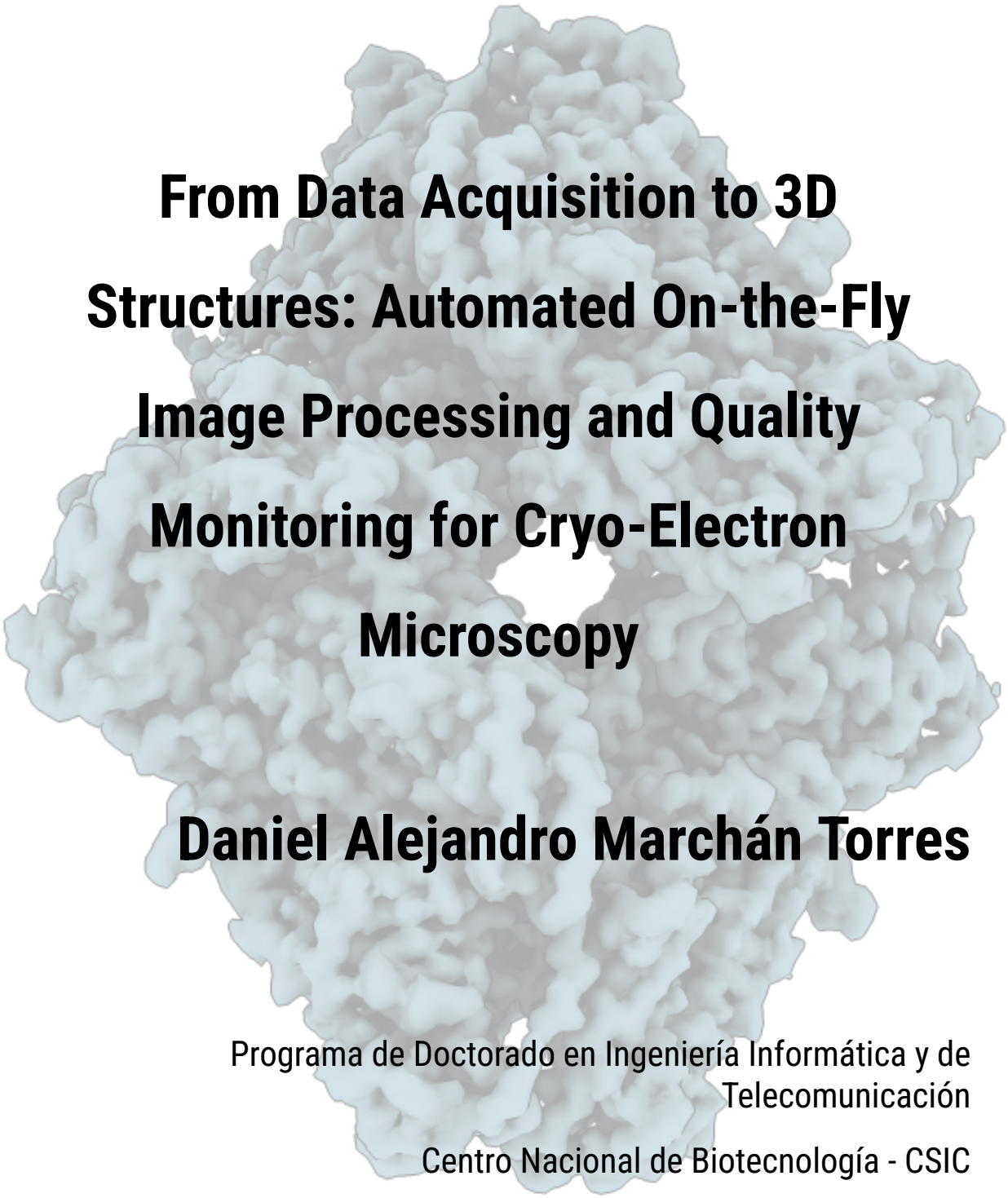
Universidad Autónoma de Madrid | Escuela de Doctorado



