreaker becomes a tool commonly used during the data collection sessions, we could build a large database of high-quality particles with additional information about the ice conditions from which they were picked. The long-term goal would be to use this information to train a machine learning model to automatically select the optimal ice conditions for data collection, even at lower magnification levels.

The developed and published model validation tool is a new addition to the CCPEM software suite. It offers the functions to automatically check the protein backbone fit into the cryo-EM density and the option to remove residues fitted poorly or into the background noise. The software evaluates only the backbone atom positions, and the side chains or rotamer positions are not considered. This approach should work reliably for the lower resolution range of 2.8-3.5 Å, where the misfit of the backbone usually causes the side chain misplacement. It was shown that in some cases, it could perform better than the Atom Inclusion score used as a validation during map and model deposition but relies on the selection of interpretation level, which in many cases might not be optimal. With this software, we score the relative position of atomic coordinates according to the cryo-EM density. The other useful functionality would be to check for sequence register errors, as even if the residue is incorrect but fits into density, it might receive a high score. The preliminary results do not show a strong correlation between the confidence score and local changes in resolution, which can also be caused by protein flexibility or data heterogeneity. Further development in that direction could improve the performance of this software and help to identify problems which originate from incorrect refinement or from the quality of collected data. In general, most of the automated model-building tools do not consider local resolution changes. The ‘pruning’ option removes poorly scored residues and can be useful for iterative model building. Incorrectly built parts of the current model can be automatically deleted, and the remaining part can be used as input for the next model building round. The comparison with other commonly used validation tools showed this tool could be complementary to them and identify some of the issues missed by other methods.

5.3 Closing remarks

The cryo-electron microscopy technique for protein structure determination is becoming more popular and accessible to users. The quality and number of deposited maps have improved greatly in recent years. In 2020, the 1.22 Å apoferritin map was obtained as the

first cryo-EM structure at atomic resolution. What since the 1970s was considered wishful thinking supported only by theoretical calculations finally became reality.

The high-throughput data collection and improved quality of recorded data bring new challenges for the cryo-EM. The vital task is to optimise the data collection strategies to make the most of the microscope time. This can be done by automatically targeting the regions of cryo-EM grids, which can result in high-quality reconstruction. Currently, one of the major concerns for data processing is the structural heterogeneity and anisotropic resolution, which could not even be determined at lower resolutions a few years ago. If the particles in the sample have different conformations and it is not identified and properly classified, it might not be possible to obtain high-resolution 3D reconstruction. The local changes in resolution can heavily affect the automated model-building techniques, leading to incorrect or incomplete models.

The EMDataResource and Electron Microscopy Public Image Archive (EMPIAR) data repositories are invaluable resources of historical cryo-EM data with details about the experiment setup. The datasets deposited in a unified way are useful for identifying common issues and can be a great starting point for the development of new methods. Initiatives like EMDataBank map building and model validation challenges are great platforms to bring together the cryo-EM community to benchmark the state-of-the-art tools for cryo-EM data processing. This way, new procedures and workflows can be established and presented to users with instructions and recommendations on how to effectively use them. In the research presented in this thesis, we often used targets from such challenges as they allowed us to evaluate our results, see the improvements, and identify flaws in our approach. Unfortunately, in many cases, the databases do not provide enough bad and low-quality data, which is also essential for new methods of development and validation.

The future developments in the cryo-EM could be done in two major directions. The first approach is improvements in hardware equipment, including new types of detectors, that would allow obtaining higher quality images, which can lead to better quality and resolution of the reconstructions. The development of new sample deposition devices and procedures could help to improve the sample stability and reproducibility. The other approach is software development with the rise of machine learning algorithms that could be applied to automate the data collection and processing pipeline even further.

High-end cryo-EM microscopes are still very expensive to operate, and only a few facilities worldwide can afford them. The total operational cost of running the cryo-EM microscope and equipment can be from £3000-£5000 per day. Recent developments in cryo-EM imaging show that final maps with resolutions better than 5Å (3.4 Å resolution from 16 500 particles of DPS protein with tetrahedral symmetry) are achievable with lower-end equipment and 100keV accelerating voltage[50, p. 100]. The possibility of better availability of cryo-EM instruments for users at reduced costs can be beneficial for better planning and optimisation of data collection. In most cases, the achievable resolution should be sufficient for a high-throughput preliminary screening of drug candidates or the initial investigation of the specimen parameters and behaviour in the sample before high-resolution data collection at the high-end facility. This approach, if commonly adopted, could help to make the most of the true cryo-EM potential with high-end, high-resolution processing. At the same time, lower-end microscopes could serve for training purposes to familiarise people with cryo-EM data collection procedures and serve as a platform for testing and developing new software for automated data collection optimisation and processing without requiring costly downtime of the high-end devices.

6 Bibliography

- [1] C. H. Wu, H. Huang, L.-S. L. Yeh, and W. C. Barker, 'Protein family classification and functional annotation', *Computational Biology and Chemistry*, vol. 27, no. 1, pp. 37–47, Feb. 2003, doi: 10.1016/S1476-9271(02)00098-1.
- [2] R. Apweiler *et al.*, 'UniProt: the Universal Protein knowledgebase', *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D115–D119, Jan. 2004, doi: 10.1093/nar/gkh131.
- [3] A. L. Lehninger, D. L. Nelson, and M. M. Cox, *Lehninger principles of biochemistry*, 6th ed. New York: W.H. Freeman, 2013.
- [4] D. N. Forthal, 'Functions of Antibodies', *Microbiol Spectr*, vol. 2, no. 4, pp. 1–17, Aug. 2014.
- [5] M. L. Chiu, D. R. Goulet, A. Teplyakov, and G. L. Gilliland, 'Antibody Structure and Function: The Basis for Engineering Therapeutics', *Antibodies((Basel))*, vol. 8, no. 4, p. 55, Dec. 2019, doi: 10.3390/antib8040055.
- [6] P. K. Robinson, 'Enzymes: principles and biotechnological applications', *Essays Biochem*, vol. 59, pp. 1–41, Nov. 2015, doi: 10.1042/bse0590001.
- [7] R. Sharma, Y. Chisti, and U. C. Banerjee, 'Production, purification, characterization, and applications of lipases', *Biotechnology Advances*, vol. 19, no. 8, pp. 627–662, Dec. 2001, doi: 10.1016/S0734-9750(01)00086-6.
- [8] 'Industrial Uses of Enzymes'. Accessed: Sep. 28, 2022. [Online]. Available: <https://www.biotecharticles.com/Applications-Article/Industrial-Uses-of-Enzymes-930.html>
- [9] A. Tripathi and V. A. Bankaitis, 'Molecular Docking: From Lock and Key to Combination Lock', *J Mol Med Clin Appl*, vol. 2, no. 1, p. 10.16966/2575-0305.106, 2017.
- [10] S. Martínez Cuesta, S. A. Rahman, N. Furnham, and J. M. Thornton, 'The Classification and Evolution of Enzyme Function', *Biophys J*, vol. 109, no. 6, pp. 1082–1086, Sep. 2015, doi: 10.1016/j.bpj.2015.04.020.
- [11] M. D. Shoulders and R. T. Raines, 'COLLAGEN STRUCTURE AND STABILITY', *Annu Rev Biochem*, vol. 78, pp. 929–958, 2009, doi: 10.1146/annurev.biochem.77.032207.120833.
- [12] W. Zhang and Y. Fan, 'Structure of Keratin', *Methods Mol Biol*, vol. 2347, pp. 41–53, 2021, doi: 10.1007/978-1-0716-1574-4_5.
- [13] M. J. Lopez and S. S. Mohiuddin, 'Biochemistry, Essential Amino Acids', in *StatPearls [Internet]*, StatPearls Publishing, 2024. Accessed: Sep. 04, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK557845/>
- [14] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, 'The Shape and Structure of Proteins', *Molecular Biology of the Cell. 4th edition*, 2002, Accessed: Sep. 28, 2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK26830/>
- [15] E. C. Griffith and V. Vaida, 'In situ observation of peptide bond formation at the water–air interface', *Proc Natl Acad Sci U S A*, vol. 109, no. 39, pp. 15697–15701, Sep. 2012, doi: 10.1073/pnas.1210029109.
- [16] G. Faure, A. Bornot, and A. G. de Brevern, 'Protein contacts, inter-residue interactions and side-chain modelling', *Biochimie*, vol. 90, no. 4, pp. 626–639, Apr. 2008, doi: 10.1016/j.biochi.2007.11.007.
- [17] L. Pauling, R. B. Corey, and H. R. Branson, 'The Structure of Proteins', *Proc Natl Acad Sci U S A*, vol. 37, no. 4, pp. 205–211, Apr. 1951.

- [18] C.-I. Brändén and J. Tooze, *Introduction to protein structure*, 2nd ed. New York: Garland Pub, 1999.
- [19] P.-N. Cheng, J. D. Pham, and J. S. Nowick, ‘The Supramolecular Chemistry of β -Sheets’, *Journal of the American Chemical Society*, vol. 135, no. 15, p. 5477, Apr. 2013, doi: 10.1021/ja3088407.
- [20] L. E. Donate, S. D. Rufino, L. H. Canard, and T. L. Blundell, ‘Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction.’, *Protein Science: A Publication of the Protein Society*, vol. 5, no. 12, p. 2600, Dec. 1996, doi: 10.1002/pro.5560051223.
- [21] B. L. Sibanda and J. M. Thornton, ‘ β -Hairpin families in globular proteins’, *Nature*, vol. 316, no. 6024, pp. 170–174, Jul. 1985, doi: 10.1038/316170a0.
- [22] I. Rehman, C. C. Kerndt, and S. Botelho, ‘Biochemistry, Tertiary Protein Structure’, in *StatPearls [Internet]*, StatPearls Publishing, 2024. Accessed: Sep. 04, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK470269/>
- [23] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, ‘The Protein Folding Problem’, *Annual review of biophysics*, vol. 37, p. 289, Jun. 2008, doi: 10.1146/annurev.biophys.37.092707.153558.
- [24] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, ‘Stereochemistry of polypeptide chain configurations’, *Journal of Molecular Biology*, vol. 7, no. 1, pp. 95–99, Jul. 1963, doi: 10.1016/S0022-2836(63)80023-6.
- [25] A. Q. Zhou, C. S. O’Hern, and L. Regan, ‘Revisiting the Ramachandran plot from a new angle’, *Protein Sci*, vol. 20, no. 7, pp. 1166–1171, Jul. 2011, doi: 10.1002/pro.644.
- [26] G. J. Kleywegt and T. A. Jones, ‘Phi/psi-chology: Ramachandran revisited’, *Structure*, vol. 4, no. 12, pp. 1395–1400, Dec. 1996, doi: 10.1016/s0969-2126(96)00147-5.
- [27] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North, ‘Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis’, *Nature*, vol. 185, no. 4711, pp. 416–422, Feb. 1960, doi: 10.1038/185416a0.
- [28] J. Dubochet and A. W. McDowell, ‘VITRIFICATION OF PURE WATER FOR ELECTRON MICROSCOPY’, *Journal of Microscopy*, vol. 124, no. 3, pp. 3–4, Dec. 1981, doi: 10.1111/j.1365-2818.1981.tb02483.x.
- [29] W. Kühlbrandt, ‘The Resolution Revolution’, *Science*, vol. 343, no. 6178, pp. 1443–1444, Mar. 2014, doi: 10.1126/science.1251652.
- [30] EMDB, ‘Electron Microscopy Data Bank’, Electron Microscopy Data Bank. Accessed: Sep. 04, 2024. [Online]. Available: https://www.ebi.ac.uk/emdb/statistics/emdb_resolution_year
- [31] J. Hughes, S. Rees, S. Kalindjian, and K. Philpott, ‘Principles of early drug discovery’, *Br J Pharmacol*, vol. 162, no. 6, pp. 1239–1249, Mar. 2011, doi: 10.1111/j.1476-5381.2010.01127.x.
- [32] M. A. Rezaei, Y. Li, D. Wu, X. Li, and C. Li, ‘Deep Learning in Drug Design: Protein-Ligand Binding Affinity Prediction’, *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 19, no. 1, p. 407, Feb. 2022, doi: 10.1109/TCBB.2020.3046945.
- [33] G. Klebe, ‘Recent developments in structure-based drug design’, *J Mol Med*, vol. 78, no. 5, pp. 269–281, Jul. 2000, doi: 10.1007/s001090000084.
- [34] G. Weissenberger, R. J. M. Henderikx, and P. J. Peters, ‘Understanding the invisible hands of sample preparation for cryo-EM’, *Nat Methods*, vol. 18, no. 5, pp. 463–471, May 2021, doi: 10.1038/s41592-021-01130-6.

- [35] I. Angert, C. Burmester, C. Dinges, H. Rose, and R. R. Schröder, 'Elastic and inelastic scattering cross-sections of amorphous layers of carbon and vitrified ice', *Ultramicroscopy*, vol. 63, no. 3, pp. 181–192, Jul. 1996, doi: 10.1016/0304-3991(96)00036-8.
- [36] W. J. Rice *et al.*, 'Routine determination of ice thickness for cryo-EM grids', *Journal of Structural Biology*, vol. 204, no. 1, pp. 38–44, Oct. 2018, doi: 10.1016/j.jsb.2018.06.007.
- [37] J. V. Peck, J. F. Fay, and J. D. Strauss, 'High-speed high-resolution data collection on a 200 keV cryo-TEM', *IUCrJ*, vol. 9, no. Pt 2, pp. 243–252, Jan. 2022, doi: 10.1107/S2052252522000069.
- [38] 'Comparison of Crystallography, NMR and EM'. Accessed: Sep. 28, 2022. [Online]. Available: https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm
- [39] Borlinghaus, Rolf T, 'Super-Resolution - On a Heuristic Point of View About the Resolution of a Light Microscope', *Analytik NEWS*, 02 2015.
- [40] X. Chen, B. Zheng, and H. Liu, 'Optical and digital microscopic imaging techniques and applications in pathology', *Analytical Cellular Pathology ((Amsterdam))*, vol. 34, no. 1–2, p. 5, 2011, doi: 10.3233/ACP-2011-0006.
- [41] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, 'Preparation of slides and coverslips for microscopy', *CSH Protoc*, vol. 2008, p. pdb.prot4988, May 2008, doi: 10.1101/pdb.prot4988.
- [42] *Principles of Protein X-Ray Crystallography*. New York, NY: Springer, 2007. doi: 10.1007/0-387-33746-6.
- [43] M. A. Dessau and Y. Modis, 'Protein Crystallization for X-ray Crystallography', *J Vis Exp*, no. 47, p. 2285, Jan. 2011, doi: 10.3791/2285.
- [44] A. M. Davis, S. J. Teague, and G. J. Kleywegt, 'Application and Limitations of X-ray Crystallographic Data in Structure-Based Ligand and Drug Design', *Angew. Chem. Int. Ed.*, vol. 42, no. 24, pp. 2718–2736, Jun. 2003, doi: 10.1002/anie.200200539.
- [45] T. C. Terwilliger, 'Using prime-and-switch phasing to reduce model bias in molecular replacement', *Acta Crystallogr D Biol Crystallogr*, vol. 60, no. 12, pp. 2144–2149, Dec. 2004, doi: 10.1107/S0907444904019535.
- [46] W. A. Hendrickson, 'Synchrotron crystallography', *Trends in Biochemical Sciences*, vol. 25, no. 12, pp. 637–643, Dec. 2000, doi: 10.1016/S0968-0004(00)01721-7.
- [47] 'Basics of NMR Sample preparation and analysis of NMR analysis data - Mesbah Energy'. Accessed: Aug. 13, 2024. [Online]. Available: <https://www.irisotope.com/en/blog/33/basics-of-nmr-sample-preparation-and-analysis-of-nmr-analysis-data>
- [48] J. C. Chatham and S. J. Blackband, 'Nuclear Magnetic Resonance Spectroscopy and Imaging in Animal Research', *ILAR Journal*, vol. 42, no. 3, pp. 189–208, Jan. 2001, doi: 10.1093/ilar.42.3.189.
- [49] D. Kumar, B. Singh, K. Bauddh, and J. Korstad, 'Title: Bio-oil and biodiesel as biofuels derived from microalgal oil and their characterization by using instrumental techniques', 2015, pp. 87–96.
- [50] K. Naydenova *et al.*, 'CryoEM at 100 keV: a demonstration and prospects', *IUCrJ*, vol. 6, no. Pt 6, pp. 1086–1098, Oct. 2019, doi: 10.1107/S2052252519012612.
- [51] D. Bhella, 'Cryo-electron microscopy: an introduction to the technique, and considerations when working to establish a national facility', *Biophys Rev*, vol. 11, no. 4, pp. 515–519, Aug. 2019, doi: 10.1007/s12551-019-00571-w.
- [52] L. Y. Kim *et al.*, 'Benchmarking cryo-EM Single Particle Analysis Workflow', *Front Mol Biosci*, vol. 5, p. 50, Jun. 2018, doi: 10.3389/fmolb.2018.00050.

- [53] N. D. Lang and W. Kohn, ‘Theory of Metal Surfaces: Work Function’, *Phys. Rev. B*, vol. 3, no. 4, pp. 1215–1223, Feb. 1971, doi: 10.1103/PhysRevB.3.1215.
- [54] Y. Luo *et al.*, ‘Electron work function: an indicative parameter towards a novel material design methodology’, *Sci Rep*, vol. 11, no. 1, p. 11565, Jun. 2021, doi: 10.1038/s41598-021-90715-4.
- [55] S. Morishita, J. Yamasaki, and N. Tanaka, ‘Measurement of spatial coherence of electron beams by using a small selected-area aperture’, *Ultramicroscopy*, vol. 129, pp. 10–17, Jun. 2013, doi: 10.1016/j.ultramic.2013.02.019.
- [56] ‘thermionic-emission gun | Glossary | JEOL Ltd.’, thermionic-emission gun | Glossary | JEOL Ltd. Accessed: Aug. 14, 2024. [Online]. Available: <https://www.jeol.com/>
- [57] ‘cold ((cathode) field-emission electron gun | Glossary | JEOL Ltd.’, cold ((cathode) field-emission electron gun | Glossary | JEOL Ltd. Accessed: Aug. 14, 2024. [Online]. Available: <https://www.jeol.com/>
- [58] R. M. Glaeser, ‘Specimen Behavior in the Electron Beam’, in *Methods in Enzymology*, vol. 579, Elsevier, 2016, pp. 19–50. doi: 10.1016/bs.mie.2016.04.010.
- [59] T. Grant and N. Grigorieff, ‘Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6’, *eLife*, vol. 4, p. e06980, May 2015, doi: 10.7554/eLife.06980.
- [60] S. H. Scheres, ‘Beam-induced motion correction for sub-megadalton cryo-EM particles’, *eLife*, vol. 3, p. e03665, Aug. 2014, doi: 10.7554/eLife.03665.
- [61] M. J. Peet, R. Henderson, and C. J. Russo, ‘The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules’, *Ultramicroscopy*, vol. 203, pp. 125–131, Aug. 2019, doi: 10.1016/j.ultramic.2019.02.007.
- [62] G. A. Meeke, *Practical Electron Microscopy for Biologists*. Wiley, 1976.
- [63] F. Zernike, ‘Phase contrast, a new method for the microscopic observation of transparent objects’, *Physica*, vol. 9, no. 7, pp. 686–698, Jul. 1942, doi: 10.1016/S0031-8914(42)80035-X.
- [64] ‘Zernike Phase Contrast Cryo-Electron Microscopy and Tomography for Structure Determination at Nanometer and Sub-Nanometer Resolutions - PMC’. Accessed: Sep. 04, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2925294/>
- [65] E. Majorovits, B. Barton, K. Schultheiß, F. Pérez-Willard, D. Gerthsen, and R. R. Schröder, ‘Optimizing phase contrast in transmission electron microscopy with an electrostatic ((Boersch) phase plate’, *Ultramicroscopy*, vol. 107, no. 2, pp. 213–226, Feb. 2007, doi: 10.1016/j.ultramic.2006.07.006.
- [66] R. H. Wade, ‘A brief look at imaging and contrast transfer’, *Ultramicroscopy*, vol. 46, no. 1, pp. 145–156, Oct. 1992, doi: 10.1016/0304-3991(92)90011-8.
- [67] G. Zanetti, J. D. Riches, S. D. Fuller, and J. A. G. Briggs, ‘Contrast transfer function correction applied to cryo-electron tomography and sub-tomogram averaging’, *Journal of Structural Biology*, vol. 168, no. 2, pp. 305–312, Nov. 2009, doi: 10.1016/j.jsb.2009.08.002.
- [68] P. A. Thuman-Commike and W. Chiu, ‘Reconstruction principles of icosahedral virus structure determination using electron cryomicroscopy’, *Micron*, vol. 31, no. 6, pp. 687–711, Dec. 2000, doi: 10.1016/S0968-4328(99)00077-3.
- [69] Y. Cong and S. J. Ludtke, ‘Chapter Eight - Single Particle Analysis at High Resolution’, in *Methods in Enzymology*, vol. 482, G. J. Jensen, Ed., in Cryo-EM, Part B: 3-D Reconstruction, vol. 482. , Academic Press, 2010, pp. 211–235. doi: 10.1016/S0076-6879(10)82009-9.
- [70] J. Zivanov, T. Nakane, and S. H. W. Scheres, ‘Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1’, *IUCrJ*, vol. 7, no. Pt 2, pp. 253–267, Feb. 2020, doi: 10.1107/S2052252520000081.

- [71] L. N. Thibos, A. Bradley, D. L. Still, X. Zhang, and P. A. Howarth, 'Theory and measurement of ocular chromatic aberration', *Vision Research*, vol. 30, no. 1, pp. 33–49, Jan. 1990, doi: 10.1016/0042-6989(90)90126-6.
- [72] R. G. Efremov and A. Stroobants, 'Coma-corrected rapid single-particle cryo-EM data collection on the CRYO ARM 300', *Acta Crystallogr D Struct Biol*, vol. 77, no. Pt 5, pp. 555–564, Apr. 2021, doi: 10.1107/S2059798321002151.
- [73] X. Zhang and Z. H. Zhou, 'Limiting factors in atomic resolution cryo electron microscopy: No simple tricks', *J Struct Biol*, vol. 175, no. 3, pp. 253–263, Sep. 2011, doi: 10.1016/j.jsb.2011.05.004.
- [74] Y. Cheng, N. Grigorieff, P. A. Penczek, and T. Walz, 'A Primer to Single-Particle Cryo-Electron Microscopy', *Cell*, vol. 161, no. 3, pp. 438–449, Apr. 2015, doi: 10.1016/j.cell.2015.03.050.
- [75] A. Rohou and N. Grigorieff, 'CTFFIND4: Fast and accurate defocus estimation from electron micrographs', *Journal of Structural Biology*, vol. 192, no. 2, pp. 216–221, Nov. 2015, doi: 10.1016/j.jsb.2015.08.008.
- [76] R. S. Ruskin, Z. Yu, and N. Grigorieff, 'Quantitative characterization of electron detectors for transmission electron microscopy', *J Struct Biol*, vol. 184, no. 3, p. 10.1016/j.jsb.2013.10.016, Dec. 2013, doi: 10.1016/j.jsb.2013.10.016.
- [77] G. McMullan, S. Chen, R. Henderson, and A. R. Faruqi, 'Detective quantum efficiency of electron area detectors in electron microscopy', *Ultramicroscopy*, vol. 109, no. 9, pp. 1126–1143, Aug. 2009, doi: 10.1016/j.ultramic.2009.04.002.
- [78] S. Wu, J.-P. Armache, and Y. Cheng, 'Single-particle cryo-EM data acquisition by using direct electron detection camera', *Microscopy((Oxford, England))*, vol. 65, no. 1, pp. 35–41, Feb. 2016, doi: 10.1093/jmicro/dfv355.
- [79] G. McMullan *et al.*, 'Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector', *Ultramicroscopy*, vol. 109, no. 9–3, pp. 1144–1147, Aug. 2009, doi: 10.1016/j.ultramic.2009.05.005.
- [80] G. McMullan, A. R. Faruqi, D. Clare, and R. Henderson, 'Comparison of optimal performance at 300 keV of three direct electron detectors for use in low dose electron microscopy', *Ultramicroscopy*, vol. 147, pp. 156–163, Dec. 2014, doi: 10.1016/j.ultramic.2014.08.002.
- [81] J. H. Mendez, A. Mehrani, P. Randolph, and S. Stagg, 'Throughput and resolution with a next-generation direct electron detector', *IUCrJ*, vol. 6, no. Pt 6, pp. 1007–1013, Oct. 2019, doi: 10.1107/S2052252519012661.
- [82] 'Improving DQE with Counting and Super-Resolution | Gatan, Inc.' Accessed: Sep. 04, 2024. [Online]. Available: <https://www.gatan.com/improving-dqe-counting-and-super-resolution>
- [83] D. J. De Rosier and A. Klug, 'Reconstruction of Three Dimensional Structures from Electron Micrographs', *Nature*, vol. 217, no. 5124, pp. 130–134, Jan. 1968, doi: 10.1038/217130a0.
- [84] S. Niebling *et al.*, 'Biophysical Screening Pipeline for Cryo-EM Grid Preparation of Membrane Proteins', *Front Mol Biosci*, vol. 9, p. 882288, Jun. 2022, doi: 10.3389/fmolb.2022.882288.
- [85] K. Neselu, B. Wang, W. J. Rice, C. S. Potter, B. Carragher, and E. Y. D. Chua, 'Measuring the effects of ice thickness on resolution in single particle cryo-EM', *J Struct Biol X*, vol. 7, p. 100085, Jan. 2023, doi: 10.1016/j.yjsbx.2023.100085.
- [86] L. A. Passmore and C. J. Russo, 'Specimen Preparation for High-Resolution Cryo-EM', *Methods in Enzymology*, vol. 579, pp. 51–86, 2016, doi: 10.1016/bs.mie.2016.04.011.
- [87] G. G. Sgro and T. R. D. Costa, 'Cryo-EM Grid Preparation of Membrane Protein Samples for Single Particle Analysis', *Front Mol Biosci*, vol. 5, p. 74, Jul. 2018, doi: 10.3389/fmolb.2018.00074.

