Using cryo-EM methods to uncover structure and function of bacteriophages

Ouyang, R.

2

Chapter 2

Cryo-EM as a tool for structural analysis of structure and function of biological samples

Introduction

In recent years, cryo-electron microscopy (cryo-EM) has gained immense popularity as a powerful tool for studying the structures of biological molecules and their interactions in situ. The development of direct electron detectors and improvements in software and hardware have significantly advanced the field, enabling the high-resolution structural analysis of proteins and macromolecular assemblies that were previously difficult to study. In particular, cryo-EM has provided a breakthrough in the structural analysis of membrane proteins, which are often difficult to crystallize. Cryo-EM's versatility in studying a wide range of samples, including proteins, protein assemblies, intact viruses, and small cells, has further contributed to its success. With continuous advancements in sample preparation techniques, cryo-EM's applications are expected to continue expanding, thereby further revolutionizing the field of structural and cellular biology.

The use of electrons instead of photons for imaging biological samples is not a novel concept. The first transmission electron microscope (TEM) was developed in 1931[1]. However, imaging biological samples using an electron beam requires the sample to be imaged in a vacuum to avoid deflection of the beam by air molecules. To image biological specimen, elaborate sample preparation procedures had to be applied. These typically include dehydration, heavy metal staining, plastic embedding and sectioning. Although these techniques provide valuable insights into cellular architectures, they do not allow for the investigation of fine structural details. High resolution imaging became possible with the advent of cryo-electron microscopy (cryo-EM) which circumvents the traditional sample preparation steps. This imaging method, developed by Dubochet and colleagues in the 1980s[2], flash-freezes the sample in a cryogen. This preparation method immobilizes biological samples in a glassy (vitreous) ice without altering the structure of the sample. Cryo-EM allows the study of protein structures that do not crystallize, as well as heterogeneous samples that are not accessible to other techniques such as X-ray crystallography. However, obtaining high-quality datasets may be limited by the sample's dimensions and heterogeneity. The resolution of cryo-EM is a determining factor to understand the function of proteins and macromolecular machines.

In order to determine the structures of biological molecules in solution, a variety of imaging techniques about cryo-EM have been developed. Among these, single-particle

analysis (SPA) and cryo-electron tomography (cryo-ET) are powerful methods that have been used to obtain three-dimensional (3D) maps of biological samples at near-atomic resolution. SPA involves the analysis of individual particles, such as proteins or viruses, and allows for the determination of their 3D structures in solution. Cryo-ET, on the other hand, is particularly useful for the study of heterogeneous samples, such as the architectures of bacteria, viruses with varying shapes, and cellular components in *situ*[3-5]. Cryo-ET has the unique ability to produce 3D reconstructions of pleomorphic objects, which provides opportunities to capture intermediate biological events in the cellular environment. Furthermore, the spatial relationships between macromolecular complexes within a cellular tomogram can be determined using cryo-ET[6]. These techniques have significantly advanced the field of structural biology by enabling the visualization of biological structures in their native environments and are poised to continue making significant contributions to our understanding of complex biological processes.

Cryo-EM is now a standard tool in structural biology that can capture high-resolution structural information of diverse sample types. For example, it has been used to study small proteins such as membrane transporters[7-10], enzymes[11-13], and signaling proteins[14], as well as large macromolecular complexes like ribosomes[15,16], the spliceosome[17,18], and the nuclear pore complex[19-22]. Virus particles, including HIV[23-25], influenza[26-28], and Zika[29,30], have also been extensively studied using cryo-EM. Furthermore, entire cells and organelles, such as the components of the mitochondrial respiratory chain[31-33], have been analyzed using this technique. Overall, cryo-EM has provided new insights into the assembly and operation of cellular machinery, as well as the coordination of essential cellular functions such as transport mechanisms and signaling pathways. Cryo-EM also has played a prominent role in coronavirus research, providing crucial insights into the structure and function of many coronavirus proteins. Structural studies have focused on the spike protein, which enables viruses to bind and enter host cells and is a major target for vaccine development[34-37]. Additionally, studies have focused on the replication and transcription complex (RTC), a highly conserved enzymatic complex driving coronavirus RNA synthesis[38-41], which constitutes an attractive target for antiviral drug development. These studies have provided a structural basis for the development of antiviral strategies[42], and have thus opened new avenues for therapeutic interventions against coronavirus infections. Overall, the continuing advances in cryo-EM have revolutionized the field of structural biology, providing

unprecedented insights into the molecular details of biological systems and offering new opportunities for drug development and therapeutic interventions.