From Data Acquisition to 3D Structures: Automated On-the-Fly Image Processing and Quality Monitoring for Cryo-Electron Microscopy

Daniel Alejandro Marchán Torres

Programa de Doctorado en Ingeniería Informática y de Telecomunicación

Centro Nacional de Biotecnología - CSIC

Departamento de Ingeniería Informática

Madrid, 2026

El trabajo presentado en esta memoria ha sido realizado en el Departamento de Estructura de Macromoléculas del Centro Nacional de Biotecnología, bajo la dirección del Prof. José María Carazo García y del Dr. Carlos Óscar Sorzano Sánchez.

DECLARATION

Hereby, I declare that this thesis is my original work, developed independently during my PhD studies. All sources, references, and materials used in the preparation of this thesis have been properly cited and are listed in accordance with academic standards.

ACKNOWLEDGMENT

Acknowledgements are written at the end, when everything can be seen in perspective and with the peace of mind that comes from having all the countless documents gathered and the thesis finally completed. From this point of view, I can say that this has been an incredible journey, full of experiences, learning, new friendships, and growth both professionally and personally. I consider myself extremely fortunate to have been part of a research group that has provided me with all of this.

First of all, I would like to thank my thesis supervisors, Carlos Óscar and José María, who have always placed complete trust in me and in my work. A trust that was somewhat daunting at the beginning, but that proved to be fundamental to my growth and to believing that I was capable of taking on any challenge I set for myself.

Carlos, thank you for your teaching and for your time. I remember the early days, when you taught me everything from how to create a protocol in Scipion to how to run Xmipp from the command line. Thank you for guiding me through that learning curve, for always being the person to turn to when obstacles appeared along the way, and for consistently taking the time to sit down with me to work through doubts and think through solutions together. José María, thank you for your guidance and for offering me a perspective that I consider essential in the world of science: the importance of understanding the *why* behind things, defining a strategy, and establishing a plan to make the most of our developments.

To my colleagues and friends at the Bioinformatics Unit of the Centro Nacional de Biotecnología, it has been a true pleasure to share these four and a half years with you. Thank you for the support, patience, warmth, and everything you have taught me throughout this time. I entered as one person and left as a much more complete one, in every sense.

To all the collaborators with whom I have been fortunate to work, especially my colleagues during my research stay at the ESRF (Isai, Ludo, among others), thank you for sharing ideas, for exposing me to new ways of thinking, and for showing me that there is an international enthusiasm for doing science that truly matters.

Finally, I would like to thank Sara, my family, and my friends for always being that constant source of support and love that gave me the energy needed to carry this personal project forward. Thank you for everything, this achievement would not have been possible without you.

AGRADECIMIENTOS

Los agradecimientos se escriben al final, cuando se puede mirar con perspectiva y con la tranquilidad de tener toda la innumerable documentación y la tesis ya escrita. Desde esta distancia, puedo decir que este ha sido un viaje increíble, lleno de experiencias, aprendizaje, nuevas amistades y crecimiento tanto profesional como personal. Considero que he tenido la enorme suerte de haber formado parte de un grupo de investigación que me ha brindado todo ello.

En primer lugar, quiero agradecer a mis directores de tesis, Carlos Óscar y José María, quienes siempre han depositado una confianza plena en mí y en mi trabajo. Algo que al principio imponía respeto, pero que ha sido clave para mi crecimiento y para creer que era capaz de afrontar todo lo que me propusiera.

A Carlos, gracias por tu enseñanza y por tu tiempo. Recuerdo los comienzos, cuando me enseñabas desde cómo crear un protocolo en Scipion hasta cómo llamar a Xmipp por línea de comandos. Gracias por acompañarme en esa curva de aprendizaje, por ser siempre la persona a la que acudir cuando aparecía una piedra en el camino y por tomarte siempre el tiempo de sentarte conmigo para resolver dudas y pensar soluciones juntos. A José María, gracias por tu guía y por ofrecerme una visión que considero esencial en el mundo de la ciencia: la importancia de dar un porqué a las cosas, definir una estrategia y trazar un plan que permita sacar el máximo provecho a nuestros desarrollos.