- [88] C. J. Russo and L. A. Passmore, ‘Ultrastable gold substrates: Properties of a support for high-resolution electron cryomicroscopy of biological specimens’, *Journal of Structural Biology*, vol. 193, no. 1, pp. 33–44, Jan. 2016, doi: 10.1016/j.jsb.2015.11.006.
- [89] K. Naydenova, M. J. Peet, and C. J. Russo, ‘Multifunctional graphene supports for electron cryomicroscopy’, *Proceedings of the National Academy of Sciences*, vol. 116, no. 24, pp. 11718–11724, Jun. 2019, doi: 10.1073/pnas.1904766116.
- [90] Y. Han *et al.*, ‘High-yield monolayer graphene grids for near-atomic resolution cryoelectron microscopy’, *Proceedings of the National Academy of Sciences*, vol. 117, no. 2, pp. 1009–1014, Jan. 2020, doi: 10.1073/pnas.1919114117.
- [91] K. Naydenova and C. J. Russo, ‘Integrated wafer-scale manufacturing of electron cryomicroscopy specimen supports’, *Ultramicroscopy*, vol. 232, p. 113396, Jan. 2022, doi: 10.1016/j.ultramic.2021.113396.
- [92] R. A. Grassucci, D. J. Taylor, and J. Frank, ‘Preparation of macromolecular complexes for cryo-electron microscopy’, *Nat Protoc*, vol. 2, no. 12, pp. 3239–3246, 2007, doi: 10.1038/nprot.2007.452.
- [93] S. T. Huber, E. Sarajlic, R. Huijink, F. Weis, W. H. Evers, and A. J. Jakobi, ‘Nanofluidic chips for cryo-EM structure determination from picoliter sample volumes’, *eLife*, vol. 11, p. e72629, Jan. 2022, doi: 10.7554/eLife.72629.
- [94] J. L. Rubinstein *et al.*, ‘Shake-it-off: a simple ultrasonic cryo-EM specimen-preparation device’, *Acta Crystallographica Section D: Structural Biology*, vol. 75, no. 12, pp. 1063–1070, Dec. 2019, doi: 10.1107/S2059798319014372.
- [95] W. F. Tivol, A. Briegel, and G. J. Jensen, ‘An Improved Cryogen for Plunge Freezing’, *Microsc Microanal*, vol. 14, no. 5, pp. 375–379, Oct. 2008, doi: 10.1017/S1431927608080781.
- [96] M. J. Dobro, L. A. Melanson, G. J. Jensen, and A. W. McDowall, ‘Chapter Three - Plunge Freezing for Electron Cryomicroscopy’, in *Methods in Enzymology*, vol. 481, G. J. Jensen, Ed., in *Cryo-EM Part A Sample Preparation and Data Collection*, vol. 481, Academic Press, 2010, pp. 63–82. doi: 10.1016/S0076-6879(10)81003-1.
- [97] T. Engstrom *et al.*, ‘High-resolution single-particle cryo-EM of samples vitrified in boiling nitro-gen’, *IUCrJ*, vol. 8, no. Pt 6, pp. 867–877, Sep. 2021, doi: 10.1107/S2052252521008095.
- [98] T. Jain, P. Sheehan, J. Crum, B. Carragher, and C. S. Potter, ‘Spotiton: A prototype for an integrated inkjet dispense and vitrification system for cryo-TEM’, *Journal of Structural Biology*, vol. 179, no. 1, pp. 68–75, Jul. 2012, doi: 10.1016/j.jsb.2012.04.020.
- [99] M. C. Darrow, J. P. Moore, R. J. Walker, K. Doering, and R. S. King, ‘Chameleon: Next Generation Sample Preparation for CryoEM based on Spotiton’, *Microsc Microanal*, vol. 25, no. S2, pp. 994–995, Aug. 2019, doi: 10.1017/S1431927619005701.
- [100] H. Wei *et al.*, ‘Optimizing “self-wicking” nanowire grids’, *Journal of Structural Biology*, vol. 202, no. 2, pp. 170–174, May 2018, doi: 10.1016/j.jsb.2018.01.001.
- [101] V. P. Dandey *et al.*, ‘Time-resolved cryo-EM using Spotiton’, *Nature Methods*, vol. 17, no. 9, pp. 897–900, Sep. 2020, doi: 10.1038/s41592-020-0925-6.
- [102] S. A. Arnold *et al.*, ‘Blotting-free and lossless cryo-electron microscopy grid preparation from nanoliter-sized protein samples and single-cell extracts’, *Journal of Structural Biology*, vol. 197, no. 3, pp. 220–226, Mar. 2017, doi: 10.1016/j.jsb.2016.11.002.
- [103] B. Carragher *et al.*, ‘CURRENT OUTCOMES WHEN OPTIMIZING “STANDARD” SAMPLE PREPARATION FOR SINGLE-PARTICLE CRYO-EM’, *J Microsc*, vol. 276, no. 1, pp. 39–45, Oct. 2019, doi: 10.1111/jmi.12834.

- [104] I. Drulyte *et al.*, ‘Approaches to altering particle distributions in cryo-electron microscopy sample preparation’, *Acta Crystallographica Section D: Structural Biology*, vol. 74, no. 6, pp. 560–571, Jun. 2018, doi: 10.1107/S2059798318006496.
- [105] A. J. Noble *et al.*, ‘Routine single particle CryoEM sample and grid characterization by tomography’, *eLife*, vol. 7, p. e34257, May 2018, doi: 10.7554/eLife.34257.
- [106] Y. Liu, X. Meng, and Z. Liu, ‘Deformed grids for single-particle cryo-electron microscopy of specimens exhibiting a preferred orientation’, *Journal of Structural Biology*, vol. 182, no. 3, pp. 255–258, Jun. 2013, doi: 10.1016/j.jsb.2013.03.005.
- [107] C. Suloway *et al.*, ‘Automated molecular microscopy: The new Legimon system’, *Journal of Structural Biology*, vol. 151, no. 1, pp. 41–60, Jul. 2005, doi: 10.1016/j.jsb.2005.03.010.
- [108] X. Li, S. Zheng, D. A. Agard, and Y. Cheng, ‘Asynchronous data acquisition and on-the-fly analysis of dose fractionated cryoEM images by UCSFImage’, *Journal of Structural Biology*, vol. 192, no. 2, pp. 174–178, Nov. 2015, doi: 10.1016/j.jsb.2015.09.003.
- [109] W. J. Rice *et al.*, ‘Routine Determination of Ice Thickness for Cryo-EM Grids’, *Journal of structural biology*, vol. 204, no. 1, pp. 38–44, Oct. 2018, doi: 10.1016/j.jsb.2018.06.007.
- [110] R. J. M. Henderikx *et al.*, ‘VitroJet: new features and case studies’, *Acta Cryst D*, vol. 80, no. 4, Art. no. 4, Apr. 2024, doi: 10.1107/S2059798324001852.
- [111] H. G. Brown and E. Hanssen, ‘MeasureIce: accessible on-the-fly measurement of ice thickness in cryo-electron microscopy’, *Commun Biol*, vol. 5, no. 1, pp. 1–9, Aug. 2022, doi: 10.1038/s42003-022-03698-x.
- [112] L. A. Baker and J. L. Rubinstein, ‘Chapter Fifteen - Radiation Damage in Electron Cryomicroscopy’, in *Methods in Enzymology*, vol. 481, G. J. Jensen, Ed., in Cryo-EM Part A Sample Preparation and Data Collection, vol. 481., Academic Press, 2010, pp. 371–388. doi: 10.1016/S0076-6879(10)81015-8.
- [113] R. M. Glaeser, ‘Limitations to significant information in biological electron microscopy as a result of radiation damage’, *Journal of Ultrastructure Research*, vol. 36, no. 3, pp. 466–482, Aug. 1971, doi: 10.1016/S0022-5320(71)80118-1.
- [114] J. F. Conway, B. L. Trus, F. P. Booy, W. W. Newcomb, J. C. Brown, and A. C. Steven, ‘The Effects of Radiation Damage on the Structure of Frozen Hydrated HSV-1 Capsids’, *Journal of Structural Biology*, vol. 111, no. 3, pp. 222–233, Nov. 1993, doi: 10.1006/jsbi.1993.1052.
- [115] H.-G. Heide, ‘Observations on ice layers’, *Ultramicroscopy*, vol. 14, no. 3, pp. 271–278, Jan. 1984, doi: 10.1016/0304-3991(84)90095-0.
- [116] E. R. Wright, C. V. Iancu, W. F. Tivol, and G. J. Jensen, ‘Observations on the behavior of vitreous ice at ~82 and ~12 K’, *Journal of Structural Biology*, vol. 153, no. 3, pp. 241–252, Mar. 2006, doi: 10.1016/j.jsb.2005.12.003.
- [117] ‘Chapter 4: Cryo-EM Data Collection’, Cryo EM 101. Accessed: Sep. 26, 2022. [Online]. Available: <https://cryoem101.org/chapter-4/>
- [118] ‘Fast and accurate defocus modulation for improved tunability of cryo-EM experiments’, *IUCrJ*, vol. 7, pp. 566–574, Jan. 2020, doi: 10.1107/S205225252000408X.
- [119] R. Danev, B. Buijsse, M. Khoshouei, J. M. Plitzko, and W. Baumeister, ‘Volta potential phase plate for in-focus phase contrast transmission electron microscopy’, *Proceedings of the National Academy of Sciences*, vol. 111, no. 44, pp. 15635–15640, Nov. 2014, doi: 10.1073/pnas.1418377111.
- [120] O. Schwartz, J. J. Axelrod, S. L. Campbell, C. Turnbaugh, R. M. Glaeser, and H. Müller, ‘Laser phase plate for transmission electron microscopy’, *Nat Methods*, vol. 16, no. 10, Art. no. 10, Oct. 2019, doi: 10.1038/s41592-019-0552-2.

- [121] R. Henderson and R. M. Glaeser, ‘Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals’, *Ultramicroscopy*, vol. 16, no. 2, pp. 139–150, Jan. 1985, doi: 10.1016/0304-3991(85)90069-5.
- [122] P. B. Rosenthal and R. Henderson, ‘Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy’, *Journal of Molecular Biology*, vol. 333, no. 4, pp. 721–745, Oct. 2003, doi: 10.1016/j.jmb.2003.07.013.
- [123] X. Li *et al.*, ‘Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM’, *Nature Methods*, vol. 10, no. 6, pp. 584–590, Jun. 2013, doi: 10.1038/nmeth.2472.
- [124] S. Q. Zheng, E. Palovcak, J.-P. Armache, K. A. Verba, Y. Cheng, and D. A. Agard, ‘MotionCor2 - anisotropic correction of beam-induced motion for improved cryo-electron microscopy’, *Nature methods*, vol. 14, no. 4, pp. 331–332, Apr. 2017, doi: 10.1038/nmeth.4193.
- [125] S. Zheng, ‘MotionCor2 User Manual’. May 31, 2022.
- [126] J. L. Rubinstein and M. A. Brubaker, ‘Alignment of cryo-EM movies of individual particles by optimization of image translations’, *Journal of Structural Biology*, vol. 192, no. 2, pp. 188–195, Nov. 2015, doi: 10.1016/j.jsb.2015.08.007.
- [127] J. Zivanov, T. Nakane, and S. H. W. Scheres, ‘A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis’, *IUCrJ*, vol. 6, no. 1, pp. 5–17, Jan. 2019, doi: 10.1107/S205225251801463X.
- [128] K. Zhang, ‘Gctf: Real-time CTF determination and correction’, *J Struct Biol*, vol. 193, no. 1, pp. 1–12, Jan. 2016, doi: 10.1016/j.jsb.2015.11.003.
- [129] P. A. Penczek, ‘Image Restoration in Cryo-electron Microscopy’, *Methods Enzymol*, vol. 482, pp. 35–72, 2010, doi: 10.1016/S0076-6879(10)82002-6.
- [130] J. L. Dickerson, P.-H. Lu, D. Hristov, R. E. Dunin-Borkowski, and C. J. Russo, ‘Imaging biological macromolecules in thick specimens: The role of inelastic scattering in cryoEM’, *Ultramicroscopy*, vol. 237, p. 113510, Jul. 2022, doi: 10.1016/j.ultramic.2022.113510.
- [131] A. Wlodawer, M. Li, and Z. Dauter, ‘High-resolution cryo-EM maps and models – a crystallographer’s perspective’, *Structure*, vol. 25, no. 10, pp. 1589-1597.e1, Oct. 2017, doi: 10.1016/j.str.2017.07.012.
- [132] X. Bai *et al.*, ‘An atomic structure of human γ -secretase’, *Nature*, vol. 525, no. 7568, pp. 212–217, Sep. 2015, doi: 10.1038/nature14892.
- [133] ‘OpenCV: Laplace Operator’. Accessed: Aug. 23, 2022. [Online]. Available: https://docs.opencv.org/4.x/d5/db5/tutorial_laplace_operator.html
- [134] T. Grant, A. Rohou, and N. Grigorieff, ‘cisTEM, user-friendly software for single-particle image processing’, *eLife*, vol. 7, p. e35383, Mar. 2018, doi: 10.7554/eLife.35383.
- [135] A. M. Roseman, ‘FindEM—a fast, efficient program for automatic selection of particles from electron micrographs’, *Journal of Structural Biology*, vol. 145, no. 1, pp. 91–99, Jan. 2004, doi: 10.1016/j.jsb.2003.11.007.
- [136] S. H. W. Scheres, ‘Semi-automated selection of cryo-EM particles in RELION-1.3’, *Journal of Structural Biology*, vol. 189, no. 2, pp. 114–122, Feb. 2015, doi: 10.1016/j.jsb.2014.11.010.
- [137] M. Shatsky, R. J. Hall, S. E. Brenner, and R. M. Glaeser, ‘A method for the alignment of heterogeneous macromolecules from electron microscopy’, *Journal of Structural Biology*, vol. 166, no. 1, pp. 67–78, Apr. 2009, doi: 10.1016/j.jsb.2008.12.008.
- [138] R. Henderson, ‘Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise’, *Proceedings of the National Academy of Sciences*, vol. 110, no. 45, pp. 18037–18041, Nov. 2013, doi: 10.1073/pnas.1314449110.

- [139] T. Bepler, K. Kelley, A. J. Noble, and B. Berger, ‘Topaz-Denoise: general deep denoising models for cryoEM and cryoET’, *Nature Communications*, vol. 11, no. 1, p. 5208, Oct. 2020, doi: 10.1038/s41467-020-18952-1.
- [140] T. Bepler *et al.*, ‘Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs’, *Nat Methods*, vol. 16, no. 11, Art. no. 11, Nov. 2019, doi: 10.1038/s41592-019-0575-8.
- [141] T. Wagner *et al.*, ‘SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM’, *Commun Biol*, vol. 2, no. 1, Art. no. 1, Jun. 2019, doi: 10.1038/s42003-019-0437-z.
- [142] J. Lehtinen *et al.*, ‘Noise2Noise: Learning Image Restoration without Clean Data’, Oct. 29, 2018, *arXiv*: arXiv:1803.04189. Accessed: Sep. 04, 2024. [Online]. Available: <http://arxiv.org/abs/1803.04189>
- [143] ‘janni tutorial []’. Accessed: Sep. 04, 2024. [Online]. Available: https://sphire.mpg.de/wiki/doku.php?id=janni_tutorial
- [144] T. Wagner and S. Raunser, ‘The evolution of SPHIRE-crYOLO particle picking and its application in automated cryo-EM processing workflows’, *Communications Biology*, vol. 3, no. 1, pp. 1–5, Feb. 2020, doi: 10.1038/s42003-020-0790-y.
- [145] F. Wang *et al.*, ‘DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM’, *Journal of Structural Biology*, vol. 195, no. 3, pp. 325–336, Sep. 2016, doi: 10.1016/j.jsb.2016.07.006.
- [146] Y. Zhu, Q. Ouyang, and Y. Mao, ‘A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy’, *BMC Bioinformatics*, vol. 18, no. 1, p. 348, Jul. 2017, doi: 10.1186/s12859-017-1757-y.
- [147] S. H. W. Scheres, ‘Maximum-likelihood methods in cryo-EM. Part II: application to experimental data’, *Methods Enzymol*, vol. 482, pp. 295–320, 2010, doi: 10.1016/S0076-6879(10)82012-9.
- [148] ‘Reference-free 2D class averaging — RELION documentation’. Accessed: Aug. 24, 2022. [Online]. Available: https://relion.readthedocs.io/en/latest/SPA_tutorial/Class2D.html#
- [149] P. A. Penczek, ‘Fundamentals of three-dimensional reconstruction from projections’, *Methods Enzymol*, vol. 482, pp. 1–33, 2010, doi: 10.1016/S0076-6879(10)82001-4.
- [150] M. Radermacher, T. Wagenknecht, A. Verschoor, and J. Frank, ‘Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*’, *J Microsc*, vol. 146, no. Pt 2, pp. 113–136, May 1987, doi: 10.1111/j.1365-2818.1987.tb01333.x.
- [151] P. A. Penczek, J. Zhu, and J. Frank, ‘A common-lines based method for determining orientations for $N > 3$ particle projections simultaneously’, *Ultramicroscopy*, vol. 63, no. 3–4, pp. 205–218, Jul. 1996, doi: 10.1016/0304-3991(96)00037-x.
- [152] E. Levin, T. Bendory, N. Boumal, J. Kileel, and A. Singer, ‘3D ab initio modeling in cryo-EM by autocorrelation analysis’, in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Apr. 2018, pp. 1569–1573. doi: 10.1109/ISBI.2018.8363873.
- [153] D. Kimanius, L. Dong, G. Sharov, T. Nakane, and S. H. W. Scheres, ‘New tools for automated cryo-EM single-particle analysis in RELION-4.0’, *Biochem J*, vol. 478, no. 24, pp. 4169–4185, Dec. 2021, doi: 10.1042/BCJ20210708.
- [154] E. Nogales and S. H. W. Scheres, ‘Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity’, *Molecular Cell*, vol. 58, no. 4, pp. 677–689, May 2015, doi: 10.1016/j.molcel.2015.02.019.

- [155] T. Nakane, D. Kimanius, E. Lindahl, and S. H. Scheres, ‘Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION’, *eLife*, vol. 7, p. e36861, Jun. 2018, doi: 10.7554/eLife.36861.
- [156] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali, ‘Protein Structure Fitting and Refinement Guided by Cryo-EM Density’, *Structure*, vol. 16, no. 2, pp. 295–307, Feb. 2008, doi: 10.1016/j.str.2007.11.016.
- [157] A. P. Joseph *et al.*, ‘Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment’, *Methods((San Diego, Calif.)*, vol. 100, pp. 42–49, May 2016, doi: 10.1016/j.ymeth.2016.03.007.
- [158] E. D. Zhong, T. Bepler, B. Berger, and J. H. Davis, ‘CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks’, *Nat Methods*, vol. 18, no. 2, Art. no. 2, Feb. 2021, doi: 10.1038/s41592-020-01049-4.
- [159] E. Villa and K. Lasker, ‘Finding the right fit: chiseling structures out of cryo-electron microscopy maps’, *Current Opinion in Structural Biology*, vol. 25, pp. 118–125, Apr. 2014, doi: 10.1016/j.sbi.2014.04.001.
- [160] G. HARAUZ and M. VAN HEEL, ‘Exact filters for general geometry three dimensional reconstruction’, *Optik ((Stuttg.)*, vol. 73, no. 4, pp. 146–156, 1986.
- [161] ‘Local-resolution estimation — RELION documentation’. Accessed: Aug. 31, 2022. [Online]. Available: https://relion.readthedocs.io/en/release-3.1/SPA_tutorial/Validation.html
- [162] E. F. Pettersen *et al.*, ‘UCSF Chimera—a visualization system for exploratory research and analysis’, *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/jcc.20084.
- [163] A. Kucukelbir, F. J. Sigworth, and H. D. Tagare, ‘Quantifying the local resolution of cryo-EM density maps’, *Nature Methods*, vol. 11, no. 1, pp. 63–65, Jan. 2014, doi: 10.1038/nmeth.2727.
- [164] T. Kato, N. Terahara, and K. Namba, ‘EMPIAR-10204 dataset’, Aug. 2018.
- [165] A. J. Jakobi, M. Wilmanns, and C. Sachse, ‘Model-based local density sharpening of cryo-EM maps’, *eLife*, vol. 6, p. e27131, Oct. 2017, doi: 10.7554/eLife.27131.
- [166] K. Cowtan, ‘Cowtan, K. The Buccaneer software for automated model building. Acta Crystallogr. D 62, 1002-1011’, *Acta crystallographica. Section D, Biological crystallography*, vol. 62, pp. 1002–11, Oct. 2006, doi: 10.1107/S0907444906022116.
- [167] G. G. Langer, S. X. Cohen, V. S. Lamzin, and A. Perrakis, ‘Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7’, *Nat Protoc*, vol. 3, no. 7, pp. 1171–1179, 2008, doi: 10.1038/nprot.2008.91.
- [168] T. C. Terwilliger *et al.*, ‘Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard’, *Acta Crystallogr D Biol Crystallogr*, vol. 64, no. Pt 1, pp. 61–69, Jan. 2008, doi: 10.1107/S090744490705024X.
- [169] K. Cowtan, ‘Fast Fourier feature recognition’, *Acta Cryst D*, vol. 57, no. 10, pp. 1435–1444, Oct. 2001, doi: 10.1107/S0907444901010812.
- [170] G. N. Murshudov *et al.*, ‘REFMAC5 for the refinement of macromolecular crystal structures’, *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. Pt 4, pp. 355–367, Apr. 2011, doi: 10.1107/S0907444911001314.
- [171] J. Jumper *et al.*, ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
- [172] K. Jamali, L. Käll, R. Zhang, A. Brown, D. Kimanius, and S. H. W. Scheres, ‘Automated model building and protein identification in cryo-EM maps’, *Nature*, vol. 628, no. 8007, pp. 450–457, Apr. 2024, doi: 10.1038/s41586-024-07215-4.

- [173] C. J. Williams *et al.*, ‘MolProbity: More and better reference data for improved all-atom structure validation’, *Protein Science: A Publication of the Protein Society*, vol. 27, no. 1, pp. 293–315, 2018, doi: 10.1002/pro.3330.
- [174] M. G. Prisant, C. J. Williams, V. B. Chen, J. S. Richardson, and D. C. Richardson, ‘New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink “waters,” and NGL Viewer to recapture online 3D graphics’, *Protein Science*, vol. 29, no. 1, pp. 315–329, 2020, doi: <https://doi.org/10.1002/pro.3786>.
- [175] E. Ramírez-Aportela *et al.*, ‘FSC-Q: a CryoEM map-to-atomic model quality validation based on the local Fourier shell correlation’, *Nature Communications*, vol. 12, no. 1, p. 42, Dec. 2021, doi: 10.1038/s41467-020-20295-w.
- [176] G. Pintilie, ‘Measurement of Atom Resolvability in CryoEM Maps with Q-scores’, *Microscopy and Microanalysis*, vol. 26, no. S2, pp. 2316–2316, Aug. 2020, doi: 10.1017/S1431927620021170.
- [177] S. Malhotra, A. P. Joseph, J. Thiyagalingam, and M. Topf, ‘Assessment of protein–protein interfaces in cryo-EM derived assemblies’, *Nat Commun*, vol. 12, no. 1, p. 3399, Jun. 2021, doi: 10.1038/s41467-021-23692-x.
- [178] C. L. Lawson *et al.*, ‘Outcomes of the 2019 EMDDataResource model challenge: validation of cryo-EM models at near-atomic resolution’, *bioRxiv*, p. 2020.06.12.147033, Jan. 2020, doi: 10.1101/2020.06.12.147033.
- [179] C. L. Lawson *et al.*, ‘Cryo-EM model validation recommendations based on outcomes of the 2019 EMDDataResource challenge’, *Nature Methods*, vol. 18, no. 2, pp. 156–164, Feb. 2021, doi: 10.1038/s41592-020-01051-w.
- [180] R. M. Glaeser, ‘PROTEINS, INTERFACES, AND CRYO-EM GRIDS’, *Curr Opin Colloid Interface Sci*, vol. 34, pp. 1–8, Mar. 2018, doi: 10.1016/j.cocis.2017.12.009.
- [181] W. Baumann and L. Reimer, ‘Comparison of the noise of different electron detection systems using a scintillator-photomultiplier combination’, *Scanning*, vol. 4, no. 3, pp. 141–151, 1981, doi: 10.1002/sca.4950040304.
- [182] A. S. Frangakis, ‘It’s noisy out there! A review of denoising techniques in cryo-electron tomography’, *Journal of Structural Biology*, vol. 213, no. 4, p. 107804, Dec. 2021, doi: 10.1016/j.jsb.2021.107804.
- [183] ‘Measuring the effects of ice thickness on resolution in single particle cryo-EM - ScienceDirect’. Accessed: Mar. 11, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2590152423000016?via%3Dihub>
- [184] M. Olek, K. Cowtan, D. Webb, Y. Chaban, and P. Zhang, ‘IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice’, *Structure*, vol. 30, no. 4, pp. 522–531.e4, Apr. 2022, doi: 10.1016/j.str.2022.01.005.
- [185] S. W. Hoh, T. Burnley, and K. Cowtan, ‘Current approaches for automated model building into cryo-EM maps using Buccaneer with CCP-EM’, *Acta Cryst D*, vol. 76, no. 6, Art. no. 6, Jun. 2020, doi: 10.1107/S2059798320005513.
- [186] D. Liebschner *et al.*, ‘Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix’, *Acta Crystallogr D Struct Biol*, vol. 75, no. 10, pp. 861–877, Oct. 2019, doi: 10.1107/S2059798319011471.
- [187] T. C. Terwilliger, P. D. Adams, P. V. Afonine, and O. V. Sobolev, ‘Cryo-EM map interpretation and protein model-building using iterative map segmentation’, *Protein Science*, vol. 29, no. 1, pp. 87–99, 2020, doi: <https://doi.org/10.1002/pro.3740>.
- [188] J. Pfab, N. M. Phan, and D. Si, ‘DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes’, *Proceedings of the National Academy of Sciences*, vol. 118, no. 2, Jan. 2021, doi: 10.1073/pnas.2017525118.