Sample Preparation

As a crucial step in cryo-EM, the sample preparation process is essential for preserving the native structure of biological samples and obtaining high-resolution images. The process involves several key steps. Firstly, a thin layer of the sample is applied to a support film typically made of carbon or graphene oxide, and the excess solution is most commonly removed by blotting with filter paper, resulting in a thin layer of the sample[43]. Next, the sample is rapidly frozen by plunging it into a cryogen such as liquid ethane or ethane-propane mixtures[44]. This vitrification step is critical for preventing the formation of ice crystals and preserving the native structure of the sample. Finally, the vitrified sample is transferred to the cryo-EM microscope while maintaining cryogenic temperatures of the sample to prevent any thawing or sublimation. High-energy electrons are then used to generate two-dimensional projections of the sample. Subsequently, these images can be combined to create a three-dimensional reconstruction of the sample.

Careful sample handling, vitrification, and imaging are essential to preserve the native structure of the sample and obtain high-quality structural information. One critical factor is the thickness of the sample, which must be thin enough for most electrons to pass through without being deflected or undergoing multiple interactions, but thick enough to avoid detrimental particle interactions at the air-water interface[45]. For single-particle analysis (SPA) using cryo-EM, the ideal sample thickness is such that it contains only the particles of interest, while for cryo-electron tomography (cryo-ET), thicker samples up to several hundred nanometers can yield useful biological information, but at the cost of achievable resolution. Additionally, the sample must be quickly frozen to prevent the formation of ice crystals, which can distort the sample and affect the resolution of the final images. Another critical factor is the control of sample heterogeneity, which can result in a mixture of particles with varying conformations and orientations, making it difficult to obtain a high-resolution 3D reconstruction. Furthermore, contamination of the sample with dust, protein aggregates, or other debris can negatively impact the quality of the final images.

Both SPA and cryo-ET have unique advantages in studying the structural details of biological molecules. While SPA can provide atomic or near-atomic resolution views of individual factors performing their biological functions[46-48], cryo-tomography enables the visualization of molecular ultra-structures in their physiologically relevant environment[49]. Therefore, both approaches are essential in providing critical contributions to our understanding of biological processes.

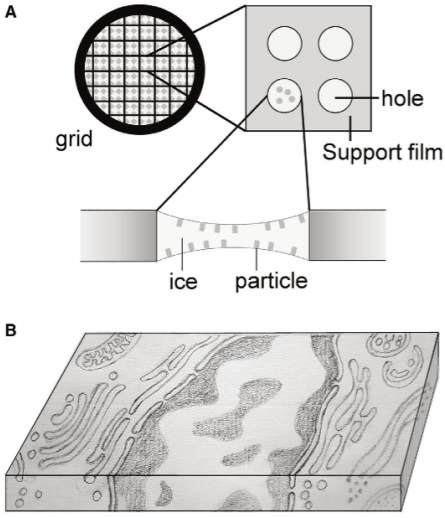

Figure 1. Schematics of single particles in ice for SPA and cell section for cryo-ET. (A) An electron microscopy (EM) grid, with a single square enlarged to show holes in the support film, and an edge-on view of a hole with a meniscus of ice containing particles that have partitioned to the air-water interfaces. A typical grid square would be ~40 μm across, and a hole would be ~1 μm. **(B)** A section of part of a vitrified eukaryotic cell, showing part of the nucleus and surrounding organelles, representing a lamella produced by focused ion beam milling. A lamella could be ~10 μm across.