A mis compañeros y amigos de la Unidad de Bioinformática del Centro Nacional de Biotecnología, quiero decirles que ha sido un auténtico placer compartir estos cuatro años y medio con ustedes. Gracias por el apoyo, la paciencia, el cariño y todo lo que me han enseñado durante este tiempo. Entré siendo una persona y salgo siendo alguien mucho más completo, en todos los ámbitos.

A todos los colaboradores con los que he tenido la suerte de trabajar, en especial a mis compañeros de estancia en el ESRF (Isai, Ludo, entre otros), gracias por compartir ideas, por abrirme a nuevas formas de pensar y por mostrarme que existe un entusiasmo internacional por hacer ciencia que importa.

Por último, quiero agradecer a Sara, a mi familia y a mis amigos por ser siempre esa fuente constante de apoyo y amor que me ha dado la energía necesaria para sacar adelante este proyecto personal. Gracias por todo, este logro no habría sido posible sin ustedes.

ABSTRACT

Single-particle analysis (SPA) by cryogenic electron microscopy (cryo-EM) has become a cornerstone of structural biology; yet, the workflow from data acquisition to an interpretable 3D structure remains highly manual, slow, and dependent on expert intervention. In high-throughput facilities, researchers frequently invest days of valuable microscope time collecting massive datasets with little to no real-time feedback on sample quality or the ultimate feasibility of achieving a high-resolution reconstruction. To address these challenges, we developed a fully unattended, end-to-end on-the-fly processing pipeline within the Scipion framework, designed not merely to automate sequential tasks but to function as an intelligent diagnostic system for modern cryo-EM environments.

The pipeline integrates streaming execution, multi-stage quality control filters, a novel consensus-based strategy for generating a data-specific particle-picking model, and a parallel 2D/3D validation scheme to ensure robust processing outcomes. Its design delivers a comprehensive, analysis-ready output, including curated micrographs, particle sets, high-quality 2D classes, and a preliminary 3D map, providing users with an immediate and reproducible entry point for downstream high-resolution refinement. To support real-time decision-making, we additionally developed a centralized, interactive quality-monitoring dashboard that consolidates processing metadata into an intuitive interface. This tool makes automated curation transparent, reduces cognitive load during acquisition, and offers actionable diagnostics that clarify why data are accepted or rejected, while also enabling future automation through its structured, model-oriented data representation.

Extensive benchmarking on 32 datasets (CryoPPP) demonstrated the pipeline's robustness and adaptability, achieving a 94% processing success rate and producing high-quality 3D reconstructions in 78% of cases across diverse and challenging samples. Deployment during a three-month International PhD Stay at the ESRF Cryo-EM Facility further validated the workflow under operational conditions. At the CM01 beamline, the pipeline consistently processed data faster than acquisition, delivering preliminary 3D maps within three hours. Approximately 70% of user experiments converged to interpretable structures, half of which reached 3–4 Å resolution, despite the inherent complexity of facility-collected samples. These results underscore its value as a real-time diagnostic tool capable of providing researchers with rapid, actionable feedback and identifying both promising and problematic datasets early enough to adjust acquisition strategies when appropriate, thereby accelerating structural determination and optimizing microscope time.

A major technical contribution of this work is the development of a granular, per-action HPC (High Performance Computing) job-submission layer that extends Scipion's queue system capabilities. This fine-grained scheduling removes static GPU allocation, maximizes throughput, and allows the system to scale seamlessly from single workstations to multi-GPU clusters, an essential requirement for reliable automation in facility settings. All workflow templates have been made publicly available through WorkflowHub ("CryoEM Facility Workflows"), ensuring long-term maintainability, transparency, and compliance with FAIR (Findable, Accessibles, Interoperables and Reutilizables) principles.