- [189] G. Reggiano, W. Lugmayr, D. Farrell, T. C. Marlovits, and F. DiMaio, ‘Residue-level error detection in cryo-electron microscopy models’, *Structure*, vol. 31, no. 7, pp. 860–869.e4, Jul. 2023, doi: 10.1016/j.str.2023.05.002.
- [190] G. Chojnowski, ‘Sequence-assignment validation in cryo-EM models with checkMySequence’, *Acta Cryst D*, vol. 78, no. 7, pp. 806–816, Jul. 2022, doi: 10.1107/S2059798322005009.
- [191] V. B. Chen *et al.*, ‘MolProbity: all-atom structure validation for macromolecular crystallography’, *Acta Cryst D*, vol. 66, no. 1, Art. no. 1, Jan. 2010, doi: 10.1107/S0907444909042073.
- [192] P. V. Afonine *et al.*, ‘New tools for the analysis and validation of cryo-EM maps and atomic models’, *Acta Crystallographica Section D: Structural Biology*, vol. 74, no. 9, pp. 814–840, Sep. 2018, doi: 10.1107/S2059798318009324.
- [193] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, ‘Features and development of *Coot*’, *Acta Crystallographica Section D Biological Crystallography*, vol. 66, no. 4, pp. 486–501, Apr. 2010, doi: 10.1107/S0907444910007493.
- [194] I. Lagerstedt *et al.*, ‘Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB’, *Journal of Structural Biology*, vol. 184, no. 2, pp. 173–181, Nov. 2013, doi: 10.1016/j.jsb.2013.09.021.
- [195] N. Grigorieff, ‘Resolution measurement in structures derived from single particles’, *Acta Cryst D*, vol. 56, no. 10, Art. no. 10, Oct. 2000, doi: 10.1107/S0907444900009549.
- [196] S. H. W. Scheres and S. Chen, ‘Prevention of overfitting in cryo-EM structure determination’, *Nat Methods*, vol. 9, no. 9, pp. 853–854, Sep. 2012, doi: 10.1038/nmeth.2115.
- [197] W. A. Havelka, R. Henderson, and D. Oesterhelt, ‘Three-dimensional structure of halorhodopsin at 7 Å resolution’, *Journal of Molecular Biology*, vol. 247, no. 4, pp. 726–738, Apr. 1995, doi: 10.1016/S0022-2836(05)80151-2.
- [198] S. Kaur *et al.*, ‘Local computational methods to improve the interpretability and analysis of cryo-EM maps’, *Nat Commun*, vol. 12, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41467-021-21509-5.
- [199] M. Beckers, A. J. Jakobi, and C. Sachse, ‘Thresholding of cryo-EM density maps by false discovery rate control’, *IUCrJ*, vol. 6, no. 1, pp. 18–33, Jan. 2019, doi: 10.1107/S2052252518014434.
- [200] Y. Benjamini and D. Yekutieli, ‘The control of the false discovery rate in multiple testing under dependency’, *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, Aug. 2001, doi: 10.1214/aos/1013699998.
- [201] ‘Confidence maps: statistical inference of cryo-EM maps - PMC’. Accessed: Mar. 11, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7137106/>
- [202] M. Wojdyr, ‘project-gemmi/gemmi’. GEMMI, 2017. Accessed: Dec. 09, 2020. [Online]. Available: <https://github.com/project-gemmi/gemmi>
- [203] ‘molmap’. Accessed: Sep. 04, 2024. [Online]. Available: <https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/molmap.html>
- [204] M. Beckers, C. M. Palmer, and C. Sachse, ‘Confidence maps: statistical inference of cryo-EM maps’, *Acta Crystallogr D Struct Biol*, vol. 76, no. Pt 4, pp. 332–339, Mar. 2020, doi: 10.1107/S2059798320002995.


Appendix A

University of York
York Graduate Research School
Research Degree Thesis Statement of Authorship

Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

Candidate name	Mateusz Olek
Department	Chemistry
Thesis title	Improved methods for the annotation and processing of Cryo-EM images and for atomic model validation


Title of the work (paper/chapter)	IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice	
Publication status	Published	x
	Accepted for publication	
	Submitted for publication	
	Unpublished and unsubmitted	
Citation details (if applicable)	Olek M, Cowtan K, Webb D, Chaban Y, Zhang P. IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice. Structure. 2022 Apr 7;30(4):522-531.e4. doi: 10.1016/j.str.2022.01.005. Epub 2022 Feb 11. PMID: 35150604; PMCID: PMC9033277.	


Description of the candidate's contribution to the work*	Conceived and designed research, developed the IceBreaker software tool, processed and analysed data, and wrote the manuscript.
Approximate percentage contribution of the candidate to the work (if possible to describe in this way)	
Signature of the candidate	
Date (DD/MM/YY)	04/09/24

Co-author contributions


By signing this Statement of Authorship, each co-author agrees that:

- (i) the candidate has accurately represented their contribution to the work;
- (ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co-author	K. Cowtan
Contact details of co-author	kathryn.cowtan@york.ac.uk
Description of the co-author's contribution to the work*	Provided feedback in a supervisory role during the execution of the work and on drafts of the paper but did not contribute text, analysis or computer code.
Approximate percentage contribution of the co-author to the work (<i>if possible to describe in this way</i>)	
Signature of the co-author	
Date (DD/MM/YY)	04/09/24

Name of co-author	Yuriy Chaban
Contact details of co-author	yuriy.chaban@diamond.ac.uk
Description of the co-author's contribution to the work*	Conceived and designed research, and contributed to data analysis and manuscript writing. The corresponding author of this paper.
Approximate percentage contribution of the co-author to the work (<i>if possible to describe in this way</i>)	
Signature of the co-author	
Date (DD/MM/YY)	03/09/24

Name of co-author	Peijun Zhang
Contact details of co-author	peijun.zhang@diamond.ac.uk
Description of the co-author's contribution to the work*	Conceived and designed research and contributed to data

	analysis and manuscript writing. The corresponding author and lead contact of this paper.
Approximate percentage contribution of the co-author to the work (if possible to describe in this way)	
Signature of the co-author	
Date (DD/MM/YY)	04/09/24

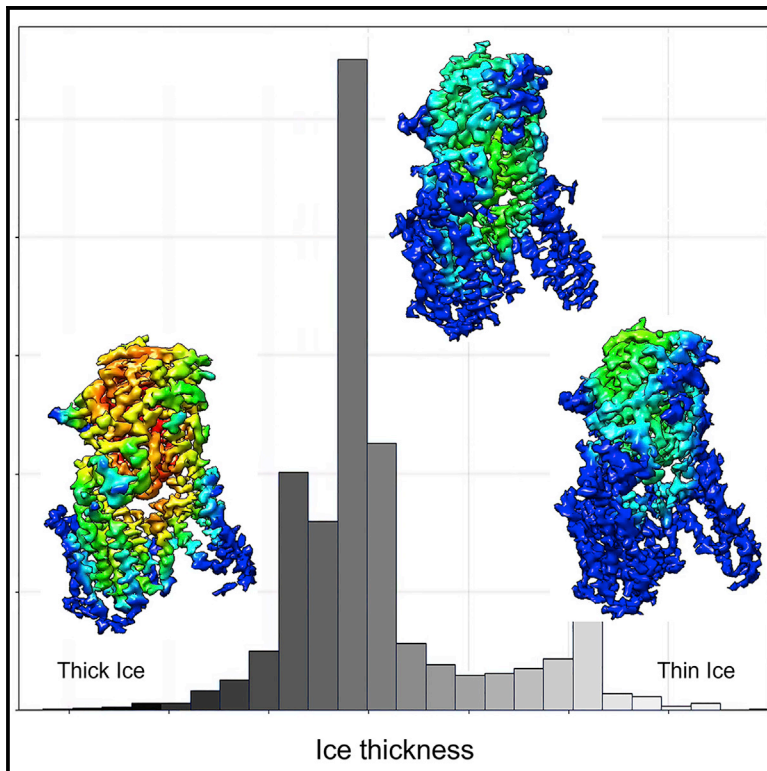
Copy and paste additional co-author panels as needed.

*The description of the candidate and co-authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates and co-authors may find it helpful to consider the [CRediT \(Contributor Roles Taxonomy\)](#) approach to recognising individual author contributions.

Structure

IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice

Graphical abstract



Authors

Mateusz Olek, Kevin Cowtan,
Donovan Webb, Yuriy Chaban,
Peijun Zhang

Correspondence

yuriy.chaban@diamond.ac.uk (Y.C.),
peijun.zhang@strubi.ox.ac.uk (P.Z.)

In brief

Olek et al. present a software tool, IceBreaker, for handling non-uniform ice thickness in cryo-EM micrographs. Ice thickness is believed to be a crucial factor that affects the quality of cryo-EM reconstructions. IceBreaker provides empirical estimation of the ice distribution and introduces an ice thickness parameter to the cryo-EM processing pipeline.

Highlights

- Develop a software tool for image segmentation based on estimated ice thickness
- Present a method to detect and annotate ice contamination in the dataset
- Show a procedure to equalize contrast on the micrographs with the non-uniform ice
- Demonstrate a workflow to identify optimal ice for data collection/particle selection



Resource

IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice

Mateusz Olek,^{1,2} Kevin Cowtan,² Donovan Webb,¹ Yuriy Chaban,^{1,*} and Peijun Zhang^{1,3,4,5,*}

¹Electron Bio-Imaging Centre, Diamond Light Source, Harwell Science and Innovation Campus, Didcot OX11 0DE, UK

²Department of Chemistry, University of York, York, UK

³Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

⁴Chinese Academy of Medical Sciences Oxford Institute, University of Oxford, Oxford OX3 7BN, UK

⁵Lead contact

*Correspondence: yuriy.chaban@diamond.ac.uk (Y.C.), peijun.zhang@strubi.ox.ac.uk (P.Z.)

<https://doi.org/10.1016/j.str.2022.01.005>

SUMMARY

Despite the abundance of available software tools, optimal particle selection is still a vital issue in single-particle cryoelectron microscopy (cryo-EM). Regardless of the method used, most pickers struggle when ice thickness varies on a micrograph. IceBreaker allows users to estimate the relative ice gradient and flatten it by equalizing the local contrast. It allows the differentiation of particles from the background and improves overall particle picking performance. Furthermore, we introduce an additional parameter corresponding to local ice thickness for each particle. Particles with a defined ice thickness can be grouped and filtered based on this parameter during processing. These functionalities are especially valuable for on-the-fly processing to automatically pick as many particles as possible from each micrograph and to select optimal regions for data collection. Finally, estimated ice gradient distributions can be stored separately and used to inspect the quality of prepared samples.

INTRODUCTION

Advancements in cryoelectron microscopy (cryo-EM) instrumentation, detector development, and data processing algorithms have allowed reconstructions to be obtained at atomic resolution (Nakane et al., 2020). The final quality of the cryo-EM reconstruction depends on several factors at different stages from the sample preparation and data collection to the data processing. One of the crucial features is the thickness and variance of the vitreous ice across the grid. The ice parameters in principle can be optimized at the sample preparation stage by the adjustments of plasma exposure time, blot force, and time (Passmore and Russo, 2016). Despite recent advancements in instrumentation, the vitrification process is still highly variable and not reproducible (Dandey et al., 2020; Drulyte et al., 2018; Rubinstein et al., 2019; Tan and Rubinstein, 2020). The overall quality of the prepared cryo-grids needs to be assessed before the data collection. Currently, user tools in data collection software such as EPU can be helpful in the automated selection of the best areas of the grid and excluding damaged areas. More advanced routines to estimate the ice thickness using energy filter, the aperture limited scattering method (Rice et al., 2018), diffraction patterns (Ahn et al., 2020), or classification routines based on machine learning algorithms for the images at low magnification (Yokoyama et al., 2020) allow targeting only the grid areas with desired ice thickness. This can lead to improvements in the final resolution and reduce the data collection time, but most of the methods need to be optimized for each project and microscope (Rheinberger et al., 2021).

The ideal setup for single-particle analysis would have the particles distributed in a thin, vitreous ice layer. The surface of the ice in the data collection areas should be flat and normal to the electron beam. Particles should occupy most of the grid holes, be oriented randomly, and not overlap with each other (Noble et al., 2018). Areas with too thin ice can be devoid of proteins or the proteins can be damaged or denatured on the air-water interface (D'Imprima et al., 2019). Thicker ice results in low SNR, errors in defocus determination, and limits the final resolution. Even though it is recommended to make the grids with thinnest-possible ice that can still support the specimen, in many cases the particles will be pushed to thicker ice areas (Wu et al., 2016), or, in other cases, the particles will have preferred orientation(s) (Cianfrocco and Kellogg, 2020; Glaeser and Han, 2017). Generally, the collected dataset will include images of variable ice thickness that affects signal-to-noise ratio (Baxter et al., 2009). Recently, image processing techniques or artificial intelligence-(AI)-based denoising software tools have been developed to improve the interpretability of the micrographs (Bepier et al., 2020). The denoised micrographs allow for picking additional particles that were otherwise not distinguishable from the noise (Wagner and Raunser, 2020). The problem of preferred orientation and missing angular projections of the specimen can limit the final resolution and affect the performance of the map reconstruction algorithms even with the large number of picked particles (Rosenthal and Henderson, 2003; Sorzano et al., 2021).

One main shortcoming common to most of the state-of-the-art automated tools is the fact that they do not take into consideration the fact that particles distributed in different ice thickness



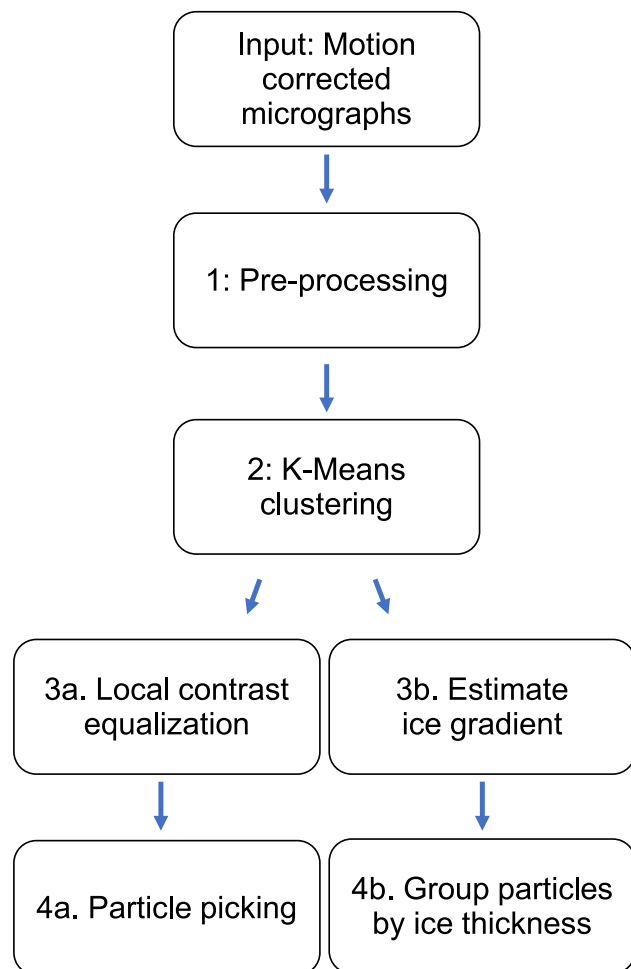


Figure 1. The IceBreaker workflow

The required input is a set of motion-corrected micrographs. The pre-processing stage includes low-pass filtering and further feature flattening done by local averaging. The output image is used for the K-means clustering to obtain segmented micrographs. From the segmented micrographs, the user can create local masks for local contrast improvement, which can lead to improved particle picking, or empirically estimate ice gradient and use this information as an additional parameter for the processing.

regions may have different quality and features. After the processing, most of that information, which could lead to the improvement of the final resolution, cannot be recovered. Currently there is no software tool that allows the user to easily connect the ice thickness parameter with the quality and state of the particles in different areas of the prepared sample.

In this work, we present a software tool, IceBreaker, for the ice thickness estimation and digital ice gradient removal on the cryo-EM micrographs. The software allows the segmentation of the micrographs and grouping areas with similar ice features. It can be used for local image processing as filtering or contrast enhancement, as well as annotating and removal of the ice contamination and/or carbon film fringes. Importantly, it introduces the empirical ice thickness parameter that can be associated with each particle based on the picked coordinates. The described tool can be used as a stand-alone image processing

software or as an external job in the integrated Relion workflow (Zivanov et al., 2018).

RESULTS

The IceBreaker workflow

The IceBreaker software allows segmentation of the cryo-EM micrographs based on the distribution of the pixel intensities recorded by the detector. The term “estimated ice thickness value” is introduced to describe and group the areas of the micrograph with similar pixel intensities. This information can be exploited during the later stages of the cryo-EM processing pipeline; e.g., particle picking, 2D classification, or 3D refinement. An overview of the workflow is presented in Figure 1 with examples of the resulting images. Each of the steps is described below.

Input data: the required input is a set of motion-corrected cryo-EM micrographs. The IceBreaker can be run as an external job of the Relion project or as a stand-alone tool from the command line. It can be used as a part of the data collection pipeline or performed on historical data. Various tools for motion correction (Grant and Grigorieff, 2015; Li et al., 2013; Zheng et al., 2017; Zivanov et al., 2019) can be installed separately. They do not affect IceBreaker results, as long as the whole dataset is processed with the same setup. The pixel intensity values from the input images are used to estimate the distribution of the ice thickness in a given dataset.

Step 1. Pre-processing: filtering and feature flattening: the 20 Å low-pass filter is applied to each micrograph to remove the high-frequency noise and reveal features such as particles, ice contamination, foil hole edges, and the ice gradient. Then, the micrograph is divided into a pre-defined number of patches: 40 in x and 40 in y direction, which is independent of the size of the micrograph. Within each patch, an average value of pixel intensities is calculated. This way local features are reduced to 1/1,600 of the micrograph area on top of the initial 20 Å filter. In our test cases, this was sufficient to reveal trends and low-frequency changes in the background, which represent the changes in the ice thickness. Additionally, the super-pixels represented by each patch can be used to reduce the size of the micrographs and improve the computation speed. Micrographs processed this way are used as input to the next stage of the processing.

Step 2. K-means clustering: the K-means clustering algorithm is used to group together the areas of the rescaled, feature-flattened micrograph with similar values. By default, each micrograph is divided into 16 segments. Then, the segmented image is upscaled to match the original size of the micrograph. This results in a micrograph with 16 discrete regions with unique values of the intensities of the pixels. Each group populates the pixels that originally represented similar background features in a given neighborhood. The segmented micrographs are saved and can be used for further processing in two ways. First, for masking and local processing of the original micrographs, and second, as a reference to identify the micrograph quality in the neighborhood of the coordinates selected during the particle picking.

Step 3. Local processing for contrast improvement or ice gradient estimation: the groups defined in the previous step allow the local processing of the original dataset. Each segment represents an area with similar background features and can be

used as a local mask that can be applied to the original, motion-corrected micrographs. Within each mask, the image processing operations such as contrast improvement can be performed. The application of the contrast equalization in different areas of the micrograph separately results in the final image with a similar ratio between the particles and the background features. This also alleviates the problem of oversaturation of parts of the image when it is equalized as a whole. The resulting image has a similar ratio between the particles and the background, which can be beneficial for the particle picking tools based on the template matching algorithms.

Another use of the presented approach allows estimating the average ice thickness in segmented micrographs. The defined local masks can be applied to the motion-corrected images. Within each mask, an average value of the pixel intensities can be calculated to estimate the ice thickness in the selected region. This way, a set of segmented micrographs with the estimated ice distribution is created and can be used to associate the picked particle coordinates with the background intensity in the area where they come from. The empirical ice thickness parameter describes whether the particle was picked from the area with high signal-to-noise ratio (which would correspond to the thin ice conditions) or low for the particles embedded in thicker ice. It also allows filtering and selecting subsets of particles of similar quality.

The performance of the IceBreaker was tested using several datasets available from the Electron Microscopy Public Image Archive (EMPIAR) database. The presented results are focused on the main features of the software: (1) local contrast enhancement to improve the particle picking; (2) evaluation of the micrographs' quality and identification of the ice contaminations and foil hole edges; and (3) the cryo-EM data processing with the newly introduced empirical ice thickness parameter.

Local contrast enhancement

One of the main challenges when processing cryo-EM micrographs with non-uniform ice distribution is the fact that the contrast levels between the particles and the background features vary in different parts of the image. This can affect the performance of the automated particle pickers, especially those using a single value threshold to detect false-positives. In order to normalize the local contrast between the particles and the background across the whole micrograph, IceBreaker segments low-pass-filtered micrographs into areas of similar overall intensity. The procedure of local contrast enhancement is presented in Figure 2. The input motion-corrected micrograph (Figure 2A) is pre-processed using a low-pass filter to identify the changes in background intensities corresponding to the ice distribution (Figure 2B). The K-means clustering is applied to the low-pass-filtered micrograph to obtain a segmented image (Figure 2C) where pixels with similar intensities are grouped together. Each of the segments created this way can be used as a local mask for image processing. An example of such a mask is highlighted blue in Figure 2D. It can be applied to the low-pass-filtered micrograph to directly access pixel coordinates as shown in Figure 2E. Within each mask, the histogram equalization is performed. This procedure is repeated for each segment of the micrograph. The resulting image in Figure 2F is flattened with the ice gradient removed. Contrast between particles and the background features is improved both in the areas

that were originally dark and bright. Images curated this way can be used as a direct input for automated particle picking. As shown in Figures 2G and 2H, particle picking with crYOLO is much improved after image flattening and contrast enhancement (Figure 2H) compared with the original micrograph (Figure 2G). The particles initially skipped due to poor contrast are now included, especially those in the darker area, yielding a greater number of picked particles. While increasing the number of picked particles is valuable when the dataset is small, views with weak contrast are missing, or when performing 2D classification, users should keep in mind that the quality of the particles from thicker ice regions might be poorer and should be evaluated when aiming for the best possible resolution. IceBreaker introduces means for such evaluations, which are described below. Figure 2I shows a comparison of the number of particles picked with Relion3.1. Laplacian of Gaussian (LoG) autopicker from original micrographs, micrographs after band-pass filtration (with the setup of 20–500 Å), and micrographs after contrast equalization with the IceBreaker. The IceBreaker produces micrographs with consistent intensity distribution, which allows the pickers to perform more reliably. By contrast, the band-pass filter produces a correction that often varies over the area of the micrograph and does not equalize the contrast between particles in thin and thick areas. The improved picking from band-pass-filtered images is still affected by the changes between the micrographs, such as defocus value, as the filter parameters are set globally for the whole dataset. The IceBreaker allows us to improve the contrast for each micrograph individually and achieve better results.