Data Collection

Data collection is a critical step in the acquisition of high-resolution structural information from Cryo-EM. It is essential to optimize imaging conditions to maximize the signal-to-noise ratio and resolution of the reconstructed images. To achieve this, several critical steps in the data collection process need to be carefully controlled. Firstly, grid screening is necessary to identify suitable areas for data collection. This is typically done using a low-magnification imaging mode to search for areas with well-

distributed particles and minimal ice contamination. Accurate defocus estimation is also critical for obtaining high-quality images. This can be achieved by collecting a series of images at different defocus values to determine the optimal defocus value for data collection. During image acquisition, thousands of images of the same sample must be taken from different angles at low electron doses to prevent radiation damage and minimize the effects of electron beam-induced motion. Furthermore, acquisition parameters such as the exposure time, defocus, and pixel size need to be optimized to achieve high-quality images. Lastly, image alignment is a crucial step in generating a high-quality 3D reconstruction, requiring the accurate alignment of numerous images of the sample from different angles. However, aligning images of low-contrast features such as flexible or heterogeneous samples can pose significant challenges.

Single-Particle Analysis

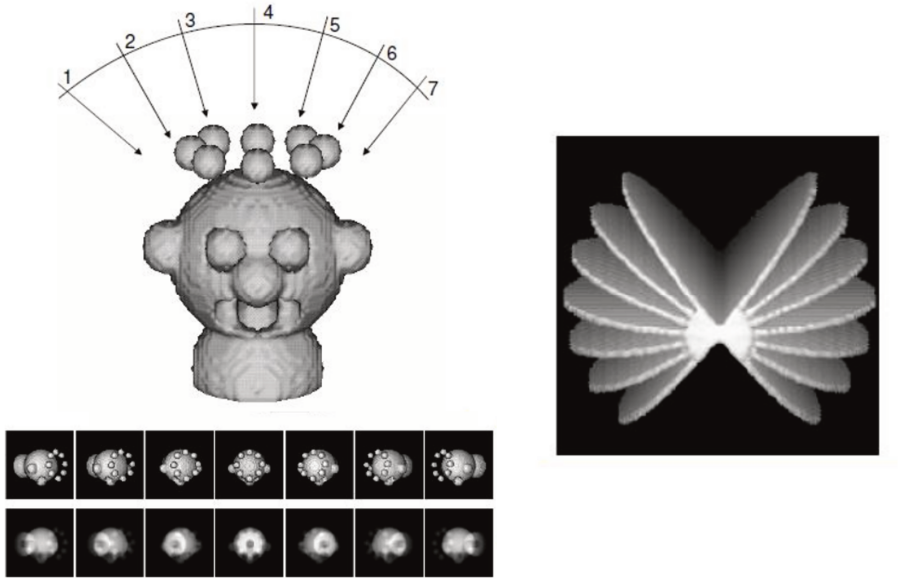

Figure 2. Central section theorem. The central section theorem is a fundamental principle in electron microscopy that relates the 2D projections of a 3D object to its 3D Fourier transform. According to this theorem, the Fourier transform of a 2D projection of an object is equal to a central section of the 3D Fourier transform of the object, which passes through the origin and is perpendicular to the direction of projection. This theorem is important in determining the 3D structure of a macromolecule from its 2D projection images obtained by electron microscopy.

The data collection process for SPA entails several crucial steps. Firstly, the sample is

vitrified on a grid through rapid freezing, followed by loading the grid into the electron microscope. Low magnification imaging is used to identify well-distributed particles on the grid, after which a series of images are captured of the particles at different angles using an electron microscope. The electron microscope parameters, such as electron dose, defocus, and astigmatism, are optimized to enhance the signal-to-noise ratio and resolution of the reconstructed images. Automated software[50-52] can be used to capture thousands of 2D projections of targets while maintaining optimal electron doses to minimize the adverse effects of radiation damage and electron beam-induced motion. The acquisition parameters, including exposure time, defocus, and pixel size, are fine-tuned to achieve high-quality images. After acquiring the images, alignment is performed to generate a high-quality 3D reconstruction.