Together, the automated pipeline, the centralized monitoring dashboard, and the new HPC execution layer establish a modern, intelligent, and reproducible approach to cryo-EM data processing. They transform cryo-EM acquisition from a passive process into an active, data-driven experiment and lay the groundwork for future developments in adaptive acquisition, automated threshold optimization, and machine-learning-driven decision support. This work represents a significant step toward a more efficient, transparent, and accessible cryo-EM ecosystem for both facilities and individual laboratories.

KEY WORDS

Structural biology, Cryogenic electron microscopy (CryoEM), Single Particle Analysis (SPA), On-the-fly processing, Automation, Scipion.

RESUMEN

El Análisis de Partículas Individuales (SPA) mediante criomicroscopía electrónica (cryo-EM) se ha consolidado como una técnica esencial en biología estructural. Sin embargo, el flujo de trabajo que va desde la adquisición de datos hasta la obtención de una estructura 3D interpretable sigue siendo en gran medida manual, lento y altamente dependiente de la intervención de un experto. En instalaciones de alto rendimiento, los investigadores suelen invertir días de valioso tiempo de microscopio recolectando grandes volúmenes de datos sin disponer de información clara y en tiempo real sobre la calidad de la muestra o la viabilidad de obtener una reconstrucción 3D de alta resolución. Para afrontar estos desafíos, desarrollamos un flujo de procesamiento completamente autónomo y de principio a fin para análisis en tiempo real (on-the-fly) dentro de la plataforma de software de Scipion. Este sistema no solo automatiza tareas secuenciales, sino que actúa como una herramienta diagnóstica inteligente adaptada a los entornos actuales de cryo-EM.

El flujo de procesamiento integra una ejecución en streaming, filtros de control de calidad en múltiples etapas del procesamiento, una estrategia de consenso para generar un modelo de picado de partículas específico para cada muestra y un esquema de validación paralelo 2D/3D que garantiza resultados robustos. Su implementación ofrece una salida completa y lista para su análisis a posteriori: micrografías curadas, conjuntos de partículas filtradas, clases 2D de alta calidad y un mapa 3D preliminar, proporcionando a los usuarios un punto de partida inmediato y reproducible para llegar a un refinamiento de mayor resolución. Para facilitar la toma de decisiones en tiempo real, desarrollamos también un panel centralizado de monitoreo de calidad, interactivo, que agrupa todos los metadatos de procesamiento en una interfaz intuitiva. Esta herramienta hace transparente la limpieza de datos automática, reduce la carga cognitiva durante la adquisición y proporciona diagnósticos para toma de decisiones que explican por qué los datos son aceptados o rechazados, a la vez que sienta las bases para futuras automatizaciones gracias a su diseño estructurado y orientado a modelos.

Un extenso análisis comparativo realizado sobre 32 conjuntos de datos (CryoPPP) demostró la robustez y adaptabilidad del flujo de procesamiento, alcanzando una tasa de éxito del 94% y generando reconstrucciones 3D de alta calidad en el 78% de los casos, incluyendo muestras diversas y desafiantes. Su implementación durante la estancia internacional del doctorado de tres meses en la instalación de Cryo-EM del ESRF validó aún más el flujo de trabajo en condiciones reales de operación. En el beamline CM01, el flujo de procesamiento analiza los datos constantemente más rápido de lo que son adquiridos, produciendo mapas 3D preliminares en menos de tres horas. Aproximadamente el 70% de los experimentos convergieron hacia

estructuras interpretables, y la mitad de ellos alcanzaron resoluciones de 3–4 Å (alta resolución), pese a la elevada complejidad de las muestras típicas en una adquisición de datos. Estos resultados ponen en manifiesto su valor como herramienta diagnóstica en tiempo real, capaz de ofrecer retroalimentación rápida y útil, de identificar conjuntos de datos prometedores o problemáticos con suficiente antelación y permitir ajustes en la estrategia de adquisición cuando sea necesario. Todo ello acelera la determinación de estructuras y optimiza el tiempo de uso del microscopio.