Micrograph quality evaluation and ice contamination detection

The segmented micrographs can be used to evaluate the overall quality of the collected dataset, in addition to CTF estimation. Figure 3A shows the distribution of the pixel intensities, which represents the background for a subset of 20 micrographs from the beta-galactosidase dataset EMPIAR-10204 (Kato et al., 2018). This analysis revealed several features of the data, which are discussed on selected examples of the micrographs and their 3D profiles presented in Figure 3B: (1) micrographs with darker backgrounds, associated with the thicker ice in these areas of the grid, can be easily separated from the ones with a lighter background and thinner ice; (2) a symmetrical box plot indicates a uniform background as in micrograph no.3, while a skewed box plot in micrograph no. 17 or 3D presentation suggests an ice gradient; (3) the outliers in box plot representing micrograph no.10 and the corresponding 3D representation indicate there are ice contaminations. Such analysis provides information that can improve further processing. Micrographs with lower quality can be excluded. The outlier analysis can be helpful to set thresholds for the particle pickers to avoid ice contaminations or remove them from the already-picked set of coordinates. Figure S1A shows a segmented micrograph with the ice contamination in the field of view. The contaminations can be easily identified by checking the pixel intensities distribution (Figure S1B). The coordinates picked with the LoG include areas associated with the contamination, which can be easily removed based on the pixel intensities distribution thresholding (Figures S1C and S1D). Associating the particles' coordinates with local background values can also help to exclude false-positive

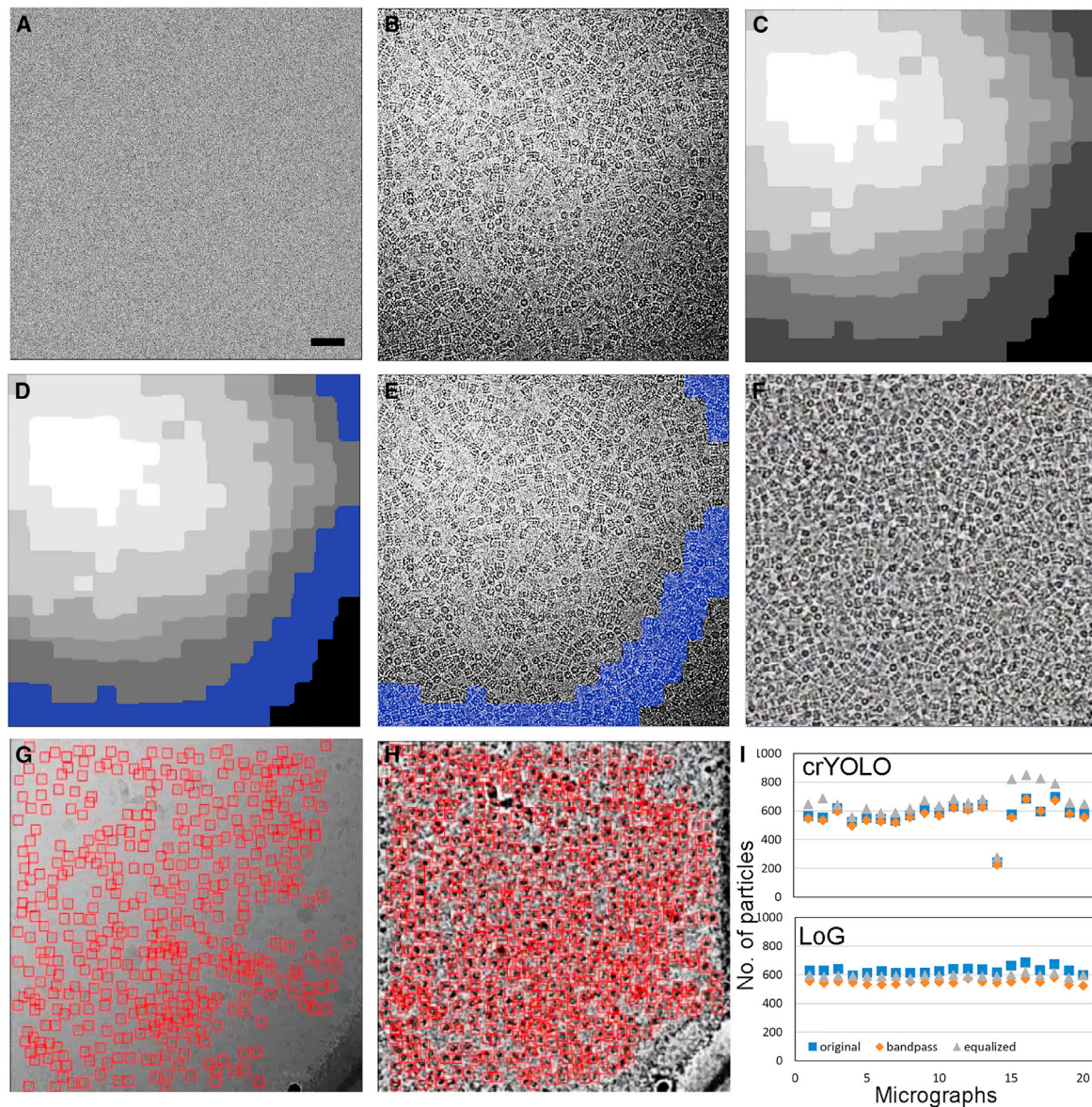


Figure 2. IceBreaker contrast enhancement

(A) A raw micrograph of T20S (EMPIAR-10025) used as an input.

(B) A 20 Å low-pass filtered micrograph, revealing non-uniform distribution of ice.

(C) A segmented micrograph, where each segment can be used as a local mask.

(D and E) Local mask (blue) applied to a corresponding example segment of the micrograph.

(F) The micrograph after contrast equalization.

(G and H) Automated particle picking using crYOLO on the original micrograph (G) and after local contrast equalization (H).

(I) Number of particles picked by crYOLO (top) and LoG (bottom) from original (blue), 20–500 Å band-pass-filtered (orange), and local contrast-equalized (gray) images randomly selected from the dataset (10%). Scale bar, 50 nm.

particle positions with automated pickers based on template matching or machine learning.

Processing based on the ice thickness parameter

The information about the distribution of the background pixel intensities can be associated with the coordinates of the particles estimated using any available picking tool. With the IceBreaker, we introduce a new empirical particle parameter representing the

estimated ice thickness based on the background features of the area where the particle is located. Users can check the overall distribution of the particles and their orientations with respect to their background quality. Figure 4 presents such analysis using the T20S proteasome dataset EMPIAR-10025 (Campbell et al., 2015). The histogram in Figure 4A shows the number of particles associated with different ice thickness values. These values are calculated from the segmented micrographs as an

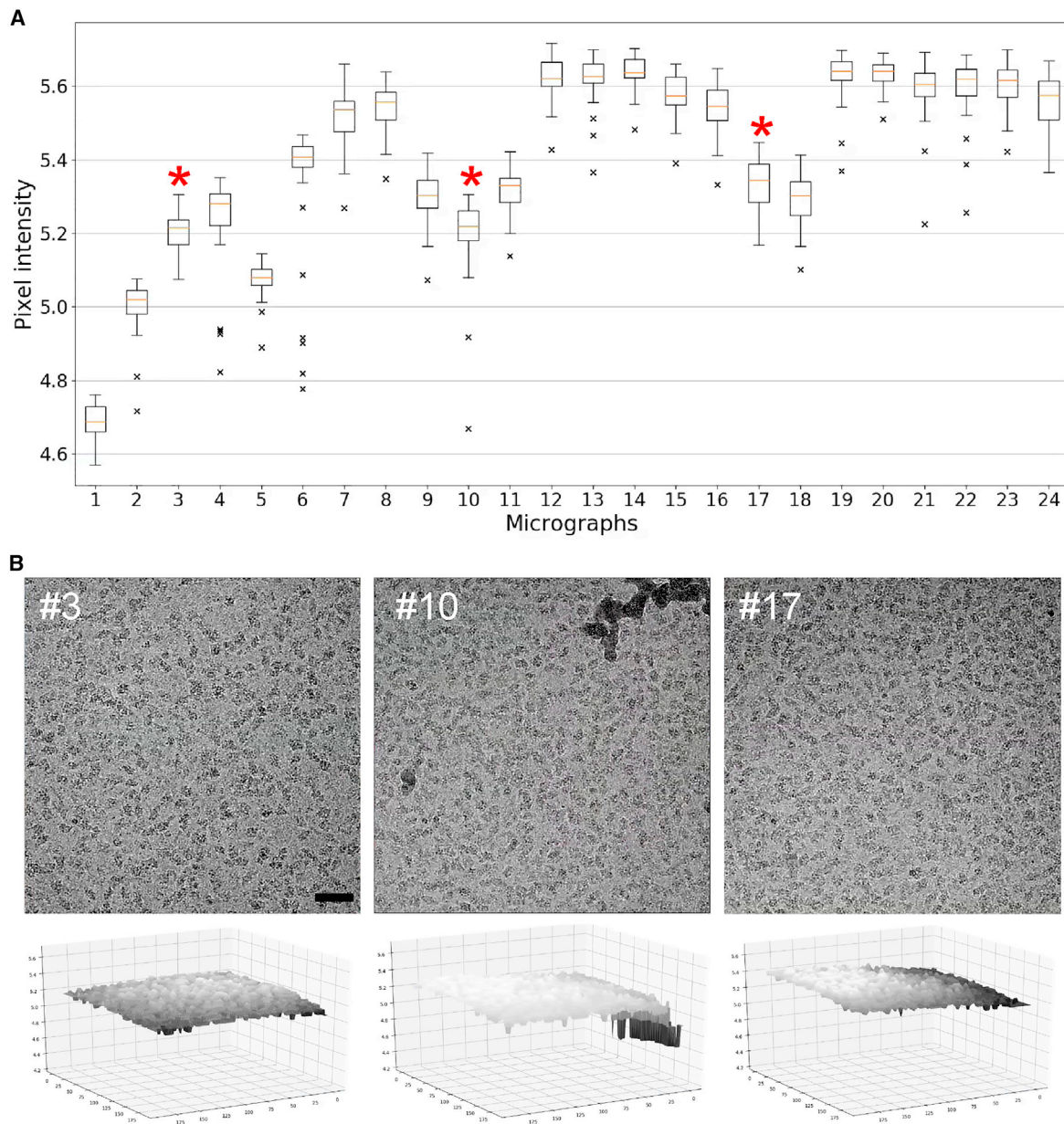


Figure 3. Assess the distribution of the ice by IceBreaker

(A) Box plots for a subset of 24 micrographs of β -Gal from EMPIAR-10204, showing pixel intensities distribution in the micrographs after segmentation. (B) Images and corresponding 3D ice distribution profiles of selected micrographs. Asterisks (*): micrograph no. 3 with no ice contamination and uniform ice distribution, micrograph no. 10 with the ice contamination indicated by the outliers on the box-plot, micrograph no. 17 with the non-uniform ice gradient represented by the skewed distribution. Scale bar, 50 nm. The size of each of the boxes in the box plots (equivalent of error bar) corresponds to the values of the first and the third quartile; orange bar represents median value of the given micrograph. The whiskers indicate datapoints that fall into the 1.5 interquartile range (IQR) and the outliers (marked with black X) represent datapoints that significantly differ from the dataset.

average value of pixel intensities in each segment (Figure 4B). The histogram shows that the majority of the particles were picked from the intermediate ice thickness values. There is an apparent skewness in the particle distribution due to the absence of particles in very thin ice, which is possibly too thin to embed T20Ss proteasome particles. The set of over 120,000 automatically picked particles, after 3D refinement in Relion with the D7 symmetry, were split into 20 groups based on the

ice thickness parameter. This allows us to assess how the particles behave in different ice thickness conditions, as shown in the particle angular distribution (Naydenova and Russo, 2017) plots (Figure 4C). For presentation clarity and to match the lowest populated group I, each plot is done for a randomly selected subset of 100 particles. In group I, which represents the thicker ice area, the number of picked particles is low, but both top views and side views of the T20S proteasome are present. As

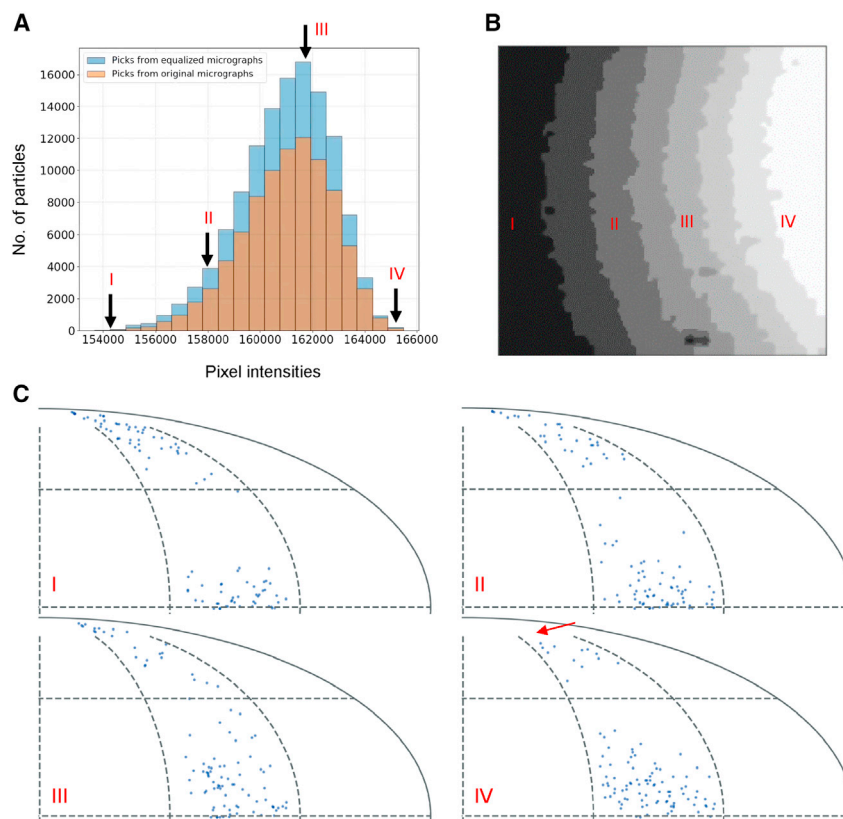


Figure 4. Distribution of T20S particles (EMPIAR-10025) in different ice thickness

(A) Distribution of the number particles picked with crYOLO from original micrographs (gold) and from contrast-equalized micrographs (cyan).

(B) An example of segmented micrograph with strong ice gradient, from the thick (I) to the thin (IV) ice area.

(C) Angular distribution of particles in selected ice thickness areas (I–IV). For each region, 100 particles were selected randomly to match the lowest populated group, I. The red arrow shows that the top views of the particles are not supported in the thinnest ice group, IV, and particles orientation are shifted toward equatorial area.

the ice gets too thin to support the top view (group IV), the angular plot shows a shift from the pole (top view) toward the equatorial area (side view). The selection of the particles from the regions can lead to under-representation of specific views, or preferred orientation, even if the signal-to-noise ratio is better. The most populated groups in the intermediate ice thickness show good support for most of the angular views required for an isotropic reconstruction (groups II–III), still the quality of the particles and signal-to-noise ratio may differ between the groups.

To further gauge the effect of ice thickness on 3D reconstruction, we regrouped the full dataset of picked particles into five groups based on the ice thickness parameter, as shown in Figure 5A. Figures 5B and 5C show the post-processed maps rendered in UCSF Chimera (Pettersen et al., 2004). Maps are colored by the local resolution calculated with LocRes (Kucukelbir et al., 2014) and labeled with the final resolution for each reported after Refine3D and post-processing jobs. Figure 5B shows a comparison of the densities obtained using all 121,000 particles and 66,000 particles from thinnest-ice groups (4 and 5). Particles from optimal ice conditions allowed to obtain similar resolution, 3.19 Å after refinement and 2.87 Å after post-processing, as the larger number of particles (3.19 Å and 2.90 Å respectively). From each ice thickness group, a random subset of 7,000 particles was selected for an additional round of 3D refinement with D7 symmetry followed by the post-processing with Relion. The setup parameters for each subset were the same, as well as the mask used for post-processing. There is a

clear trend that the resolution improves as the ice thickness reduces, from 4.5 Å to 3.8 Å after refinement and 4.0 Å to 3.26 Å after post-processing. This shows that associating the particles with the local ice thickness can help to identify the optimal ice thickness areas to obtain the best possible resolution for a given specimen. This also allows us to test whether preferred orientation may have been caused by recording data from areas of sub-optimal ice thickness. Finally, if the size of the data allows, resolution improvement can be achieved by selecting particles from particular ice groups.

The T20S proteasome has a D7 symmetry and may not be affected by the lower number of edge-on views in thin ice. We therefore selected another low-symmetry particle dataset, gamma-secretase (EMPIAR-10194), for the ice thickness-based refinement (Bai et al., 2015). The distribution of particles in the estimated ice thickness groups was analyzed (Figure 6A). Combining all particles from various ice thickness resulted in a density map at 4.07 Å resolution after refinement and at 3.81 Å post-processing. The particles were later divided into three groups based on the estimated ice thickness value. From each group, a subset of 60,000 particles was randomly selected and refined with C1 symmetry. In this case, a trend of resolution improving with thicker ice allowed to improve resolution from 5.60 Å to 4.59 Å after refinement and from 4.84 Å to 4.16 Å in thick ice after post-processing. This result in conjunction with the previous example shows that particles from different estimated ice regions substantially influence the quality of the cryo-EM map.

DISCUSSION

The non-uniform ice distribution on the cryo-EM micrographs affects the data processing and the quality of the final map. The thickness of ice in which the particles are embedded affects the local signal-to-noise ratio, particle quality, and behavior. The presented software, IceBreaker, aims to overcome the issues caused by the varying ice gradient. The tailored contrast enhancement can improve the micrographs' interpretability

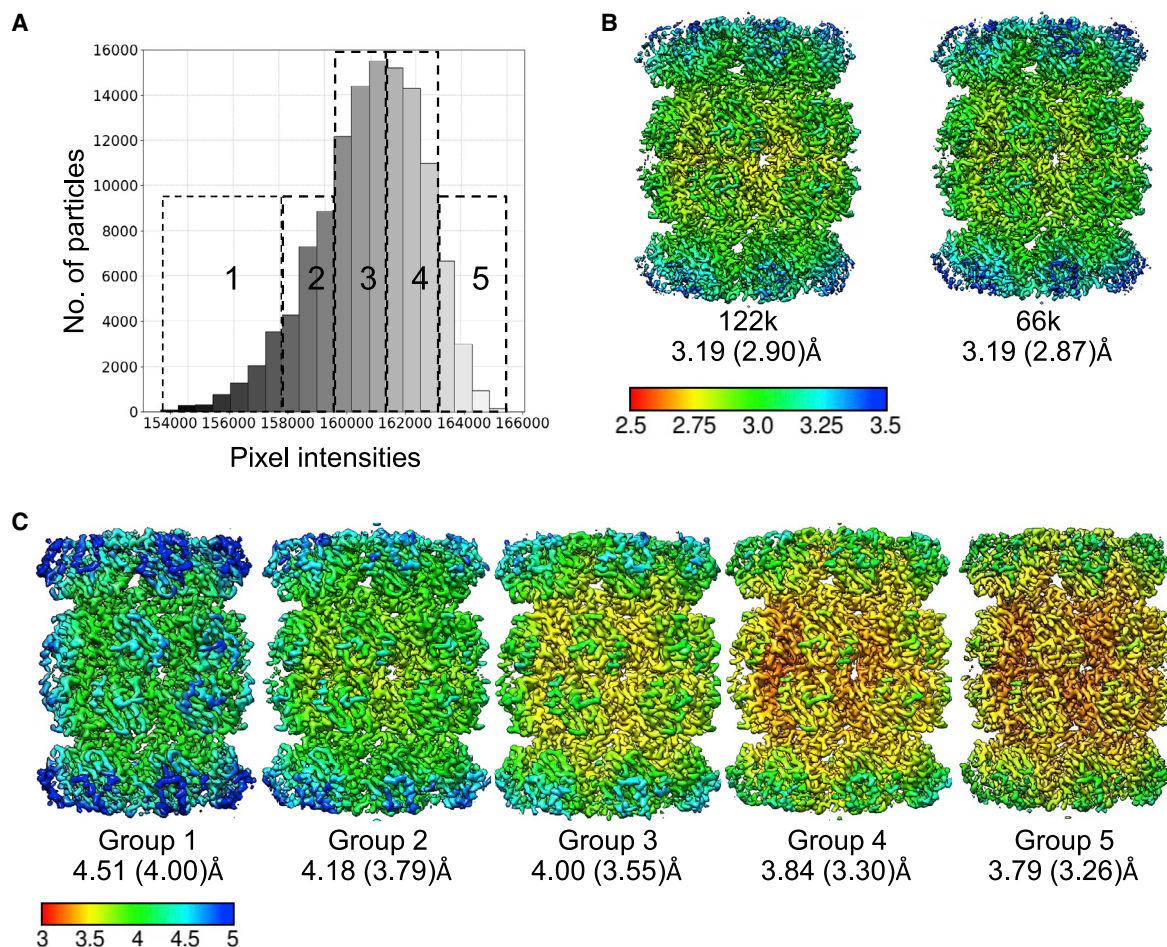


Figure 5. 3D reconstruction of T20S particles based on ice thickness

(A) Particle subsets selected from the T20s dataset (EMPIAR-10025) according to ice thickness parameter. Group 1 corresponds to the thick ice, group 5 the thin ice.

(B) Cryo-EM maps reconstructed from all 121,913 particles and 66,000 particles from thinnest ice groups, 4 and 5.

(C) Cryo-EM maps reconstructed from a set of 7,000 particles picked randomly from each ice group 1–5, D7 symmetry applied. Maps are colored according to the local resolution. For each map, reported resolutions after 3D refinement and post-processing are indicated. Temperature scale bar values are in Angstroms.

and the performance of automated particle picking tools. At the same time, it allows the application of information about the ice distribution in the original micrograph for later stages of processing. To our knowledge, currently no other software offers this level of insight into the ice gradient in the micrograph.

With the analysis of pixel intensity distribution in the segmented micrographs, users can get an insight into overall quality of the collected data. This helps the user to easily identify the micrographs with non-uniform ice distribution, ice contamination, and foil hole edges in the field of view. Based on the outlier analysis, a threshold can be applied to exclude areas of poor quality from further processing.

Our software allows determining empirically and associating the ice thickness parameter with each particle. It allows us to select optimal particles and achieve the best possible resolution for collected cryo-EM datasets. It provides users with additional information about the dataset and the possibility to determine angular distribution of particles in different ice

gradient regions. Users can filter and group the particles based on the estimated optical density of the micrographs, normally associated with amorphous ice thickness, ice contamination, or foil hole fringes. Presented results using the EMPIAR-10025 dataset as an example show improvement in the final resolution of the map with the particles picked from thinner ice. Because the T20S proteasome has high symmetry, the effect of missing orientations in thinner ice areas was less prominent. The fact that the non-symmetrical gamma-secretase dataset (EMPIAR-10194) has improved resolution of the map from thicker regions shows that the local ice conditions can affect the quality of the final map, and the thinnest ice sometimes has to be avoided. In this case, better results were obtained from thicker ice. This type of analysis can be done during the initial, small-scale data collection to determine the optimal setup for a given dataset and to target the best ice conditions, whether for the optimal angular orientation coverage or for a better signal-to-noise ratio.

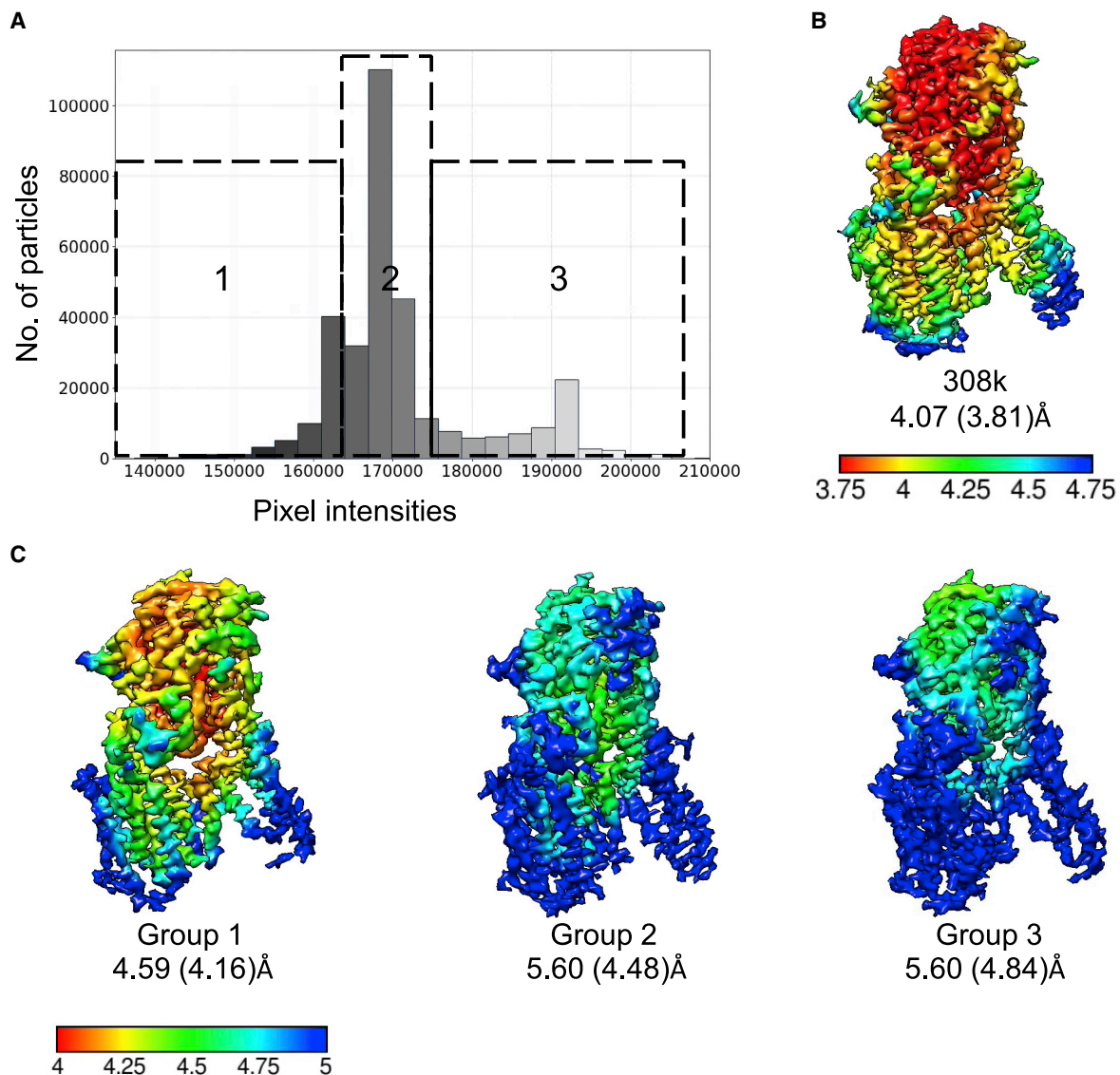


Figure 6. 3D reconstruction of gamma-secretase particles based on ice thickness

(A) Particle subsets selected from the gamma-secretase dataset (EMPIAR-10194) according to ice thickness parameter. Group 1 corresponds to the thick ice, group 3 the thin ice.