The SPA data processing workflow (Figure 3) starts with pre-processing, including motion correction, contrast transfer function (CTF) estimation, and particle selection from the collected images. The data is then subjected to alignment, classification, and 3D reconstruction, incorporating CTF correction, which depend on algorithms to detect similarity using cross-correlation or optimization of a target function. Due to the inherent beam sensitivity of biological samples, the images must be collected under low-dose conditions. This results in noisy data, requiring iterative processing and probabilistic (Bayesian) methods to avoid introducing artifacts by trapping in false minima[53]. In the particle picking step, individual particles are selected from the cryo-EM micrographs, and their images are extracted for further analysis. The extracted particle images are then subjected to 2D classification, which groups similar images together based on their structural features, and selects the highest quality images for 3D reconstruction. The 3D refinement step involves the alignment and merging of the selected particle images to generate a 3D density map of the macromolecule. The resulting map can then be used to model the atomic structure of the macromolecule. Validation of the reconstructed structure is critical, and various metrics such as resolution estimation and map validation tools are used to ensure accuracy and reliability. Software developments such as cryoSPARC[54] or Relion4[55] have enabled on-the-fly data analysis as images emerge from the microscope, continually improving the performance and speed of SPA workflows.

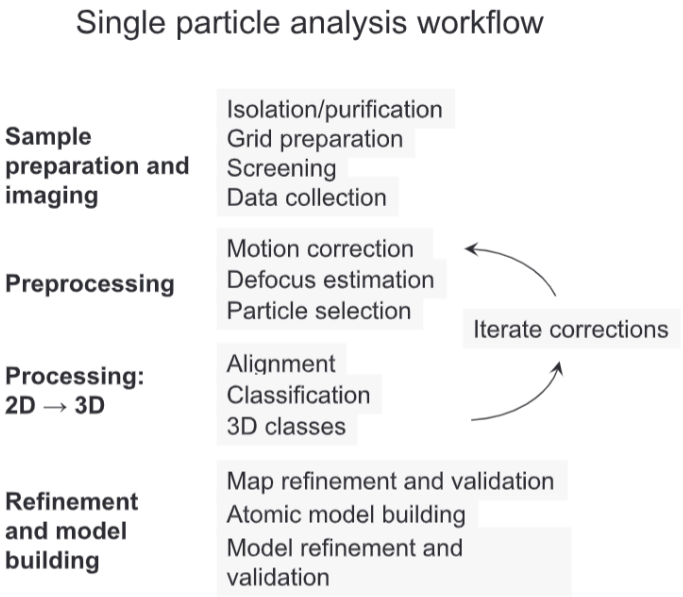

Figure 3. Single-particle workflow. The main steps in macromolecular structure determination by single-particle analysis (SPA).

Atomic Model Building and Fitting

Atomic model building is a crucial step in the cryo-electron microscopy (cryo-EM) workflow for the determination of the 3D structure of biological macromolecules. The goal of atomic model building is to fit a high-resolution atomic model into the cryo-EM map of the macromolecule. The accuracy and completeness of the atomic model building depend on the resolution of the cryo-EM map.

Numerous tools, such as Coot[56], Phenix[57], Rosetta[58] and ISOLDE[59] are available for fitting and building atomic models into cryo-EM maps. The choice of approach depends primarily on the map resolution, which determines the size of the rigid body that can be fitted. At low resolutions from 10 Å to 30 Å, cryo-EM can only provide a rough outline of the shape and size of the sample, without showing any details of its internal structure or composition. The fitting is typically done using rigid body fitting of known domains or subunits from homologous structures. While at intermediate resolution from 4 Å to 10 Å, secondary structures and some features can be fitted, such as alpha helices and beta sheets. While at higher resolutions from 2 Å to 4 Å, the fitting involves the

placement of the tertiary structures, such as side chains, loops, and disulfide bonds. At atomic resolutions below 2 Å, cryo-EM can reveal almost all features of the atomic structure of the sample, such as hydrogen atoms, water molecules, metal ions, and isotopic labels. As of 2020, the majority of the protein structures determined by cryo-EM are at a lower resolution of 3 to 4 Å[60]. However, as of 2020, the best cryo-EM resolution has been recorded at 1.22 Å[47], making it a competitor in resolution in some cases. If a related structure is available, building a homology model using the sequence of the imaged sample is the most straightforward approach. In the absence of a known homologous structure, de novo atomic model building can be used to generate an initial 3D structure from the sequence of the sample. The resolution range for building atomic models directly from cryo-EM maps is between 1.5 Å and 2.5 Å[61].

In recent years, the application of deep learning techniques has greatly advanced de novo atomic model building, which has had a significant impact on the discovery of new structures. The recent application of deep learning (DL) techniques[62] has greatly advanced de novo protein structure prediction using co-evolutionary sequence data. AlphaFold2[63,64], in particular, has achieved remarkable success in de novo structure prediction and is having a significant impact on the discovery of new structures. Predicted structures can be used to locate new components in complexes and recognize densities, which has proven highly successful in recent research[19,65,66], especially on SARS-CoV-2[67,68]. After fitting a known structure or a predicted model into the cryo-EM map, flexible fitting is often performed using molecular dynamics flexible fitting (MDFF)[69], iMODfit[70], or ISOLDE[59] software. Finally, both the model and the fit to the map should be validated, with important criteria for validation including model-to-map density correlation, model geometry, clashes, local score along the sequence, and consistency of local B factors[71].

Cryo-electron Tomography: *In Situ* Structural Biology

Cryo-electron tomography (cryo-ET) is an advanced imaging technique that has revolutionized the field of structural biology by enabling the visualization of complex cellular structures in their native, hydrated state at the nanoscale. Cryo-ET is a highly useful technique for visualizing intact protein complexes and membrane-bound structures that are difficult to isolate and study using conventional biochemical methods. The complex nature of the cellular environment, which cannot be accurately

represented by traditional 2D projection images, can instead be studied by cryo-ET. The technique has the potential to uncover transient super complexes and long-range interactions, which are essential for understanding biological processes at the molecular level. For instance, cryo-ET has been successfully used to investigate the interactions between bacteriophages and their host, shedding light on the mechanisms of viral infection. In addition, cryo-ET has led to a greater understanding of the 3D organization and architecture of bacterial cells and their components, such as membranes, organelles, cytoskeleton, and macromolecular complexes[72,73]. Specifically, the application of cryo-focused ion beam (cryo-FIB) milling to thin bacterial specimens makes cryo-ET for a high-resolution data collection, which allows deep research about cytoskeletal filament assembly, intracellular organelles, and multicellularity[74].

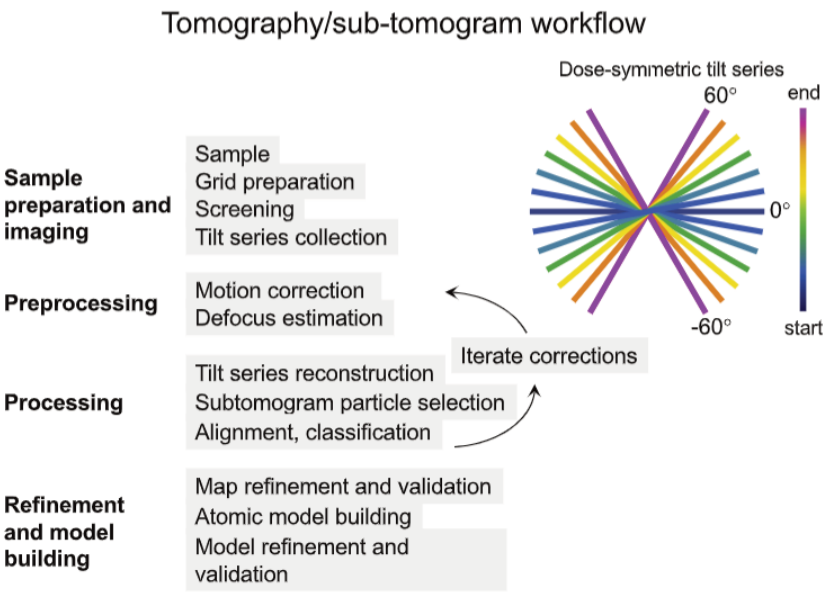

Figure 4. Tomography/Sub-tomogram Averaging Workflow. The main steps in reconstruction of complexes extracted from tomograms. A diagram of the dose-symmetric data collection scheme for tomograms is shown on the right. The order of recording the set of views at different tilt angles is represented by the color scale, with blue at the start of data collection and magenta at the end.