(B) Cryo-EM map reconstructed from all selected particles.

(C) Cryo-EM maps reconstructed from a set of 60,000 particles picked randomly from each group 1–3, C1 symmetry applied. Maps are colored according to the local resolution. For each map, reported resolutions after 3D refinement and post-processing are indicated. Temperature scale bar values are in Angstroms.

The IceBreaker can be run as an external job in an existing Relion3.1 project, as it has been integrated into Relion seamlessly, or run as a stand-alone software. Further integration with data collection pipelines, such as IspyB (Delagenière et al., 2011), can extend the use of IceBreaker for selection of the best regions for data acquisition on the fly, based on specimen properties. The software is being incorporated as a part of the data processing pipeline (Fernandez-Leiro and Scheres, 2017) and the CCP-EM software suite (Burnley et al., 2017). We demonstrate the utility of IceBreaker with a few examples shown here, and the method can be applied to any cryo-EM single-particle dataset, either already collected or being collected.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Materials availability
 - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHODS DETAILS

- IceBreaker scripts
- Image processing and analysis
- T2OS data processing
- Gamma-secretase processing
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.str.2022.01.005>.

ACKNOWLEDGMENTS

We thank Dr. Jamie Blaza, Prof. Fred Antson, and Prof. Roderick Hubbard for useful discussion. We thank Dr. Colin Palmer, Dr. Tom Burnley, Dr. Markus Gerstel, Miss Anna Horstmann, Dr. Daniel Hatton, and Diamond Scientific Computing for technical support. This work was supported by the University of York and Diamond [Light Source Ltd] joint studentship, the UK Wellcome Trust Investigator Award 206422/Z/17/Z, the UK Biotechnology and Biological Sciences Research Council grant BB/S003339/1, the European Research Council Advanced Grant 101021133, and the Medical Research Council grant MR/V000403/1.

AUTHOR CONTRIBUTIONS

M.O., Y.C., K.C., and P.Z. conceived and designed research. M.O. developed the IceBreaker software tool and processed the data. D.W. helped with the integration of IceBreaker into Relion. M.O., Y.C., K.C., and P.Z. analyzed data and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 4, 2021

Revised: December 1, 2021

Accepted: January 18, 2022

Published: February 11, 2022

REFERENCES

Ahn, J., Lee, D., Jo, I., Jeong, H., Hyun, J.-K., Woo, J.-S., Choi, S.-H., and Ha, N.-C. (2020). Real-time measurement of the liquid amount in cryo-electron microscopy grids using laser diffraction of regular 2-D holes of the grids. *Mol. Cells* **43**, 298–303.

Bai, X., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., Zhou, R., Scheres, S.H.W., and Shi, Y. (2015). An atomic structure of human γ -secretase. *Nature* **525**, 212–217.

Baxter, W.T., Grassucci, R.A., Gao, H., and Frank, J. (2009). Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. *J. Struct. Biol.* **166**, 126–132.

Bepler, T., Kelley, K., Noble, A.J., and Berger, B. (2020). Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nat. Commun.* **11**, 5208.

Bradski, G. (2000). The OpenCV library. *Dr.Dobbs Journal*.

Burnley, T., Palmer, C.M., and Winn, M. (2017). Recent developments in the CCP-EM software suite. *Acta Crystallogr. Section D Struct. Biol.* **73**, 469–477.

Burnley, T., Palmer, C., and Winn, M. (2017). Recent developments in the CCP-EM software suite. *Acta Cryst. D* **73**, 469–477.

Campbell, M.G., Veesler, D., Cheng, A., Potter, C.S., and Carragher, B. (2015). 2.8 Å resolution reconstruction of the *Thermoplasma acidophilum* 20S protease using cryo-electron microscopy. *eLife* **4**, e06380.

Cianfrocco, M.A., and Kellogg, E.H. (2020). What could go wrong? A practical guide to single-particle cryo-EM: from biochemistry to atomic models. *J. Chem. Inf. Model.* **60**, 2458–2469.

Dandey, V.P., Budell, W.C., Wei, H., Bobe, D., Maruthi, K., Kopylov, M., Eng, E.T., Kahn, P.A., Hinshaw, J.E., Kundu, N., et al. (2020). Time-resolved cryo-EM using Spotiton. *Nat. Methods* **17**, 897–900.

Delagenière, S., Brenchereau, P., Launer, L., Ashton, A.W., Leal, R., Veyrier, S., Gabadinho, J., Gordon, E.J., Jones, S.D., Levik, K.E., et al. (2011). ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics* **27**, 3186–3192.

D'Imprima, E., Floris, D., Joppe, M., Sánchez, R., Grninger, M., and Kühlbrandt, W. (2019). Protein denaturation at the air-water interface and how to prevent it. *eLife* **8**, e42747.

Drulyte, I., Johnson, R.M., Hesketh, E.L., Hurdiss, D.L., Scarff, C.A., Porav, S.A., Ranson, N.A., Muench, S.P., and Thompson, R.F. (2018). Approaches to altering particle distributions in cryo-electron microscopy sample preparation. *Acta Crystallogr. Section D Struct. Biol.* **74**, 560–571.

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

Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions* (CRC Press).

Glaeser, R.M., and Han, B.-G. (2017). Opinion: hazards faced by macromolecules when confined to thin aqueous films. *Biophys. Rep.* **3**, 1–7.

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

Kato, T., Terahara, N., and Namba, K. (2018). EMPIAR-10204 dataset.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. *Nature* **585**, 357–362.

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

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

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theor.* **28**, 129–137.

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P.M.G.E., Grigoras, I.T., Malinauskaitė, L., Malinauskas, T., Miehl, J., et al. (2020). Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156.

Naydenova, K., and Russo, C.J. (2017). Measuring the effects of particle orientation to improve the efficiency of electron cryomicroscopy. *Nat. Commun.* **8**, 629.

Noble, A.J., Dandey, V.P., Wei, H., Brasch, J., Chase, J., Acharya, P., Tan, Y.Z., Zhang, Z., Kim, L.Y., Scapin, G., et al. (2018). Routine single particle cryoEM sample and grid characterization by tomography. *eLife* **7**, e34257.

Passmore, L.A., and Russo, C.J. (2016). Specimen preparation for high-resolution cryo-EM. *Methods Enzymol.* **579**, 51–86.

Petterson, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612.

Rheinberger, J., Oostergetel, G., Resch, G.P., and Paulino, C. (2021). Optimized cryo-EM data acquisition workflow by sample thickness determination. *BioRxiv*, 2020.12.01.392100.

Rice, W.J., Cheng, A., Noble, A.J., Eng, E.T., Kim, L.Y., Carragher, B., and Potter, C.S. (2018). Routine determination of ice thickness for cryo-EM grids. *J. Struct. Biol.* **204**, 38–44.

Rohou, A., and Grigorieff, N. (2015). CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221.

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

Rubinstein, J.L., Guo, H., Ripstein, Z.A., Haydaroglu, A., Au, A., Yip, C.M., Di Trani, J.M., Benlekhir, S., and Kwok, T. (2019). Shake-it-off: a simple ultrasonic

cryo-EM specimen-preparation device. *Acta Crystallogr. Section D: Struct. Biol.* 75, 1063–1070.

Sorzano, C.O.S., Semchonok, D., Lin, S.-C., Lo, Y.-C., Vilas, J.L., Jiménez-Moreno, A., Gragera, M., Vacca, S., Maluenda, D., Martínez, M., et al. (2021). Algorithmic robustness to preferred orientations in single particle analysis by CryoEM. *J. Struct. Biol.* 213, 107695.

Tan, Y.Z., and Rubinstein, J.L. (2020). Through-grid wicking enables high-speed cryoEM specimen preparation. *Acta Crystallogr. Section D Struct. Biol.* 76, 1092–1103.

Tukey, J., W (1977). *Exploratory Data Analysis*XVI, 688 S (Mass. - Menlo Park, Cal., London, Amsterdam, Don Mills Ontario, Sydney: Addison-Wesley Publishing Company Reading).

Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D., et al. (2019). SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun. Biol* 2, 218. <https://doi.org/10.1038/s42003-019-0437-z>.

Wagner, T., and Raunser, S. (2020). The evolution of SPHIRE-crYOLO particle picking and its application in automated cryo-EM processing workflows. *Commun. Biol.* 3, 1–5.

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

Yokoyama, Y., Terada, T., Shimizu, K., Nishikawa, K., Kozai, D., Shimada, A., Mizoguchi, A., Fujiyoshi, Y., and Tani, K. (2020). Development of a deep learning-based method to identify “good” regions of a cryo-electron microscopy grid. *Biophys. Rev.* 12, 349–354.

Zheng, S.Q., Palovcak, E., Armache, J.-P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2 - anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* 14, 331–332.

Zivanov, J., Nakane, T., Forsberg, B.O., Kimanius, D., Hagen, W.J., Lindahl, E., and Scheres, S.H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* 7, e42166.

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Beta-galactosidase	(Kato et al., 2018)	EMPIAR-10204
<i>Thermoplasma acidophilum</i> 20S proteasome	(Campbell et al., 2015)	EMPIAR-10025
Human gamma-secretase	(Bai et al., 2015)	EMPIAR-10194 PDB: 5a63
The cryoEM density map of T20s proteasome with various ice thickness, subset 1 (EMPIAR-10025 reprocessing)	This paper	EMD-13309
The cryoEM density map of T20s proteasome with various ice thickness, subset 2 (EMPIAR-10025 reprocessing)	This paper	EMD-13310
The cryoEM density map of T20s proteasome with various ice thickness, subset 3 (EMPIAR-10025 reprocessing)	This paper	EMD-13311
The cryoEM density map of T20s proteasome with various ice thickness, subset 4 (EMPIAR-10025 reprocessing)	This paper	EMD-13312
The cryoEM density map of T20s proteasome with various ice thickness, subset 5 (EMPIAR-10025 reprocessing)	This paper	EMD-13313
The cryoEM density map of T20s proteasome with various ice thickness, subset 4 and 5 combined (EMPIAR-10025 reprocessing)	This paper	EMD-13902
The cryoEM density map of T20s proteasome with various ice thickness, full dataset (EMPIAR-10025 reprocessing)	This paper	EMD-13901
The cryoEM density map of human gamma-secretase complex with various ice thickness, subset 1 (EMPIAR-10194 reprocessing)	This paper	EMD-13903
The cryoEM density map of human gamma-secretase complex with various ice thickness, subset 2 (EMPIAR-10194 reprocessing)	This paper	EMD-13904
The cryoEM density map of human gamma-secretase complex with various ice thickness, subset 3 (EMPIAR-10194 reprocessing)	This paper	EMD-13905
The cryoEM density map of human gamma-secretase complex with various ice thickness, full dataset (EMPIAR-10194 reprocessing)	This paper	EMD-13907
Software and algorithms		
Relion3.1	(Zivanov et al., 2018)	https://github.com/3dem/relion
MOTIONCORR2	(Zheng et al., 2017)	https://emcore.ucsf.edu/ucsf-software
CTFFIND-4.1	(Rohou and Grigorieff, 2015)	https://grigoriefflab.umassmed.edu/ctf_estimation_ctffind_ctffilt
crYOLO	(Wagner et al., 2019)	https://pypi.org/project/cryolo/

(Continued on next page)

<i>Continued</i>		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
LocRes	(Kucukelbir et al., 2014)	http://resmap.sourceforge.net
Mrcfile	(Burnley et al., 2017)	https://github.com/ccpem/mrcfile
NumPy	(Harris et al., 2020)	https://numpy.org
OpenCV	(Bradski, 2000)	https://opencv.org
Gemmi	GEMMI - library for structural biology – Gemmi 0.5.2 documentation	https://github.com/project-gemmi/gemmi
Chimera	(Pettersen et al., 2004)	https://www.cgl.ucsf.edu/chimera/
IceBreaker	This paper	https://github.com/DiamondLightSource/python-icebreaker https://pypi.org/project/icebreaker-em/ Zenodo deposition https://doi.org/10.5281/zenodo.5743790

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Prof. Peijun Zhang (peijun.zhang@strubi.ox.ac.uk)

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the [key resources table](#). The reconstructed cryoEM density maps have been deposited at EMDB and are publicly available as of the date of publication. Accession numbers are listed in the [key resources table](#). The cryoEM density maps reconstructed from T20S proteasome particles picked from various ice-thickness areas have been deposited in the EMDB under accession code EMD-13309 for the group 1 with thickest ice, EMD-13310 for the group 2, EMD-13311 for the group 3, EMD-13312 for the group 4, EMD-13313 for the group 5 with thinnest ice, EMD-13902 for the combined group 4 and 5 and EMD-13901 for the full dataset respectively. The cryoEM density maps from human gamma-secretase particles picked from various ice-thickness areas have been deposited in the EMDB under accession code EMD-13903 for the group 1 with thickest ice, EMD-13904 for the group 2, EMD-13905 for the group 3 with thinnest ice and EMD-13907 for the full dataset.

The code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the [key resources table](#).

The software is freely available also from <https://github.com/DiamondLightSource/python-icebreaker> or can be downloaded with the Python Package Index <https://pypi.org/project/icebreaker-em/>

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All data are generated from the datasets provided in the [Key resources table](#).

METHODS DETAILS

IceBreaker scripts

The IceBreaker can be run from the command line or as an external job in Relion project. The software includes two main scripts. The `ib_job.py` can be used for image processing. It requires motion-corrected micrographs as an input. It can be run in two modes: ‘flatten’ to improve the contrast or ‘group’ to estimate the ice thickness in different areas of the micrographs. The number of threads for parallel processing can also be defined with input parameter but is limited by the number of available CPU threads. Example command which can be used with Relion is: `ib_job -o Output/Directory/ -in_mics PathToMotionCorrMicrographs.star -mode flatten -j 10`: the micrographs listed in the star file will be processed to improve the contrasts. 10 threads will be used to process 10 micrographs at the same time and speed up the processing. The output micrographs will have the same name as input files with suffix ‘_flattened.mrc’.

`ib_job -o Output/Directory/ -in_mics PathToMotionCorrMicrographs.star -mode group -j 10`: the micrographs listed in the star file will be segmented according to the background pixels intensities. Again, 10 threads will be used to process 10 micrographs at the

same time and speed up the processing. The output micrographs will have the same name as input files with suffix ‘_grouped.mrc’.

The second script **ib_group.py** is used to process the star file with particle coordinates and associate them with estimated background quality. As input, it requires a star file with particle coordinates and a set of ‘grouped’ micrographs created in the previous step with ‘ib_job.py’ in group mode. Example command to run ‘ib_group.py’ is: `ib_group -o OutputFile.star -in_mics micrographs_grouped.star -in_parts particles.star`

The output .star file has an additional column with the ‘ice-thickness’ parameter value for each particle. As for now, this new parameter is labelled as ‘_rinHelicalTubelD’. The star file can be used in Relion to select subsets of the particles in the processing pipeline.

Image processing and analysis

The IceBreaker is written in Python 3. The micrographs are processed with the mrcfile package (Burnley et al., 2017). The STAR files are handled with GEMMI. The tool requires NumPy (Harris et al., 2020) and OpenCV (Bradski, 2000) packages for data processing.

The image segmentation is done with the K-Means algorithm (Lloyd, 1982). It is a commonly used clustering algorithm which can give insight into the structure of the data, in this case the micrographs. The n observations are split into k number of sets S , where $k \leq n$. The objective is to group observations in sets in a way to minimize the sum of squared distances (variance) between the observations and the centre of the cluster to which they are assigned, according to the (Equation 1):

$$\operatorname{argmin}_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 = \operatorname{argmin}_S \sum_{i=1}^k |S_i| \operatorname{Var} S_i \quad (\text{Equation 1})$$

where x denotes observation, S_i is a set of observations and μ_i represents the mean of points in set S_i .

The contrast improvement performed in each defined local mask is based on the histogram equalization algorithm. It adjusts the contrast of the input image to evenly utilize the full range of intensities. To do so, the cumulative distribution function (cdf) calculated for the histogram normalized between 0 and 1 has to be linearised to produce a new image with a flat histogram. The (Equation 2) describes the linearised cdf:

$$\operatorname{cdf}_y(i) = (i + 1)K \text{ for } 0 \leq i \leq L \quad (\text{Equation 2})$$

where y is the corrected image, i is the pixel intensity level, K is a constant value and L is total number of intensity levels. The cumulative distribution function is increasing and continuous thus according to the definition of the inverse distribution function, if $F^{-1}(p)$, $p \in (0, 1)$, there is a real number x that $F(x) = p$, therefore $F^{-1}(F(x)) = x$ (Gilchrist, 2000). The transform which is applied to the original image to obtain corrected image is described with (Equation 3):

$$y = T(k) = \operatorname{cdf}_x(k) \quad (\text{Equation 3})$$

where y is the corrected image, x is the initial image and k is the pixel intensity level in the range $[0, L-1]$.

To evaluate the quality of the micrographs the box plots are used. They provide information about the data distribution based on the five-number summary (Tukey, 1977). It includes the minimum, the maximum, the median and the first and the third quartile. The first quartile (Q_1) represents the 25th percentile, which means that 25% of recorded observations have lower value. The third quartile (Q_3) represents the 75th percentile. The size of the box is determined by the interquartile range (IQR) which is a distance between Q_1 and Q_3 , $\operatorname{IQR} = Q_3 - Q_1$. The outliers are detected as observations outside the range:

$$[Q_1 - 1.5/\operatorname{IQR}, Q_3 + 1.5/\operatorname{IQR}] \quad (\text{Equation 4})$$

T20S data processing

The deposited dataset was averaged, therefore no further motion correction was performed. The dataset was processed with Relion3.1 pipeline. The parameters of contrast transfer function were estimated with CTFFIND-4.1 (Rohou and Grigorieff, 2015). The motion corrected micrographs had contrast equalized with the IceBreaker for particle picking. The total number of particles picked with crYOLO was 163,630. After manual selection of the best 2D classes from reference-free classification 121,913 particles were used for 3D classification. The best 3D class was used as a reference for 3D refinement with D7 symmetry which resulted in 3.19 Å resolution based on the gold standard FSC = 0.143 criterion. The post-processing with the soft mask created from low-pass filtered initial 3D class and automatically estimated negative B-factor resulted with 2.90 Å final resolution. Local resolution changes were calculated with LocRes and rendered with UCSF Chimera. After refinement the particles were divided into five subsets according to the estimated ice thickness value, from each group a set of 7,000 particles was randomly selected and refined again to see how the varying ice affects the final resolution.

Gamma-secretase processing

The dataset was processed with Relion3.1 pipeline. Motion correction was done using MotionCor2 with 5x5 patches and binning factor 2. CTFFIND-4.1 was used to estimate the parameters of contrast transfer function. 920,945 particles picked with crYOLO from 2,925 micrographs were used for reference-free 2D classification. The best 2D classes were selected manually. The initial 3D classification resulted with reported resolution 7.47 Å. 308,706 particles from the best 3D classes were used for the 3D refinement

with C1 symmetry and resulted in 4.07 Å resolution based on the gold standard FSC = 0.143 criterion. The map was sharpened using a soft mask created from the atomic model PDB 5a63 (Bai et al., 2015) and with automatically estimated negative B-factor. After sharpening, the final resolution was 3.81 Å. The changes in local resolution were calculated using LocalRes. The larger number of particles were kept to allow selection of representative subsets from different estimated ice thickness levels. The particles used for the 3D refinement were associated with the estimated ice thickness value using the IceBreaker. Three subsets of 60,000 particles each were selected randomly from groups representing thin, medium and thick ice and used for re-refinement and post-processing with the same setup.

QUANTIFICATION AND STATISTICAL ANALYSIS

The methods of statistical analysis are provided in [method details](#) and [supplemental information](#).

Appendix B

University of York
York Graduate Research School
Research Degree Thesis Statement of Authorship

Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

Candidate name	Mateusz Olek
Department	Chemistry
Thesis title	Improved methods for the annotation and processing of Cryo-EM images and for atomic model validation


Title of the work (paper/chapter)	Cryo-EM Map-Based Model Validation Using the False Discovery Rate Approach	
Publication status	Published	x
	Accepted for publication	
	Submitted for publication	
	Unpublished and unsubmitted	
Citation details (if applicable)	Olek M, Joseph AP. Cryo-EM Map-Based Model Validation Using the False Discovery Rate Approach. Front Mol Biosci. 2021 May 18;8:652530. doi: 10.3389/fmolb.2021.652530. PMID: 34084774; PMCID: PMC8167059.	

Description of the candidate's contribution to the work*	Conceived the idea, developed the tool, performed analysis on different examples, and drafted the manuscript.
Approximate percentage contribution of the candidate to the work (if possible to describe in this way)	
Signature of the candidate	
Date (DD/MM/YY)	04/09/24

Co-author contributions

By signing this Statement of Authorship, each co-author agrees that:

- (i) the candidate has accurately represented their contribution to the work;
- (ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co-author	Agnel-Praveen Joseph
Contact details of co-author	Research Complex at Harwell, Rutherford Appleton Laboratory, OX11 0FA
Description of the co-author's contribution to the work*	Help with data analysis, implementation of the tool in the CCP-EM, and manuscript writing. The corresponding author of this paper.
Approximate percentage contribution of the co-author to the work (if possible to describe in this way)	
Signature of the co-author	
Date (DD/MM/YY)	04/09/24

Copy and paste additional co-author panels as needed.

*The description of the candidate and co-authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates and co-authors may find it helpful to consider the [CRediT \(Contributor Roles Taxonomy\)](#) approach to recognising individual author contributions.



Cryo-EM Map-Based Model Validation Using the False Discovery Rate Approach

Mateusz Olek^{1,2} and Agnel Praveen Joseph^{3*}

¹Department of Chemistry, University of York, York, United Kingdom, ²Electron BiImaging Center, Rutherford Appleton Laboratory, Didcot, United Kingdom, ³Scientific Computing Department, Science and Technology Facilities Council, Research Complex at Harwell, Didcot, United Kingdom

Significant technological developments and increasing scientific interest in cryogenic electron microscopy (cryo-EM) has resulted in a rapid increase in the amount of data generated by these experiments and the derived atomic models. Robust measures for the validation of 3D reconstructions and atomic models are essential for appropriate interpretation of the data. The resolution of data and availability of software tools that work across a range of resolutions often limit the quality of derived models. Hence, the final atomic model is often incomplete or contains regions where atomic positions are less reliable or incorrectly built. Extensive manual pruning and local adjustments or rebuilding are usually required to address these issues. The presented research introduces a software tool for the validation of the backbone trace of atomic models built in the cryo-EM density maps. In this study, we use the false discovery rate analysis, which can be used to segregate molecular signals from the background. Each atomic position in the model can be associated with an FDR backbone validation score, which can be used to identify potential mistraced residues. We demonstrate that the proposed validation score is complementary to existing validation metrics and is useful especially in cases where the model is built in the maps having varying local resolution. We also discuss the application of the score for automated pruning of atomic models built *ab-initio* during the iterative model building process in Buccaneer. We have implemented this score in the CCP-EM software suite.

OPEN ACCESS

Edited by:

Giulia Palermo,
University of California, Riverside,
United States

Reviewed by:

Pavel Afonine,
Lawrence Berkeley National
Laboratory, United States
Carlos Oscar Sanchez Sorzano

*Correspondence:

Agnel Praveen Joseph
agnel-praveen.joseph@stfc.ac.uk

Specialty section:

This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: 12 January 2021

Accepted: 26 April 2021

Published: 18 May 2021

Citation:

Olek M and Joseph AP (2021) Cryo-EM Map-Based Model Validation Using the False Discovery Rate Approach. *Front. Mol. Biosci.* 8:652530. doi: 10.3389/fmolb.2021.652530

Keywords: cryo-EM, model validation, FDR map, CCP-EM, automated model building

INTRODUCTION

Improvements in cryo-EM data collection and processing techniques in recent years have enabled structure determination at near-atomic resolutions (Subramaniam, 2019). For structure interpretation, a number of tools for *ab-initio* model building have been developed and used in recent years (Hoh et al., 2020; Terwilliger et al., 2020; Pfab et al., 2021; Lawson et al., 2021). Despite the resolution revolution, the majority of maps (92%) deposited in the EM Data Bank (<https://www.ebi.ac.uk/pdbe/emdb/>) are at resolutions worse than 3 Å, and the average resolution of maps this year is around 5 Å (https://www.ebi.ac.uk/pdbe/emdb/statistics_sp_res.html/). Moreover, some local areas in the cryo-EM map can be poorly resolved. These issues may result in some parts of the derived atomic model being incorrectly built or traced into background noise. Model validation tools that are based on the analysis of stereochemical properties of the atomic model, such as MolProbity

(Williams et al., 2018), CaBLAM (Prisant et al., 2020), or Ramachandran plots, detect potential issues with the geometry of the model. The users can inspect the possible incorrect regions of the model and attempt to fix these in interactive visualization tools like Coot (Emsley et al., 2010).

Another set of validation tools evaluate the agreement of the atomic model with the cryo-EM map. Some of these scores can estimate the agreement of each residue against the area of the map covered by the residue. The agreement is either quantified as Manders' overlap coefficient in SMOG (Joseph et al., 2016), real space cross-correlation coefficient in PHENIX local CCC (Afonine et al., 2018), or a score of atomic resolvability in MapQ (Pintilie 2020). One of the observations from the recent model challenge is that the absolute values of some of these metrics are sensitive to the map resolution (Lawson et al., 2021). One reason is the underlying sensitivity of the metric toward differences in the shape of map distributions at different resolutions. Another reason is the fact that the synthetic map calculation from the model may not be optimal to represent experimental data at different resolutions.

The recently introduced FSC-Q score allows us to assess the local agreement of a model with the cryo-EM density map, and is normalized to account for local resolution variation (Ramírez-Aportela et al., 2021). The map-model local Fourier shell correlation (FSC) is normalized with respect to the local FSC obtained from the halfmaps. The FSC-Q score is calculated as the difference between these two and the values fluctuates around 0. A threshold of ± 0.5 is recommended to detect poorly fitted atoms. Although the FSC-Q calculation is not directly affected by the B-factor values used for map sharpening, the mask applied can have an effect in the local FSC calculation.

MapQ scores atoms in the residues by comparing the distance-dependent map value fall-off against a Gaussian-like reference derived from a map of apoferritin resolved at 1.54 Å and an associated well-fitted atomic model. The Q-score is calculated as a correlation between the map values and the reference Gaussian. Values close to 1 indicate that the atom is well resolved (Pintilie, 2020).

Metrics such as the atom-inclusion score (Lagerstedt et al., 2013), implemented as part of EMDB validation analysis, identify atoms in the model that are outside a selected map contour. The score is hence very sensitive to the choice of map contour, which is often subjective. Also, in cases where the local resolution varies across the map, a single contour may not be optimal to cover the entire molecular volume without including the background noise.

The resolution of cryo-EM maps may vary as a result of molecular flexibility, partial occupancy, non-uniform particle orientation, and other factors associated with the reconstruction process. Often, the resolution is better in the core and it gradually worsens toward the edges or other flexible parts of the molecular assembly. The statistical analysis used in the false discovery rate (FDR) approach allows associating confidence in distinguishing molecular signals from the background and detecting weak features in the map based on the statistical significance estimate. The FDR calculation (Beckers et al., 2019) generates confidence maps with values at each voxel reflecting the fraction of voxels expected to contain molecular

signals at this threshold (the voxel value). The 1% FDR threshold (confidence map threshold of 0.99) was demonstrated to reliably discriminate voxels associated with the molecular volume from the background noise over a wide resolution range, including maps at near-atomic resolutions to 6.8 Å and the subtomogram averages in the resolution range 7–90 Å.

In this study, we present a tool for validating the backbone trace of an atomic model by estimating the confidence that the backbone atoms are in the molecular volume rather than the background. Each residue in the model are assigned scores based on the confidence map calculated using the FDR approach. We demonstrate the utility of the approach to detect mistraced residues, using datasets from the EMDB model challenges 2015/16 and 2019, and compare it against other metrics used in the field for estimating local fit to maps.

This procedure is also useful for pruning mistraced regions of the model generated by *ab-initio* modeling tools like Buccaneer (Hoh et al., 2020). Especially at areas of the map with resolution worse than 3.5 Å, it is not uncommon that the chain may be mistraced into the background. Also, Buccaneer often traces a few polypeptide fragments in the background areas with noisy features or artifacts from map reconstruction and postprocessing. These fragments are not connected with the main chains of the model and usually are only of a few residues long. Currently, there is no automated tool to locate and remove mistraced residues. Where possible, the pruned models can then be extended with one of the automated model building tools or rebuilt in an interactive tool like Coot.

Initial results indicate that our approach is effective in detecting mistraced regions of the model and for automated pruning of models as part of Buccaneer. The FDR backbone validation score assesses whether the backbone coordinates are within the molecular volume and is complementary to existing validation tools that either assess the model quality or evaluate agreement with the map. The described tool is implemented and available as a part of the CCP-EM (Burnley et al., 2017) software suite.

METHODS

The FDR backbone validation method assesses positions of the atoms in the input model based on the confidence map derived using the FDR approach (Beckers et al., 2019). For the confidence map calculation, a processed/sharpened but unmasked map is preferred. Masked maps that exclude the majority of solvent background are not useful for confidence map calculations. The procedure estimates the background noise distribution from four density cubes placed outside of the particle volume in the x, y, and z central axes by default. Each voxel of the map is then compared against the background estimate to detect significant deviations and a *p*-value is associated to quantify significance. To account for the number of voxels and their dependencies, the *p*-values are further adjusted using false discovery rate (Benjamini and Yekutieli, 2001). Each voxel is assigned with an FDR-adjusted significance score between 0 and 1, 0 refers to noise only and 1 to a clear molecular signal. A score of 0.99 indicates that a maximum

of 1% of voxels (1% FDR) is expected to be background noise, beyond this threshold.

We use the following steps to calculate the FDR backbone score:

- (1) The minimal input for the FDR-validation is the pdb or cif/mmcif format file of the atomic model and the confidence map calculated using the FDR control approach. The confidence map can be calculated using the “confidence map” implementation in the CCP-EM software suite or using a standalone installation from the source (<https://git.embl.de/mbeckers/FDRthresholding>).
- (2) The input model coordinates are extracted and mapped onto the confidence map grid by associating the voxel(s) around the atomic coordinate (within 1 Å).
- (3) Each atom of the model is then associated with the corresponding map value from the confidence map. In the default mode, the FDR backbone score of each residue is calculated as an average of map values at the coordinates of the C-alpha, C, and N atoms. We use this approach primarily to detect mistraced residues based on the positions of backbone coordinates in the map. We exclude the backbone carbonyl oxygen as they are often associated with weak map information at resolutions worse than 3 Å. This approach can be used to detect misplacement of the side chains as well, although missing map data at the ends of acidic and highly flexible side chains can lead to false detections. For nucleic acids, the average score is calculated based on the C1', C2', C3', C4', C5', O3', O4', O5', and P atoms positions. For ligands and waters, all of the atoms are taken into consideration. The users can also choose an optional validation mode based solely on the Ca positions of the residues and C1' for nucleic acids. This mode is useful with models with only Ca atoms, usually built in low-resolution cryo-EM maps.
- (4) Additionally, this tool offers an option to prune the atomic model, which can be used to automatically remove the residues with a score lower than 0.9 as well as the preceding and following residues. A model pruned this way can be used in the next stages of the iterative model building procedures, where the missing segments can be extended or rebuilt. This is useful when dealing with the *ab-initio* models from the automated model building tools. In some cases, particularly when building in areas of the map with a local resolution worse than 3.5 Å, parts of the chains can be traced into the background.
- (5) As an output, we provide a CSV format file containing the list of residues with the associated FDR backbone scores. The models after pruning will have the low scoring residues removed. They are saved in the selected folder with the original model names with a suffix “pruned” added depending on the mode used. We also provide an attribute file that can be used to associate the FDR score for each residue in the atomic model in UCSF Chimera (Pettersen et al., 2004). The model can then be colored using the FDR score attribute to identify areas with low scores.

The FDR backbone validation tool is written in Python 3. To handle the I/O model files in pdb and cif/mmcif format the GEMMI package (Wojdyr, 2017) is used. The map files are processed with the mrcfile python package (Palmer, 2016). The tool also requires NumPy (tested with v1.16.2 (NumPy v1.16 Manual' 2019)). The GUI implementation with CCP-EM software suite was done using PyQt ('PyQt 4.9.4 Reference Guide' 2011).

In this study, we compare the FDR backbone validation approach against other metrics for estimating local fit to maps. For a fair comparison, the other metrics were also calculated only on the backbone atoms of the models. The map deposited in EMDB as a “primary” map was used for the analysis, the FSC-Q score calculation also requires the half-maps.

The Q-score values for the backbone atoms were calculated using the MapQ plugin (v1.6) for UCSF Chimera, at the resolution reported for the deposited primary map.

The FSC-Q score was calculated using the tool (validate fsc-q) integrated in the Scipion v3.0.7-Eugenius. The FSC-Q value for the backbone is calculated as an average for the C, N, and Ca atoms.

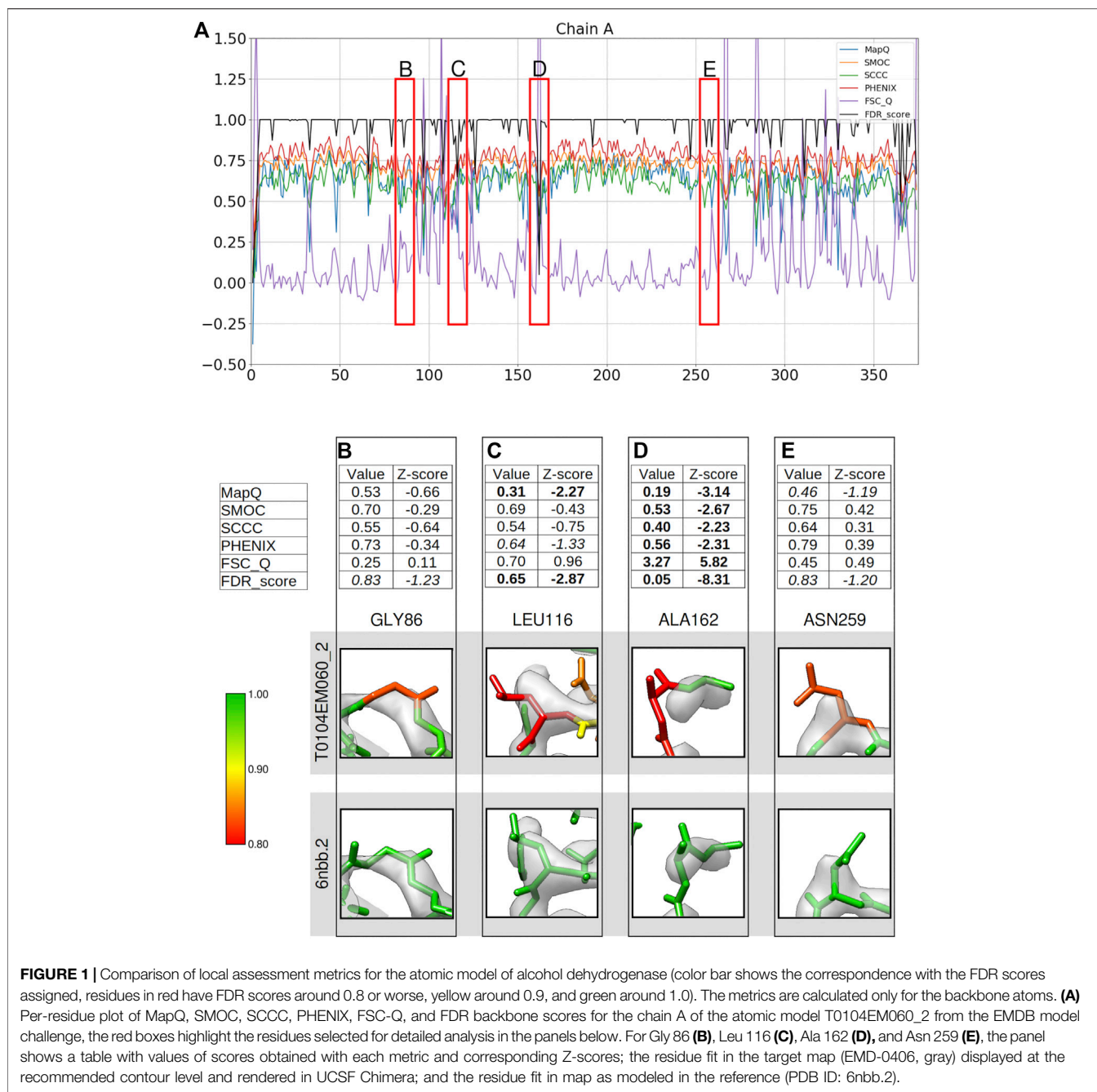
The SMOC and SCCC scores were calculated using the score_smoc.py script available from TEMPy1 in CCP-EM v1.5, for the minimal backbone of the models (C, N, and Ca atoms). The script was run with the “-distance” mode option which uses distance from the atoms for identifying voxel-covered. SMOC estimates the Manders' overlap coefficient while SCCC calculates the cross correlation coefficient.

The PHENIX map/model CCC (v1.18.2) scores were calculated on the models obtained from 10 cycles of atomic B-factor refinement with REFMAC5 (Murshudov et al., 2011) (using the keyword option “refi bonly”). This was done to ensure that the atomic B-factors are refined as PHENIX uses the atomic B-factors as part of the map calculation from the atomic model.

The box size of the input map was trimmed wherever possible to improve the speed of the computations. The FDR-validation requires a sharpened but unmasked map, with the background features present in order to estimate the noise distribution.

RESULTS

To demonstrate the application of our approach, we used the following examples, the majority of which are models submitted to the EMDB model challenges for target maps resolved at a range of resolutions. In each case, we compare the FDR backbone scores against other metrics that estimate local fit to map. Using a set of residues detected as “mistraced” by the FDR backbone score, we assess agreement with other scores and also highlight cases where there is a disagreement. We use the reference model from the model challenge to compare the backbone conformation and fit to map. Please note that the reference model does not always have the best fit to map for all residues in the model, and often several of the models submitted to the challenge have a better fit (Lawson et al., 2021). In some cases, there are obvious backbone misfits in the reference model, as discussed below. In such cases, we also compare the model of interest against other models reported with



a higher rank higher in the model challenge based on a number of validation metrics (<https://model-compare.emdataresource.org/>).

Alcohol Dehydrogenase (2.9 Å, Target T0104)

We computed per-residue backbone scores based on different metrics for the chain A of model T0104EM060_2 submitted to the Model Challenge 2019 for the target alcohol dehydrogenase map (EMD-0406) resolved at 2.9 Å resolution (Herzik et al. 2019; **Figure 1A**). We checked residues either associated with lower

confidence scores (0.95 or lower) or where the scores disagree in detecting a mistrace, and compared against the reference model used in the model challenge (PDB ID: 6nbb). The reference structure has 10 models representing local conformational variability. We chose the second model (6nbb.2) for our analysis as it has a relatively better fit with the map when inspected in UCSF Chimera, for cases we discuss in **Figures 1, 2**, especially at the N-terminus (**Figure 2B**). The metrics used for comparison includes Q-score, SMOC, SCCC, PHENIX local CCC, and FSC-Q, calculated only for the backbone atoms (see Methods). The residues highlighted in red boxes in **Figure 1A**

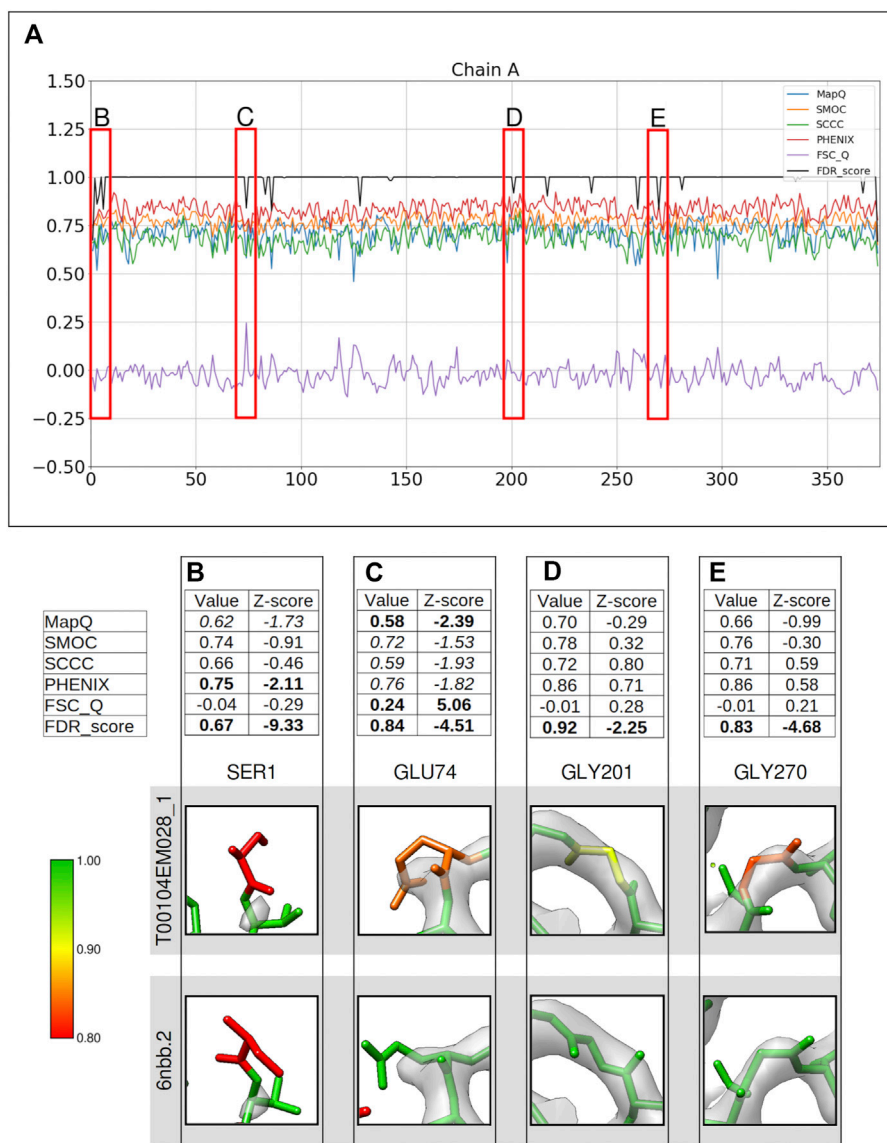


FIGURE 2 | Comparison of metrics for the atomic model of alcohol dehydrogenase (color bar shows the correspondence with the FDR scores assigned, residues in red have FDR scores around 0.8 or worse, yellow around 0.9, and green around 1.0). The metrics are calculated only for the backbone atoms. **(A)** Per-residue plot of the scores from MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the chain A of the atomic model T00104EM028_1 from the EMDB model; the red boxes highlight the residues selected for detailed analysis in the panels below. For Ser1 **(B)**, Glu 74 **(C)**, Gly 201 **(D)**, and Gly 270 **(E)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-0406, grey) displayed at the recommended contour level and rendered in UCSF Chimera; and the residue fit in map as modeled in the reference (PDB ID: 6nbb.2).

indicate some of the regions where the scores differ. We provide a detailed analysis of these residues. **Figures 1B–E** show a table with the values of each metric and corresponding Z-scores, along with a snapshot of the residue colored by the FDR backbone score. Also, the corresponding view of the residue from the reference model 6nbb.2 ((Herzik et al., 2019), model 2) is provided. The models are overlaid with the deposited cryoEM density map EMD-0406 (Herzik et al., 2019) and rendered at the author-recommended contour level 0.02 (0.6σ).

Compared to a few other models submitted to the model challenge, the model T00104EM060_2 is ranked lower by the

validation metrics used in the challenge (https://model-compare.emdataresource.org/2019/cgi-bin/em_multimer_results.cgi?target_map=T00104emd_0406). Plot of per-residue backbone scores for the chain A (**Figure 1A**), also shows that many residues in this model are associated with lower scores (drops in the plot). We investigated a few residues including cases where the metrics disagree. Gly86 is highlighted as a potential mistrace by the FDR backbone score of 0.83 (**Figure 1B**). The residue in the reference model has a high FDR score (0.991) and the backbone shows better fit to map with a different conformation involving a shift and differences in backbone dihedrals. Z-scores computed for

different metrics reflect that none of the other scores identify this mistrace with any significance (absolute value of Z-scores < 1). Note that the Z-scores for FDR backbone assessment are less reliable, especially because the majority of the residues often have a score of 1.0 and the distribution is not close to normal. We recommend using the absolute values of this score to detect potential mistraces.

Figure 1C shows Leu116 associated with a low FDR backbone score of 0.65. It can be seen that the backbone C α and C are out of the map at the recommended contour level. In comparison, the reference model shows a better fit of backbone atoms. The mistrace is also detected by the MapQ score (0.31, Z-score -2.27), while PHENIX-CCC (0.63, Z-score -1.33) and FSC-Q (0.70, Z-score 0.96) have lower scores but associated with relatively low significance (Z-scores of -1.33 and 0.96, respectively).

Another residue Ala162 is located at a relatively disordered or low-resolution area of the map (**Figure 1D**), and the backbone is partly out of the map contour when compared to the reference model. All the scores identify the mistrace with significance (absolute Z-scores > 2.0), and the residue is associated with a low FDR backbone score of 0.05. Even though the map at the recommended contour level does not fully support the backbone, the reference model shows a better fit and has an FDR backbone score of 0.978. This reflects that the FDR backbone score detects voxels covering molecular volume even in the low resolution areas of the map.

Figure 1E shows Asn259 associated with an FDR backbone score of 0.83 with part of the backbone outside the contoured map. The reference model shows a better fitted backbone conformation (**Figure 1E**). MapQ also points to the potential mistrace in the submitted model with a Q-score of 0.46 although with a less significant Z-score of -1.19. The other metrics fail to identify this issue with the backbone fit. Hence, in comparison to other metrics tested in this study, the FDR backbone score detects cases of mistrace where one or more backbone atoms are displaced into background noise.

Figure 2 presents a similar analysis of the model T0104EM028_1 submitted to the same target map. Ser1 at the N-terminus of chain A is associated with an FDR backbone score of 0.67, clearly indicating a potential mistrace. Ser1 is associated with a disordered area of the map with no prominent map information at the recommended contour (**Figure 2B**). PHENIX_CC (0.75, Z-score -2.11) and MapQ (0.62, Z-score -1.73) scores also suggest poor agreement with data. The map trace is more obvious at a lower contour level (**Supplementary Figure S1**), and the terminal N atom is outside the map even at this level. Hence, there is less confidence associated with the backbone atom positions and this is also highlighted by PHENIX_CC and MapQ scores.

Figure 2C highlights Glu74 with both backbone and side chain atoms out of the recommended contour. The residue, as modeled in the reference, shows better fit with backbone atoms (and most of the side chain) inside the recommended contour. The lack of map information for the end of side chain is a common trait observed in cryo-EM maps for negatively charged side chains. FSC-Q and MapQ indicate a backbone mistrace with

Z-score values less than -2.0 (>2.0 in case of the FSC-Q score). The other metrics also highlight this, although with a relatively lower significance (Z-score < -1.5).

Gly201 is associated with an FDR validation score of 0.92. Other scores do not seem to indicate mistrace with any significance (all Z-scores were between -1 and 1) (**Figure 2D**). This residue has a different backbone conformation in the reference model and is associated with an FDR backbone score of 1.0. The backbone has a better fit in the reference with all atoms except carbonyl oxygen inside the recommended contour. Another case where only the FDR-validation score detects a mistrace of backbone is Gly270, where the reference model shows a better fit with the map with a slight shift in atom positions (**Figure 2E**). The backbone residue shifted outside of the map density is presented in **Figure 2E**. These cases highlight that the FDR backbone score can work in complementarity to the scores that quantify agreement with the map.

The structure of alcohol dehydrogenase has zinc ions bound but the ions are not modeled in all of the structures submitted to the model challenge. **Supplementary Figure S2A** presents a comparison of two models submitted (Model Challenge IDs: T0104EM010_1 on the left panel and T0104EM028_1 on the center) where the zinc atoms are modeled, along with the reference model (PDB ID: 6nbb) on the right. In the model T0104EM010_1, the zinc atom is highlighted as a potential misfit based on our approach, and no obvious map data can be seen at this position. It can be seen that the ligands in both the model T0104EM028_1 and the reference structure are placed in a position justified by map density and supported by the higher FDR backbone scores. It is worth mentioning that many of the automated model building software do not support ligand fitting, and therefore this is often done interactively. The presented validation technique can be useful for validating the modeled ligands in cryo-EM maps.

T20s Proteasome (2.8 Å)

Another set of models used for the evaluation of the FDR backbone validation approach were chosen from those submitted to the model challenge for the target map of the T20s proteasome (EMD-6287), resolved at 2.8 Å resolution. **Figure 3A** presents the comparison of scores from difference metrics obtained for the chain L of the model T0002EM133_1. Again, the areas where the scores disagree were inspected closely.

Several residues in this model are associated with lower score values as evaluated by different metrics (**Figures 3B–E**). In this case, the Z-scores are less meaningful as the distribution of scores is likely to deviate significantly from normal because of the presence of several low-scoring residues (outliers). Therefore, we considered a less stringent absolute Z-score cutoff of 1.0 to associate significance to the scores. Again, as the FDR scores do not follow a normal distribution (often many residues have a score of 1.0 and a few scoring lower), the Z-scores are less useful. We recommend using the absolute FDR scores to detect potential mistraces.