Cryo-ET involves acquiring a series of two-dimensional (2D) images of the specimen from different angles, which are then reconstructed into a 3D model using computer

algorithms. The data collection process in cryo-ET involves several steps, starting with the preparation of the biological sample for imaging. Data processing of cryo-ET (Figure 4) starts from tilt series data. Fiducial markers can be used for accurate alignment of the tilt views since the tilt angles are known. The data processing step involves correcting for imaging artifacts, such as specimen drift and electron beam-induced motion, as well as compensating for the contrast transfer function (CTF) of the microscope. 3D defocus correction is more complex but feasible for tilted samples, e.g., as implemented in NovaCTF[75]. Once a 3D tomogram is generated, it can be visualized and analyzed using various methods such as sub-tomogram averaging and segmentation.

Sub-tomogram averaging is a powerful technique that can improve the resolution of identical cellular structures. This involves extracting 3D sub-volumes or "sub-tomograms" from the reconstructed tomogram of the target structure. Sub-tomograms are then aligned and averaged to increase the signal-to-noise ratio and create a high-resolution 3D density map of the target structure. This approach is particularly useful for large macromolecular complexes or cellular organelles that cannot be purified or are too large for single-particle analysis.

Segmentation of the tomogram can be used to identify subcellular structures or molecular complexes within the sample. For example, the IMOD[76] software allows users to create a model containing contours describing the target surface in different slices. The 3D model can be displayed by surface or volume rendering of the image data. Quantitative results, such as shape and location of subvolumes, surface areas, and nearest neighbor distances can be obtained from the model data. However, segmentation is a labor-intensive process that requires significant manual input and can be highly subjective to the individual user. Automated methods using machine learning, such as EMAN2[77], are available to assist with segmentation.

It is worth noting that segmentation and sub-tomogram averaging may not be feasible for some unique objects, such as a bundle of flexible fibers. Furthermore, a limitation of cryo-ET is the presence of an imaging artifact called "missing wedge" in Fourier space, which is caused by the inability to acquire data beyond a maximum tilt angle (typically limited to +/- 60° tilt angle for standard cryo-ET sample holders). The resulting missing data is shaped like a wedge in Fourier space. In real space, this missing data mainly results in missing information and elongation of features oriented

parallel to the electron beam. However, features in the sample plane are still well-resolved. Recent techniques have been developed to overcome this limitation, such as using deep learning algorithms to partially fill in the missing information and reduce anisotropy in tomograms[78].

Applications of AI technology in Cryo-EM

Artificial intelligence (AI) technology has been increasingly applied to Cryo-electron microscopy (Cryo-EM) in recent years, providing researchers with powerful tools to enhance the efficiency and accuracy of data analysis and interpretation[79].

In the field of Cryo-EM, there has been a growing interest in the application of AI-based approaches for image denoising and artifact correction. Cryo-EM images often suffer from low signal-to-noise ratios and other artifacts, which can result in inaccurate structure determination. To address this issue, generative adversarial networks (GANs) and denoising deep neural networks (DNNs) have been introduced. GANs can learn the underlying distribution of noise and artifacts in a set of noisy or artifact-contaminated images, allowing them to generate clean images. Similarly, DNN can be trained to process images that have the same number and order of color channels as the noisy image. Recently, several advanced deep learning-based tools, such as Topaz-Denoise[80], Noise-Transfer2Clean (NT2C)[81], have been released for image denoising and artifact correction. These approaches have demonstrated their effectiveness in enhancing the resolution and quality of Cryo-EM reconstructions.

One of the main challenges in Cryo-EM is the processing and analysis of the large amounts of data generated, which can be time-consuming and require significant computational resources. AI techniques can help to automate and accelerate these processes, enabling more efficient and accurate analysis of Cryo-EM data. A typical application of AI in Cryo-EM is the use of deep learning algorithms for particle picking. Deep learning algorithms, such as convolutional neural networks (CNNs), have been trained to recognize particles based on their characteristic features, such as size, shape, and contrast. DeepPicker[82] and DeepEM[83] are some of the earliest models of fully automated particle recognition tools. These algorithms have shown to be more accurate and efficient than traditional manual or semi-automated particle picking methods. However, these approaches still have some drawbacks. The computational costs can be

high, and these approaches are not suitable for large particles or samples with ice contamination. Recently, several advanced deep learning-based particle-picking packages have been released to address the above issues, such as TOPAZ[84], WARP[85], and crYOLO[86].