Val14 is associated with a low FDR score of 0.83, also supported by a lower MapQ score of 0.32 (Z-score -1.73)

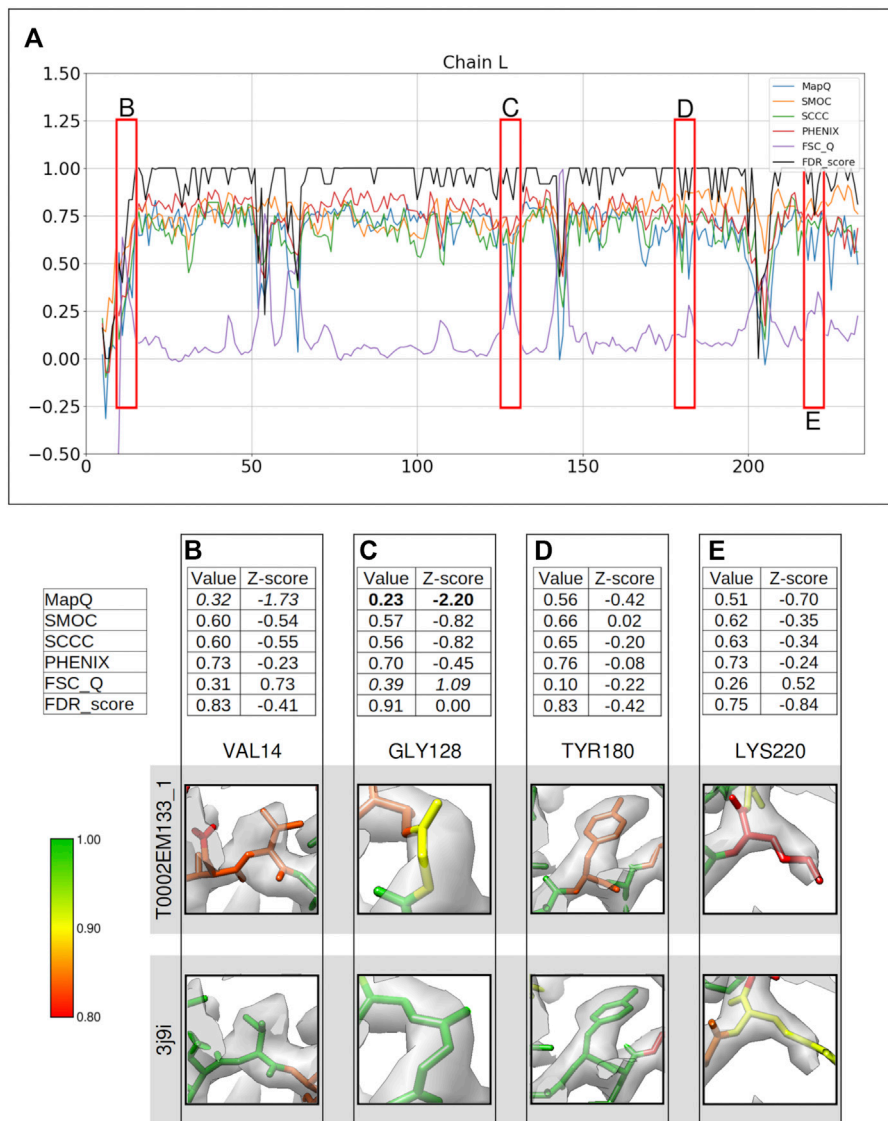


FIGURE 3 | Comparison of metrics for the atomic model of T20s proteasome, calculated only for the backbone atoms. **(A)** Comparison of the per-residue scores from MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the chain L of the atomic model T0002EM133_1 from the EMD model; the red boxes highlight the residues selected for detailed analysis in the panels below. For Val 14 **(B)**, Gly 128 **(C)**, Tyr 180 **(D)**, and Lys 220 **(E)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-6287, grey) displayed at the recommended contour level 0.025 (3.3σ) and rendered in UCSF Chimera; and the residue fit in map as modeled in the reference (PDB ID: 3j9i).

(Figure 3B). The backbone N atom is out of the map contoured at the recommended level. The reference model (PDB ID: 3j9i) shows a better fit and has an FDR score of 1.0. Hence, the FDR backbone score and MapQ detect the mistrace with a greater significance compared to other metrics.

Figure 3C shows Gly128 which has been scored lower by MapQ (0.23, Z-score: -2.20), and FSC-Q has a score of 0.39, although with a relatively less significant Z-score of 1.09. The backbone of this residue is associated with an FDR backbone score of 0.91. Visual inspection of the backbone shows that the Ca and carbonyl C atoms are partly out of the contoured map. The

reference model shows a better fit of this residue with a higher FDR score of 0.99.

Tyr180 is assigned a low FDR backbone score of 0.83 **(Figure 3D)**, and the other scores do not highlight a backbone misplacement with all absolute Z-score values less than 0.5. The backbone N atom of the modeled residue is partly outside the contoured map. The reference model (chain B) shows a better backbone and a side chain fit and has an FDR backbone score of 0.99. Hence, the FDR score detects backbone misplacements compared to other metrics used in this study and is thus effective in identifying mistraced residue backbone.

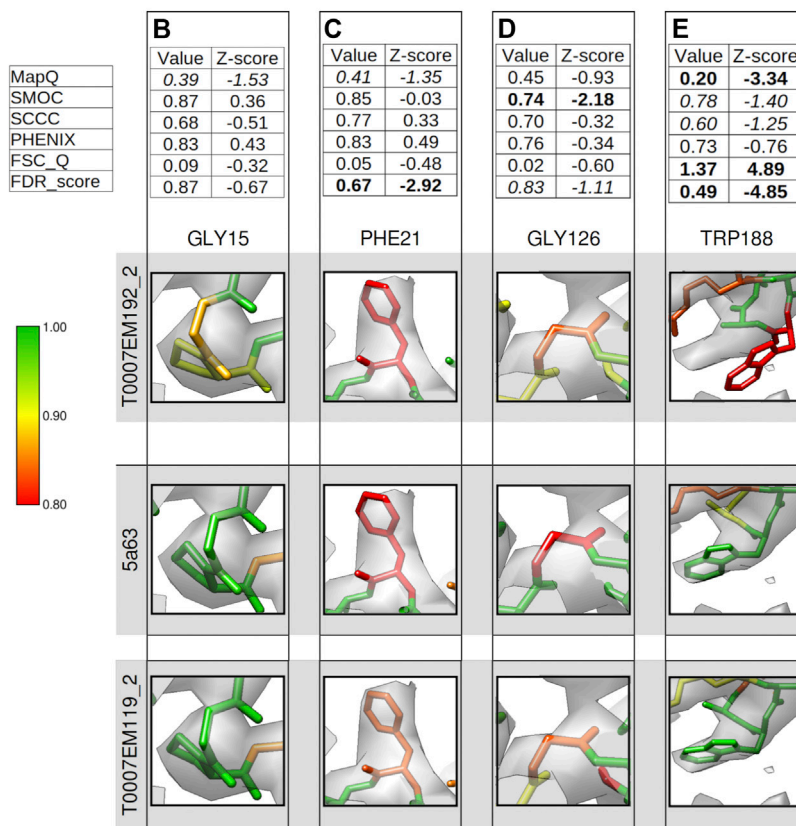
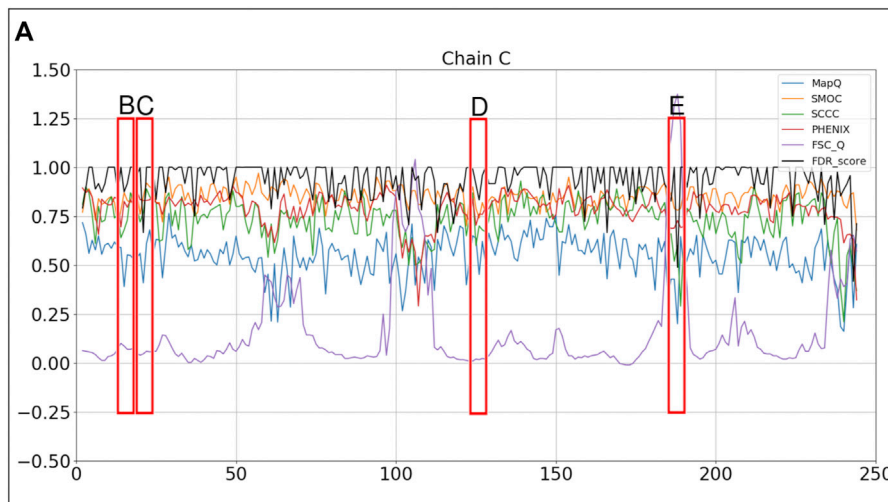


FIGURE 4 | Comparison of metrics for the atomic model of γ -secretase, calculated only for the backbone atoms. **(A)** Comparison of per-residue scores from MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the chain C of the atomic model T0007EM192_2 from the EMD model challenge; the red boxes highlight the residues selected for detailed analysis in the panels below. For Gly 15 **(B)**, Phe 21 **(C)**, Gly 126 **(D)**, and Trp 188 **(E)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-3061, grey) displayed at the recommended contour level and rendered in UCSF Chimera (first row); the residue fit in map as modeled in the reference (PDB ID: 5a63, center), and the fit of model T0007EM119_2 in the map (last row).

This is another case where the scores disagree is Lys220 **(Figure 3E)**, which is associated with a low FDR backbone score of 0.75, while the other scores do not highlight a mistrace with significance. The MapQ score is relatively lower with a value of 0.51 (Z-score: -0.70). A closer inspection and

comparison with reference suggests that the residue has a better placement in the reference with a shift of backbone atoms accompanied by better positioning of side chain. The residue backbone in the reference was assigned an FDR score of 0.92 and the residues on either side of Lys220 also score low. This

highlights the possibility of further improvement of backbone atom placement in this segment of the reference.

γ -Secretase (3.4 Å, Target T0007)

The FDR backbone validation tool was used to assess another model (Model Challenge ID: T0007EM192_2) submitted to the EMDB Model Challenge 2015/2016 (Lawson and Chiu, 2018) for the gamma secretase map EMD-3061, solved at 3.4 Å resolution (Bai et al., 2015). **Figure 4** shows the comparison of different scores associated with residues in the chain C of the model. **Figures 4B–E** provide a closer look into some of the areas of the model where the scores disagree. As discussed below, the reference model (PDB ID: 5a63) does not show a better fit for most of these cases. Hence, we also compared the backbone fit against T0007EM119_2, which is another model submitted to the model challenge for this target and ranked higher than the reference by multiple metrics used in the challenge.

Gly15 is associated with an FDR backbone score of 0.87 (**Figure 4B**) and MapQ also associates a low score of 0.39 with the backbone (Z-score: -1.53). Other scores do not highlight a mistrace of backbone atoms for this residue. Visual inspection shows that the N and C α atoms are outside the map at the recommended contour. In the reference model (PDB ID: 5a63), the backbone shows a slight shift of the backbone toward the map volume. In the model T0007EM119_2, which scored higher than the reference in the model challenge, the atoms are shifted well into the map and Gly15 has an FDR score of 1.0. Hence, the slight backbone misplacement is highlighted by FDR and MapQ scores in this case.

Phe21 is also associated with a low FDR validation score of 0.67 and MapQ associates a relatively lower Q-score of 0.41 (Z-score: 1.35) (**Figure 4C**). The reference model shows similar backbone atom positions but associated with a lower FDR backbone score (0.58). Upon closer inspection of the model at a higher contour level, we find that the carbonyl C atom is out of the map. In the model T0007EM119_2, the residue shows a slightly better fit with the backbone shifted into the map, and has an FDR backbone score of 0.83. Multiple metrics (the FDR score and MapQ) point to a potential backbone misfit and further investigation is required in this case to establish this and check for improvement upon refitting.

Figure 4D shows Gly126 with the C α atom outside the map at the recommended contour. The modeled residue is detected as potential mistrace with an FDR backbone validation score of 0.83 and a lower SMOC score of 0.74 (Z-score -2.18). Other scores do not highlight this with a significant Z-score. The backbone of Gly126 in the reference model (PDB ID: 5a63) is also partly outside the contoured map and associated with a lower FDR score (0.75). In the model T0007EM119_2, a similar scenario was found where the C α atom is partly outside the map contour. Both the FDR score and SMOC identify a misfit in this case reflecting a potential for improvement of the backbone fit. In the absence of a good reference fit, further investigation and refitting is required to confirm the backbone misplacement.

Another residue associated with a low FDR backbone score of 0.49 is Trp188 (**Figure 4E**). The backbone mistrace is evident in this case when compared to the reference structure (PDB ID:

5a63), where Trp188 is better fitted in the map (**Figure 4E**) and has an FDR backbone score of 1.0. FSC-Q and MapQ scores also detect the backbone misplacement with significant Z-scores. The model T0007EM119_2 also shows a well-fitted backbone with an FDR score of 0.995. Hence, in this case, the FDR score works in complementarity with the metrics that calculate CCC or similar (SMOC).

Supplementary Figure S2B highlights another segment of the model (T0007EM192_2) with a polysaccharide, where the atomic positions in the terminal monosaccharide units have relatively lower FDR scores. The FDR validation score is calculated as an average of scores of the atoms in each unit. These terminal units of the carbohydrate are expected to be more flexible and the range of values of the score reflects this as well, suggesting higher uncertainty of the positions at the edges for being associated with molecular signals. The terminal monosaccharide unit in the reference model (PDB ID: 5a63) is also associated with lower FDR validation scores.

RNA Polymerase Complex From SARS-CoV-2 (2.5 Å)

We also applied our approach to assess the atomic model (PDB ID: 7bv2) deposited with the recently published structure of RNA polymerase complex (EMD-30210, 2.5 Å) from SARS-CoV-2 virus (Yin et al., 2020). A few residues in the model have lower confidence scores assigned (**Figure 5A**).

Figure 5 shows a comparison of the validation metrics for residues in the chain B of the model. The FSC-Q score was not calculated for this case as the half maps were not available from EMDB. **Figures 5B–D** provide insights into selected regions of the model, fitted in the map contoured at the recommended level 0.058 (4.3σ). At this contour the map data corresponding to most of the backbone of low-scoring residues at the N-terminus is disconnected, possibly indicating relatively lower local resolutions. We also assessed the backbone atom placement at a lower contour level (0.035) (**Figures 5B–D**, second row). To check whether the disconnected map data is due to local over-sharpening (often resulting from a global sharpening factor applied to the map), we calculated a locally sharpened map using LocScale (Jakobi et al., 2017) implemented in the CCP-EM software suite (**Figures 5B–D**, last row).

Figure 5B shows Val83 highlighted as a potential mistrace by the FDR score with a value of 0.67 and MapQ (0.45, Z-score -2.30). The poor quality of fit is also indicated by other metrics including SMOC (0.69, Z-score -1.27), SCCC (0.51, Z-score -1.80), and PHENIX (0.58, Z-score -1.81). It can be seen that this residue backbone is not fully supported by the map even at a lower contour level (**Figure 5B**, second row) and the backbone peptide N atom is partly out of the map. As expected, the locally sharpened map from LocScale is less disordered with the peptide N atom at the edge of the map contour. The peptide N has a low FDR score of 0.0 compared to C α and carbonyl C which have scores of 1.0. In this case, the backbone is likely to be misplaced as highlighted by multiple scores.

A similar case involving Leu98 is presented in **Figure 5C**. Leucine 98 scores low with all other metrics (Z-scores lower than

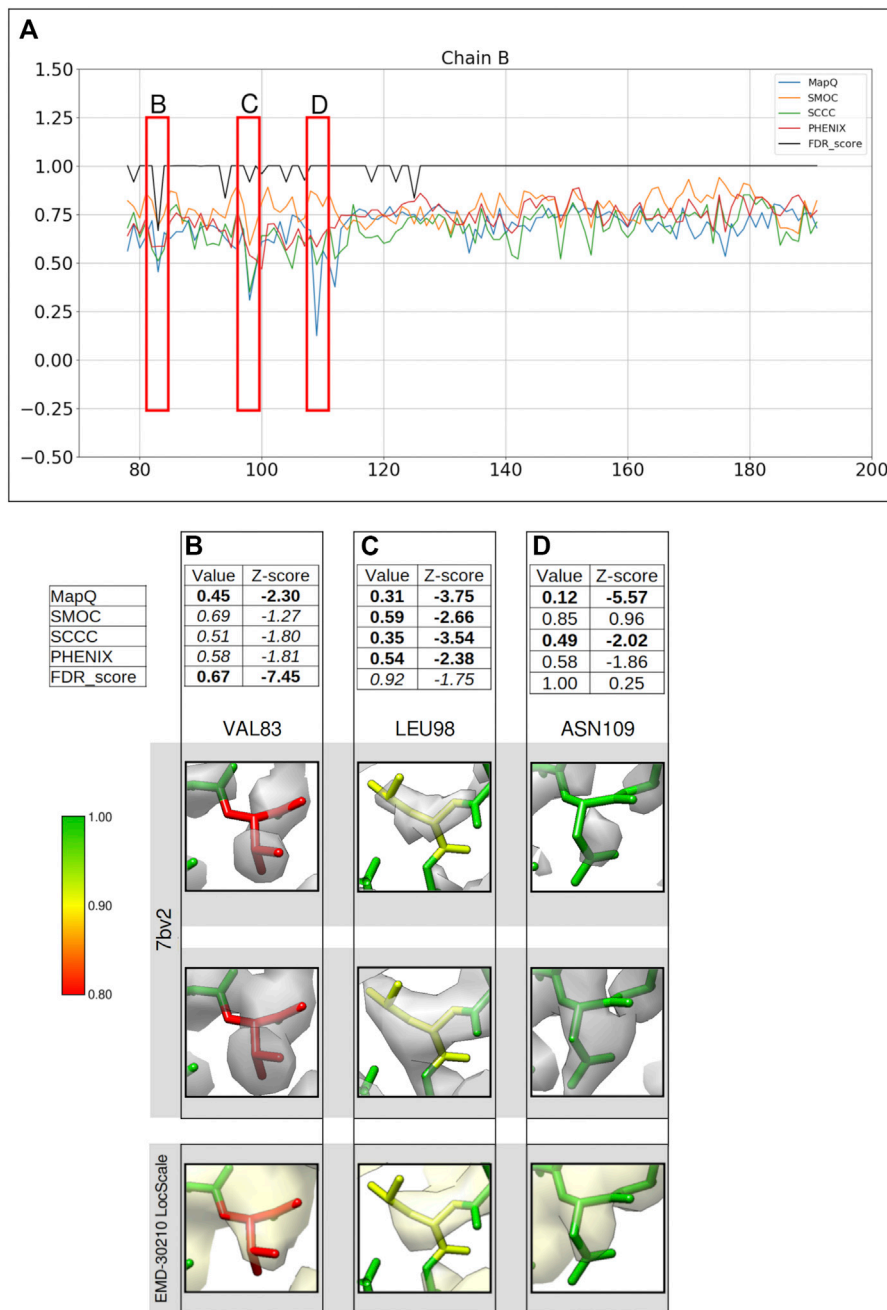


FIGURE 5 | Comparison of backbone validation metrics for the atomic model of the RNA polymerase complex (PDB ID: 7bv2). **(A)** Comparison of the per-residue scores from MapQ, SMOC, SCCC, PHENIX, and FDR backbone score for the chain B of the atomic model, the red boxes highlights the residues selected for detailed analysis in the panels below. For Val 83 **(B)**, Leu 98 **(C)**, and Asn 109 **(D)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-30210, grey) displayed at the recommended contour level and rendered in UCSF Chimera; the residues fit in the map rendered at a lower contour level.

-2.0) and this residue has an FDR score of 0.92. The position of carbonyl C atom is not fully supported by the map even at a lower contour and this atom has an FDR score of 0.75. In this case, multiple metrics highlight a potential backbone misfit and require further investigation to explore the possibility of improving the fit. The locally sharpened map also shows disconnected map trace

at the selected contour, with the carbonyl C atom placed outside the contour. In the absence of a good reference fit for this residue, Asn109 on the other hand, is highlighted as a misfit by MapQ (0.12, Z-score -5.57), SCCC (0.49, Z-score -2.02), and PHENIX (0.58, Z-score -1.86) **(Figure 5D)**. However, the FDR validation score assigns a value of 1.00 (Z-score 0.13) for this residue

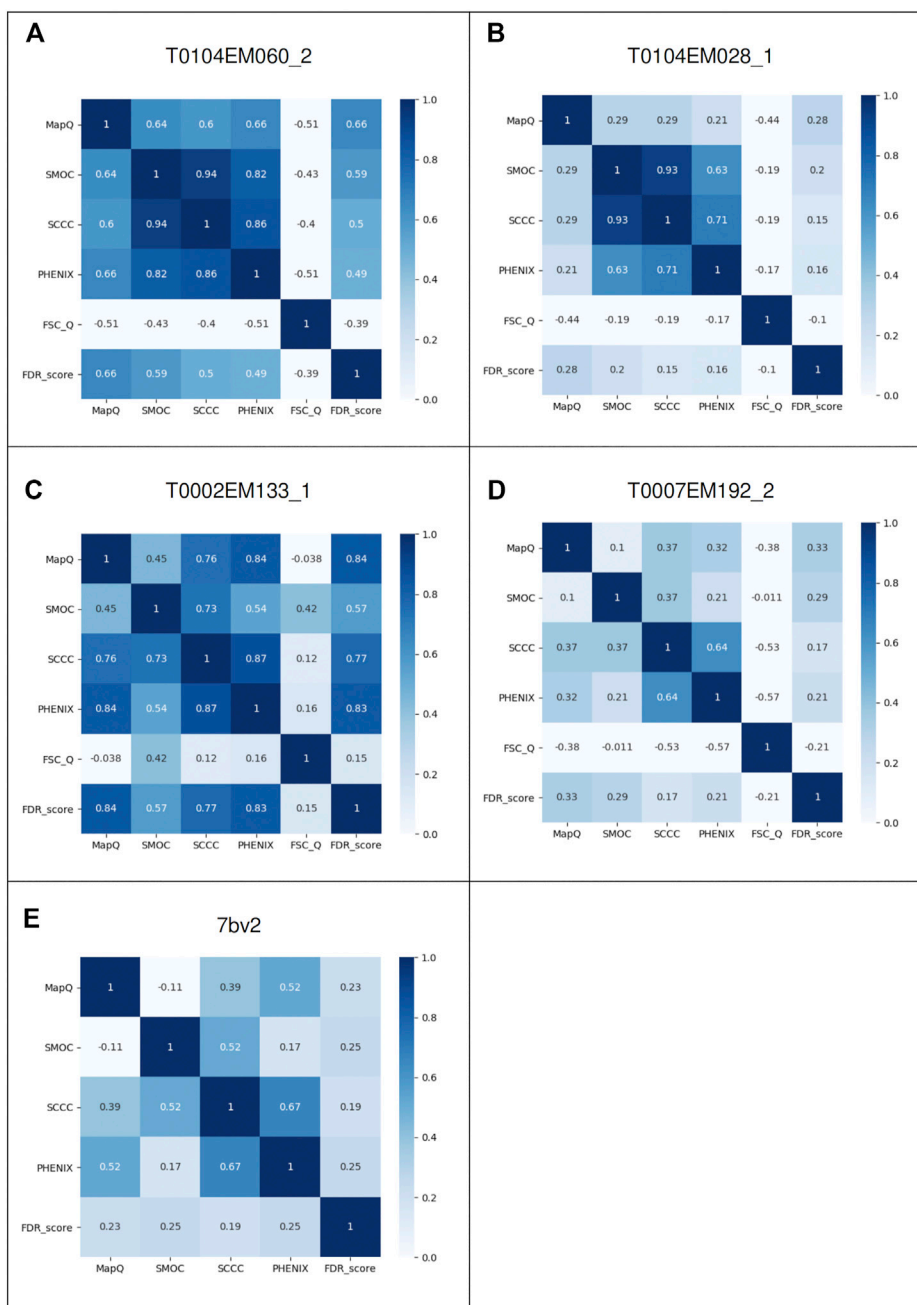


FIGURE 6 | Pairwise correlations of different metrics: MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the atomic models: **(A)** Chain A of alcohol dehydrogenase T0104EM060_2, **(B)** chain A of T0104EM028_1, **(C)** chain L of T20s proteasome T0002EM133_1, **(D)** chain C of γ -secretase T0007EM192_2, and **(E)** chain B of the RNA polymerase complex model 7bv2.

backbone, suggesting no mistrace of the backbone. A closer inspection of this position in the model shows that the lower contour level covers most of the backbone atoms of the residue, except carbonyl C atom where the map is still disconnected. The locally sharpened map is smoother with no disorder and shows a better coverage of backbone atoms. The residue is located at a low resolution area of the map and

it is likely that the backbone is within the molecular volume but the atoms are misfitted, as highlighted by the other metrics.

Supplementary Figure S2C shows an area of the modeled RNA where the terminal nucleotide has a lower FDR score. The map density is also disordered at this position likely due to the higher flexibility of this part of the RNA.

TABLE 1 | Outlier detection by different metrics for ten of the models submitted to the 2019 Model Challenge for the target alcohol dehydrogenase map (EMD-0406). For each model, the table number of residues of chain A associated with Z-scores lower than -2.0 . For the FDR backbone score, the table also shows the number of residues with an FDR backbone score less than 0.9

ModelID	CC	Method	FDR_score <0.9	FDR_score	MapQ	SMOC	SCCC	PHENIX	FSC-Q	FSC-Q
				Z-score < -2	Z-score < -2	Z-score < -2	Z-score < -2	Z-score < -2	Z-score > 2	Z-score < -2
T0104EM035_1	0.32	<i>Ab-initio</i>	3	6	15	9	8	8	12	7
T0104EM027_1	0.32	<i>Ab-initio</i>	8	7	15	11	9	10	10	6
T0104EM010_1	0.32	<i>Ab-initio</i>	7	9	14	8	13	6	11	7
T0104EM041_1	0.32	<i>Ab-initio</i>	10	10	10	9	13	9	13	5
T0104EM090_1	0.31	<i>Ab-initio</i>	21	19	12	13	12	12	12	0
T0104EM028_1	0.31	<i>Ab-initio</i>	9	15	13	12	10	10	14	2
T0104EM025_1	0.31	Optimized	8	9	12	17	14	11	14	5
T0104EM082_1	0.31	<i>Ab-initio</i>	9	9	15	18	18	12	12	2
T0104EM060_2	0.28	<i>Ab-initio</i>	39	66	17	14	14	16	8	0
T0104EM054_1	0.27	<i>Ab-initio</i>	83	26	17	16	18	17	18	3

Correlation Between Different Metrics

In the cases discussed above, we show a number of cases where different metrics disagree in the detection of backbone mistrace and cases where the FDR backbone scores can be complementary. To check how different metrics rank models based on local backbone fit to maps, we scored ten of the models submitted to the 2019 Model Challenge for the target alcohol dehydrogenase map (EMD-0406), and nine of them were built using *ab-initio* model building approaches. For each model, the number of residues of chain A associated with a Z-score lower than -2.0 were counted (Table 1). The table also shows the number of residues with the FDR backbone score less than 0.9 . The models in the table are sorted by the global CCC scores derived from the assessment results of the model challenge (https://model-compare.emdataresource.org/2019/cgi-bin/em_multimer_results.cgi?target_map=T0104emd_0406). No two metrics completely agree in the ranks assigned to the models based on the number of potential backbone misfits. However, there is a general agreement on best scoring the models and those with the lowest ranks. Note that the Z-scores are less meaningful in cases where the distribution of the score is far from normal. This is expected to affect the ranks, especially in the case of FSC-Q and MapQ where the outliers have significantly lower scores than the rest of the distribution.

To further investigate the pairwise agreement between metrics, we computed pairwise correlations between scores for the case studies discussed above. Figures 6A–E present the correlation matrices highlighting pairwise correlations between metrics for each case (corresponding to Figures 1–5). In a general scenario where an atomic model fits well overall in a map but includes a few mistraced residues, the majority of the residues have FDR scores of 1.0 and we expect lower scores for mistraced residues. Hence, the FDR score being less variable relative to other scores, the pairwise correlations involving the FDR score are expected to be low. Indeed, we observe this for most of the models except for T0104EM060_2 and T0002EM133_1 where many of the residues are associated with low backbone scores (Figure 4). In these two cases, the FDR score shows better correlation with MapQ with pairwise correlation coefficients of 0.66 and 0.84 ,

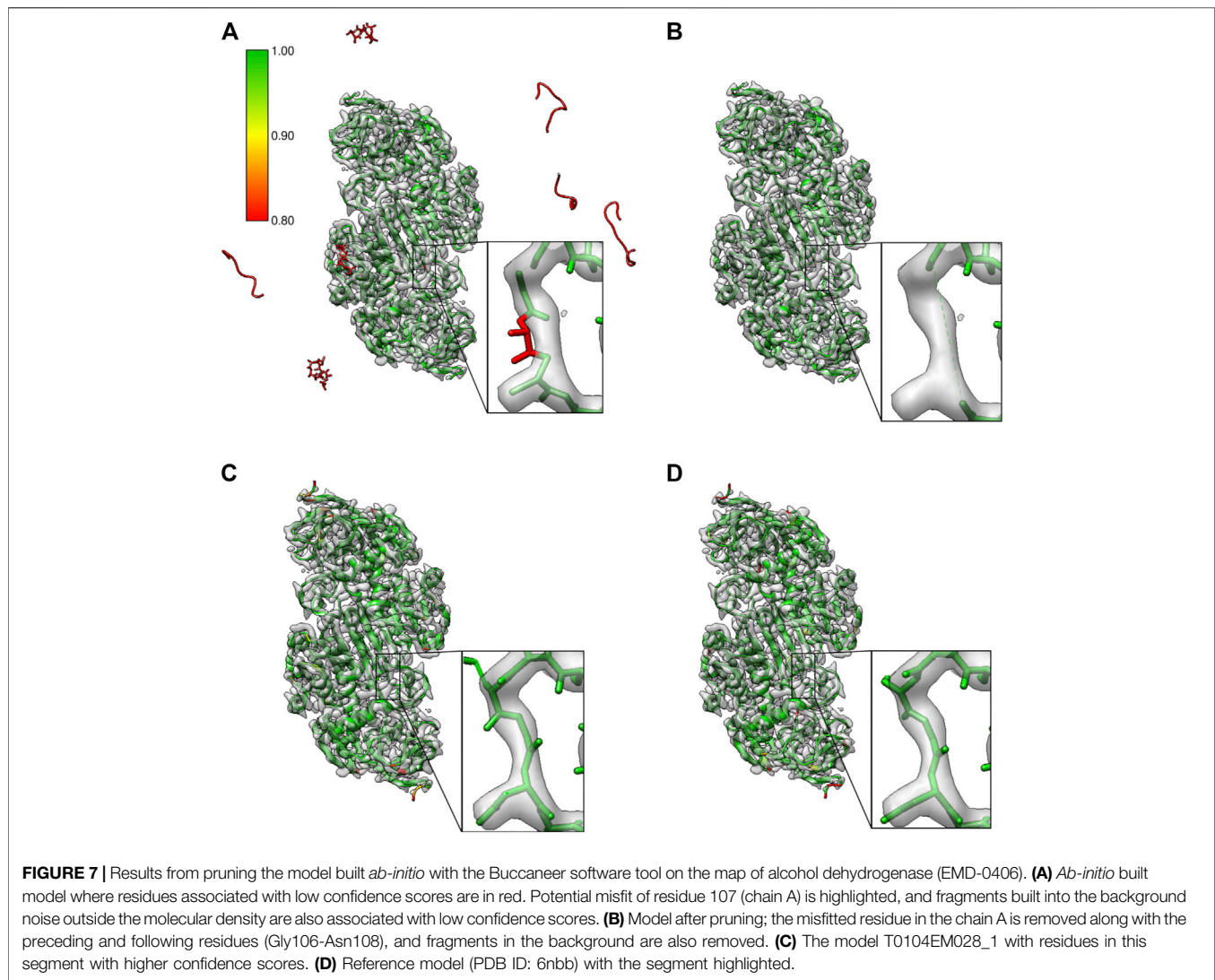
respectively. MapQ scores also correlate with SCCC and PHENIX_CC scores for these two cases.

Overall, SCCC and PHENIX CCC show a good correlation in most cases with pairwise correlations in the range 0.64 to 0.87 , which is expected as both scores involve calculation of cross correlation coefficient. SCCC and SMOC scores are largely correlated as well with pairwise correlations spread between 0.37 and 0.94 . These two scores use similar underlying procedures for synthetic map generation from model and identification of voxels covered by atoms. FSC-Q does not correlate with any of the other scores as the score reflects the model-map (and map-model) differences, unlike the other scores.

Pruning *Ab-Initio* Built Models

The proposed approach was used to prune models generated by Buccaneer (Cowtan, 2006) which is an *ab-initio* model building tool that works by an iterative process involving finding backbone seed positions, growing them to fragments, connecting and pruning fragments to chains and pruning the resulting chains. Often the final model from Buccaneer needs to be pruned interactively in Coot to remove any fragments and fix any obvious mistraces. Identifying parts of the model that are fitted into low confidence regions of the map enables automated pruning of the models.

We tested this using the *ab-initio* model built using the Buccaneer software for the 2.9 \AA reconstruction of alcohol dehydrogenase (EMD-0406). Figure 7A shows the model built from four Buccaneer cycles. The confidence map-based approach identifies fragments built into the background noise outside the molecular density (highlighted in red). The zoomed area provides a closer look at the loop where one of the residues is mistraced and backbone atoms are out of the contoured map. Figure 7B shows the same model after pruning based on our approach. All the fragments and mistraced residues were removed. Residues on either side of the low scoring residue are also removed while pruning. This helps to rebuild this whole region in the next round of the automated model building. Figures 7C,D shows the confidence scores for the same segment from the model T0104EM028_1 and the reference model (6nbb.2), respectively. The residues of these models have higher



confidence scores and the residues are fitted better in the contoured map.

DISCUSSION

The majority of cryo-EM reconstructions in EMDB are determined at resolutions worse than 3 Å and often the local resolution varies significantly in maps that are otherwise resolved at higher resolutions on an average. Hence, the chances of errors in the model are higher and validation tools that can detect errors and areas with high uncertainty, are necessary. In this study we tested an approach that evaluates the backbone trace of atomic models based on the local molecular signal (compared to background noise) in the map. The confidence scores calculated per voxel from the original map using the FDR control approach (Beckers et al., 2019) are mapped to individual backbone atoms in the model.

For the purpose of testing the approach, we used examples covering a range of reported resolutions from 2.5 to 3.4 Å. The residue backbones that have an FDR score less than 0.9 are included in **Supplementary Table S1**. Most of these models are built using model building and refinement tools commonly used in the field, as part of the EMDB model challenges. These challenges act as platforms to assess models derived using a wide range of modeling techniques and compare metrics which can be used to evaluate these atomic models. It also provides a repository of models built from a range of map targets and a reference model to compare against, which can be extremely useful for development and testing of new validation software.

We show that the FDR backbone score is complementary to existing model evaluation tools. The proposed score evaluates only the atomic positions and not the model agreement with the map. Hence, it is not useful for detecting any misfits within the molecular contour. Also, the current implementation of the score does not identify side chain rotamer misfits. However, as seen in

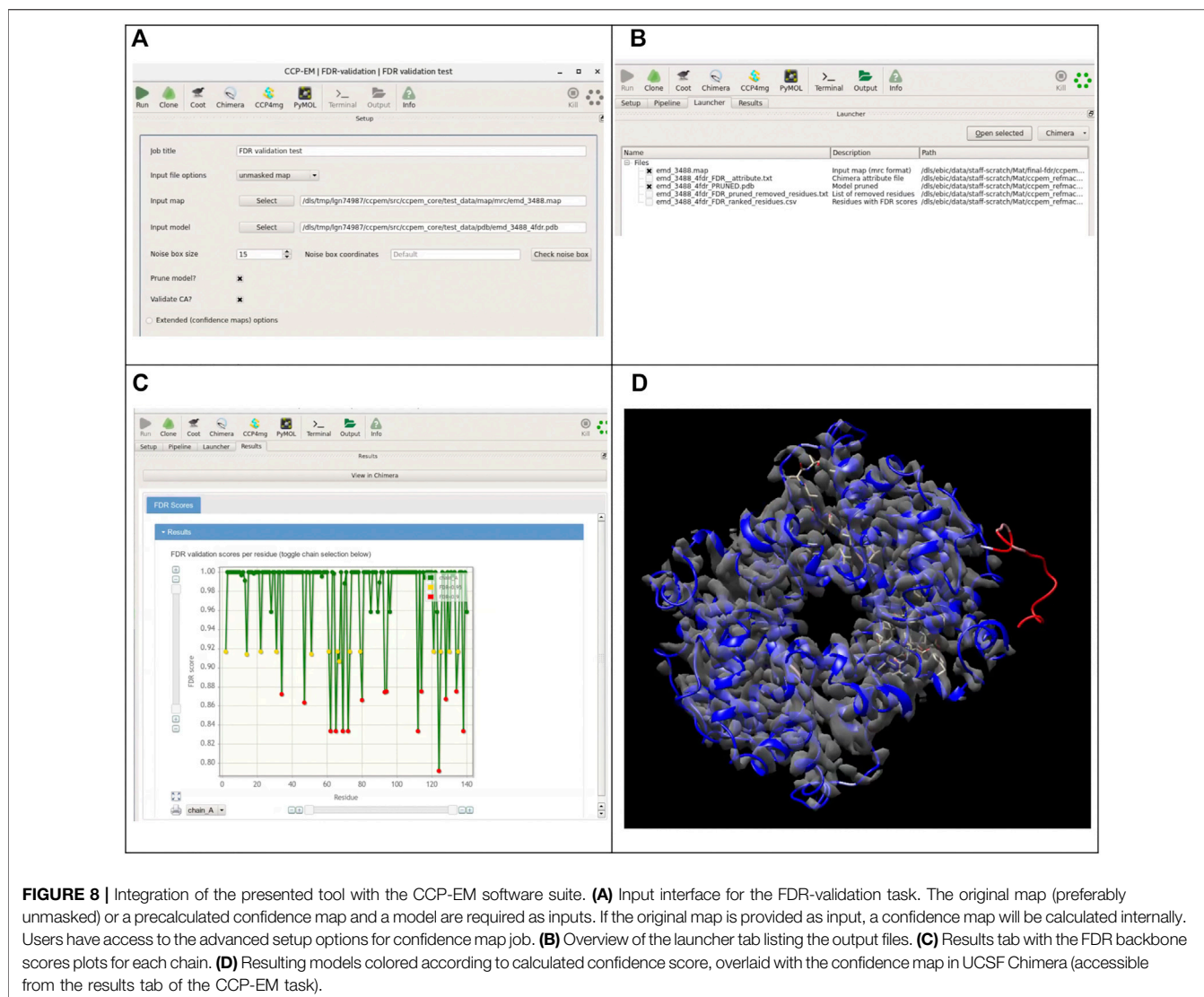


FIGURE 8 | Integration of the presented tool with the CCP-EM software suite. **(A)** Input interface for the FDR-validation task. The original map (preferably unmasked) or a precalculated confidence map and a model are required as inputs. If the original map is provided as input, a confidence map will be calculated internally. Users have access to the advanced setup options for confidence map job. **(B)** Overview of the launcher tab listing the output files. **(C)** Results tab with the FDR backbone scores plots for each chain. **(D)** Resulting models colored according to calculated confidence score, overlaid with the confidence map in UCSF Chimera (accessible from the results tab of the CCP-EM task).

many of the cases discussed in results, often backbone mistraces are associated with side chain misfits as well.

On the other hand, as demonstrated in results, some residues where one or more atoms are fitted into the background noise may still have fit-to-map scores within tolerable limits. This misplacement of backbone atoms is evident when compared to the reference, where a better backbone fit can be found. In such cases, the FDR backbone score works in a complementary manner.

Potential backbone mistraces involving a number of glycine residues were detected by the FDR backbone score (see Results), and not by other metrics. One explanation could be that glycine is often seen in flexible loops associated with low-resolution areas of the map, and some of these scores are sensitive to map resolution. In general, limiting the score calculations to backbone atoms, might also affect some of the scores like CCC, where a sufficiently large distribution of values is expected for meaningful estimation of mean and standard deviation and hence a reliable score calculation.

We also show that the approach detects weak molecular signals that are at low resolution areas of the map and not otherwise obvious. We recommend that residues associated with FDR scores less than 0.95 usually require attention and residues with scores less than 0.9 usually reflect clear cases of backbone mistrace.

We also demonstrate that the approach is useful in detecting residue mistraces in a model. Hence, the tool is useful as part of iterative model building pipelines or to evaluate the final model. Automated pruning of models based on this approach can be a useful step in the iterative model building and refinement process. Models after pruning can be also a starting point for extending or iterative building with the Buccaneer model building tool. As presented in the results, the approach is useful to validate ligands, carbohydrates, and nucleic acids as well.

The implementation of this score as a tool in the CCP-EM software suite makes it easily accessible for the cryo-EM community. The described software tool is available from the CCP-EM suite as “FDR validation task” confidence map

calculation can be run as part of this task, where the user has to provide the original map (preferably unmasked) and the model to validate. As an option, the user can adjust the size of the noise box used for calculation of background statistics. In some cases, especially if the specimen is significantly elongated in one direction, users should also check the preview of the noise boxes to make sure that the noise box does not contain any part of the molecular volume. The extended options for the confidence maps section allows to set the advanced parameters for the FDR maps calculation (Figure 8A).

If the user has already generated the FDR map, it can be used directly as an input (Figure 8B). Instead of a confidence map, any custom map can be used as well and residues will be assigned scores based on values in the map. Validation based on the scores of backbone atoms is run by default, users can additionally choose to validate only the CA positions. Optionally, the model can be pruned further to remove residues associated with low confidence scores. A model file with atomic b-factors replaced by the confidence scores and a CSV file containing the confidence scores for each residue are generated as outputs. If the option to prune the model was chosen, a pruned model is provided as the additional output, along with a text file containing the list of all removed residues. Figure 8C shows the launcher tab with a list of output files generated from the job. On the results tab, a link is included to open the resulting models directly in UCSF Chimera with the model colored based on the confidence scores. Figure 8D shows the results open directly in UCSF Chimera with the resulting model colored according to the confidence score.

For a confidence map of size $192 \times 192 \times 192$ voxels, the assignment of FDR scores to residues takes about 0.22 s on a PC with specification: Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz x 8, 8 GB RAM. The latest CCP-EM nightly release available from <https://www.ccpem.ac.uk/download.php> includes the implementation of “FDR validation task” The source code of the tool for evaluating atomic models based on confidence maps is available from (<https://github.com/m-olek/FDR-validation>).

CONCLUSION

In this study, we present a tool for validating atomic models derived from cryoEM maps. It works based on the calculation of the confidence maps, which estimates molecular signal to noise at every voxel, and also detects weak signals from the low resolution areas of the map. This helps to assess atomic positions based on the

REFERENCES

- Afonine, P. V., Klaholz, B. P., Moriarty, N. W., Poon, B. K., Sobolev, O. V., Terwilliger, T. C., et al. (2018). New Tools for the Analysis and Validation of Cryo-EM Maps and Atomic Models. *Acta Cryst. Sect D Struct. Biol.* 74 (9), 814–840. doi:10.1107/s2059798318009324
- Bai, X., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., et al. (2015). An Atomic Structure of Human γ -Secretase. *Nature* 525 (7568), 212–17. doi:10.1038/nature14892
- Beckers, M., Jakobi, A. J., and Sachse, C. (2019). Thresholding of Cryo-EM Density Maps by False Discovery Rate Control. *Int. Union Crystallogr. J.* 6 (1), 18–33. doi:10.1107/s2052252518014434

local information in the map and identify mistraced residues in the model. This approach is complementary to other validation tools that quantify agreement with the map, as it evaluates atom positions based on the local map information. We believe that, with the integration with the CCP-EM software suite, the presented tool will be a useful addition to the existing validation tools.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

MO developed the tool, performed analysis on different examples, and drafted the manuscript. AJ and MO conceived the idea. AJ helped with analysis, implementation of the tool in CCP-EM, and manuscript writing.

FUNDING

This work was supported by PhD studentship from the University of York and Diamond (eBIC) and Wellcome Trust research grant WT (208398/Z/17/Z).

ACKNOWLEDGMENTS

We thank Prof. Kevin Cowtan and Prof. Peijun Zhang for useful discussions throughout the course of this study. We thank Maxmillian Beckers, Arjen Jakobi, and Prof. Carsten Sachse for help with implementation of FDR approach in CCP-EM and useful comments on this work. We also thank Sony Malhotra for comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.652530/full#supplementary-material>

- Benjamini, Y., and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* 29 (4), 1165–1188. doi:10.1214/aos/1013699998
- Burnley, T., Palmer, C. M., and Winn, M. (2017). Recent Developments in the CCP-EM software Suite. *Acta Cryst. Sect D Struct. Biol.* 73 (6), 469–477. doi:10.1107/s2059798317007859
- Cowtan, K. (2006). The Buccaneer Software for Automated Model Building. *Acta Crystallogr. D* 62. *Acta Crystallogr. Section D, Biol. Crystallogr.* 62 (October), 1002–100211. doi:10.1107/s0907444906022116
- Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010). Features and Development of Coot. *Acta Crystallogr. D Biol. Cryst.* 66 (4), 486–501. doi:10.1107/s0907444910007493

- Herzik, M. A., Wu, M., and Lander, G. C. (2019). High-Resolution Structure Determination of Sub-100 KDa Complexes Using Conventional Cryo-EM. *Nat. Commun.* 10 (1), 1032. doi:10.1038/s41467-019-08991-8
- Hoh, S. W., Burnley, T., and Cowtan, K. (2020). Current Approaches for Automated Model Building into Cryo-EM Maps Using Buccaneer with CCP-EM. *Acta Cryst. Sect D Struct. Biol.* 76 (6), 531–541. doi:10.1107/s2059798320005513
- Jakobi, A. J., Wilmanns, M., and Sachse, C. (2017). Model-Based Local Density Sharpening of Cryo-EM Maps. *ELife* 6 (October), e27131. doi:10.7554/elife.27131
- Joseph, A. P., Malhotra, S., Burnley, T., Wood, C., Clare, D. K., Winn, M., et al. (2016). Refinement of Atomic Models in High Resolution EM Reconstructions Using Flex-EM and Local Assessment. *Methods* 100 (May), 42–49. doi:10.1016/j.ymeth.2016.03.007
- Lagerstedt, I., Moore, W. J., Patwardhan, A., Sanz-García, E., Best, C., Swedlow, J. R., et al. (2013). Web-Based Visualisation and Analysis of 3D Electron-Microscopy Data from EMDb and PDB. *J. Struct. Biol.* 184 (2), 173–181. doi:10.1016/j.jsb.2013.09.021
- Lawson, C. L., and Chiu, W. (2018). Comparing Cryo-EM Structures. *J. Struct. Biol.* 204 (3), 523–526. doi:10.1016/j.jsb.2018.10.004
- Lawson, C. L., Kryshtafovych, A., Adams, P. D., Afonine, P. V., Baker, M. L., Barad, B. A., et al. (2021). Cryo-EM Model Validation Recommendations Based on Outcomes of the 2019 EMDDataResource Challenge. *Nat. Methods* 18 (2), 156–164. doi:10.1038/s41592-020-01051-w
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Long, F., Vagin, A. A., et al. (2011). REFMAC5 for the Refinement of Macromolecular Crystal Structures. *Acta Crystallogr. Section D: Biol. Crystallogr.* 67 (Pt 4), 355–367. doi:10.1107/s0907444911001314
- Palmer, C. (2016). Mrcfile: MRC File I/O Library. Available at: <https://github.com/ccpem/mrcfile>.
- Petersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera?A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25 (13), 1605–1612. doi:10.1002/jcc.20084
- Pfab, J., Phan, N. M., and Si, D. (2021). DeepTracer for Fast De Novo Cryo-EM Protein Structure Modeling and Special Studies on CoV-Related Complexes. *Proc. Natl. Acad. Sci.* 118 (2). doi:10.1073/pnas.2017525118
- Pintilie, G. (2020). Measurement of Atom Resolvability in CryoEM Maps with Q-Scores. *Microsc. Microanal.* 26 (S2), 2316. doi:10.1017/s1431927620021170
- Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S., and Richardson, D. C. (2020). New Tools in MolProbity Validation: CaBLAM for CryoEM Backbone, UnDowser to Rethink "waters," and NGL Viewer to Recapture Online 3D Graphics. *Protein Sci.* 29 (1), 315–329. doi:10.1002/pro.3786
- Ramírez-Aportela, E., Erney, D. M., and Fonseca, Y. C. (2011). 'FSC-Q: A CryoEM Map-To-Atomic Model Quality Validation Based on the Local Fourier Shell Correlation'. *Nat. Commun.* 12 (1), 42. 10.1038/s41467-020-20295-w
- Subramaniam, S. (2019). The Cryo-EM Revolution: Fueling the Next Phase. *Int. Union Crystallogr. J.* 6 (1), 1–2. doi:10.1107/s2052252519000277
- Terwilliger, T. C., Adams, P. D., Afonine, P. V., and Sobolev, O. V. (2020). Cryo-EM Map Interpretation and Protein Model-building Using Iterative Map Segmentation. *Protein Sci.* 29 (1), 87–99. doi:10.1002/pro.3740
- Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci.* 27 (1), 293–315. doi:10.1002/pro.3330
- Wojdyr, M. (2017). Project-Gemmi/Gemmi. C++. GEMMI. Available at: <https://github.com/project-gemmi/gemmi> (Accessed April 9, 2021)
- Yin, W., Mao, C., Luan, X., Shen, D.-D., Shen, Q., Su, H., et al. (2020). Structural Basis for Inhibition of the RNA-dependent RNA Polymerase from SARS-CoV-2 by Remdesivir. *Science* 368 (6498), 1499–1504. doi:10.1126/science.abc1560
- 'NumPy Reference NumPy v1.16 Manual' (2019). Available at: <https://docs.scipy.org/doc/numpy-1.16.0/reference/> (Accessed April 9, 2021).
- 'PyQt 4.9.4 Reference Guide' (2011). 'PyQt 4.9.4 Reference Guide'. Available at: <https://doc.bccnsoft.com/docs/PyQt4/>

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Olek and Joseph. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.