While deep learning algorithms are mostly used in pre-processing steps, AI has also been used for automated model building and refinement in Cryo-EM. Recent studies have applied neural network-based approaches for determining SPA 3D reconstructions. CryoGAN[87], CryoDRGN[88], and 3D Flexible Refinement (3DFlex)[89] are three examples of such approaches, with CryoGAN modifying generative adversarial networks to reconstruct a 3D model for the continuous variability of biomolecules. CryoDRGN uses a modified variational auto-encoder for the posterior estimation of the volume, while 3DFlex adopts an auto-decoder model performing direct inference to increase the accuracy of the posterior estimation of the conformational coordinates. Although the auto-decoder approach enables the reconstruction of flexible regions with higher resolution to understand more detailed information about the dynamics of macromolecules, it requires more computational resources than the encoder-based approaches.

Recent advancements in single-particle analysis (SPA) techniques and image processing software have led to produce more and more high-resolution 3D EM maps. However, the evaluation and verification of these maps remains a challenge, as the concept of resolution in the electron microscopy field has not been fully defined and various approaches are currently in use. Fourier Shell Correlation (FSC) curves are the commonly used method to define the resolution of a 3D EM map, but it has limitations, including the requirement to set a reference threshold and insensitivity to isotropic filtering. Alternative approaches include BlocRes[90], ResMap[91], and MonoRes[92], which estimate local resolution based on sliding windows, 3D sinusoidal waves, and monogenic amplitude, respectively, but these approaches also have their own limitations. Recently, deep learning-based methods such as DeepRes[93] have been proposed to detect local changes in 3D EM maps caused by post-processing procedures with higher accuracy. Despite the ongoing debate and lack of a universally accepted approach in the field, further research and development of deep learning-based methods are necessary to advance the field of local resolution estimation.

The primary objective of structural analysis in the field of biology is to gain insights into the interactions between biological macromolecules by understanding the implications of their atomic structures. Many high-resolution EM maps are suitable for determining atomic structures using slightly modified software, such as Coot[56], Phenix[57], Rosetta[58] and ISOLDE[59]. However, the inherent properties of macromolecular complexes, such as high flexibility and multiple conformational states, make it challenging to determine atomic structures from intermediate resolution EM maps ranging from 5-8 Å. Researchers are actively addressing this issue by exploring new methods, especially by combining AI techniques. One such approach involves modeling the atomic structure by fitting a given template, such as a previously determined homologous atomic structure or a predicted structure based on amino acid sequences, to an EM map through a series of refinement processes. However, this method requires a high level of user expertise and incurs significant computational costs. To overcome these challenges, deep learning-based approaches, such as Emap2sec[94] and EMBuild[95], have been proposed to automatically build the atomic structure from relatively low-resolution EM maps. Compared to traditional approaches, these methods have shown improved accuracy in detecting secondary structures in the resulting 3D maps, and are expected to simplify structural determination in intermediate resolution cryo-EM maps. Despite the progress made in this area, further improvements are necessary to enhance the accuracy and reliability of these methods.

The application of AI technology in cryo-ET has significantly enhanced the efficiency of data processing, analysis, and validation. Machine learning algorithms have been developed to identify, classify, and exclude low-quality Cryo-ET data based on predetermined quality criteria. Innovative tools that utilize AI to automate various data analysis steps, such as particle picking, tomogram reconstruction, and local resolution determination, have also been developed. Notably, AITom[96], an open-source AI platform for cryo-electron tomography data analysis, has been established to enable fast and accurate data processing. Moreover, other tools based on machine learning algorithms have been developed to detect and extract target particles from reconstructed tomograms. These AI-based tools and algorithms have greatly facilitated the analysis of cryo-ET data, allowing for improved accuracy, speed, and reproducibility of results.