Una contribución técnica clave de este trabajo es el desarrollo de una capa más granular de envío de trabajos de cómputo (jobs) por operación de procesamiento a sistemas HPC (computación de alto rendimiento), que amplía las capacidades del sistema de colas de Scipion. Esta gestión granular más fina elimina la asignación estática de GPU, maximiza el rendimiento y permite que el sistema escale sin problemas desde estaciones de trabajo individuales hasta clústeres multi-GPU, lo cual es esencial para una automatización fiable en distintos entornos de instalación. Todas las plantillas de flujo de procesamiento se han puesto a disposición pública en WorkflowHub ("CryoEM Facility Workflows"), garantizando mantenibilidad a largo plazo, transparencia y cumplimiento de los principios FAIR (Encontrables, Accesibles, Interoperables y Reutilizables).

En conjunto, el flujo de procesamiento automatizado, el panel centralizado de monitoreo y la nueva capa de ejecución para HPCs establecen un enfoque moderno, inteligente y reproducible para el procesamiento de datos de cryo-EM. Transforman la adquisición de datos en cryo-EM de un proceso tradicionalmente pasivo, en un experimento activo y guiado por resultados, y sientan las bases para futuros avances en adquisición adaptativa, optimización automática de umbrales y toma de decisiones impulsada por aprendizaje automático. Este trabajo representa un paso significativo hacia un ecosistema de cryo-EM más eficiente, transparente y accesible, tanto para instalaciones de recolección de datos como para laboratorios individuales.

PALABRAS CLAVE

Biología estructural, Microscopía electrónica criogénica, Análisis de Partículas Individuales, Procesamiento en tiempo real (on-the-fly), Automatización, Scipion.

TABLE OF CONTENTS

LISTS

List of Figures

List of Tables

ACRONYMS

cFSC conical Fourier Shell Correlation
cryo-EM Cryo-Electron Microscopy
cryo-ET Cryo-Electron Tomography
CCD Charge Coupled Device
CNN Convolutional Neural Network
CTF Contrast Transfer Function
DED Direct Electron Detector
DL Deep Learning
DQE Detective Quantum Efficiency
EM Electron Microscopy
EM Expectation–Maximization
EMPIAR Electron Microscopy Public Image Archive
ESRF The European Synchrotron
FSC Fourier Shell Correlation
GPU Graphics Processing Unit
GUI Graphic User Interface
HPC High-Performance Computing
HDD hard drive storage
HTTPS Hypertext Transfer Protocol Secure
MRC Multi-reference classification
MSA Multivariate Statistical Analysis
MRA Multi-Reference Alignment
NMR Nuclear Magnetic Resonance
PCA Principal Components Analysis
PDB Protein Data Bank
PSD Power Spectral Density
RELION REgularised LIkelihood OptimisatioN (Software)
SGD stochastic gradient descent
SLURM Simple Linux Utility for Resource Management
SNR Signal to Noise Ratio
SPA Single Particle Analysis
SSD Solid-State Drive
TEM Transmission Electron Microscope
URL Uniform Resource Locator

CHAPTER 1 – INTRODUCTION

1.1 CryoEM Background and Relevance

1.1.1 Structural Biology and CryoEM

Nobel Laureate and Physicist Richard Feynman once stated: "*It is very easy to answer many fundamental biological questions; you just look at the thing!*" [4]. This notion captures the essence of structural biology: by observing biological systems in sufficient detail, we can discern their structures and understand the mechanisms governing complex biological processes. In service of this aim, structural biology has been instrumental in major biological discoveries throughout history [5].

Structural biology relies primarily on three experimental techniques: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (EM), as illustrated in Figure 1. For decades, X-ray crystallography and NMR spectroscopy were the principal sources of high-resolution structures for proteins and nucleic acids. Indeed, prior to 2020, X-ray crystallography alone accounted for nearly 90% of the atomic coordinate entries in the Protein Data Bank (PDB) [5].

However, both techniques possess significant limitations. X-ray crystallography requires well-ordered three-dimensional (3D) crystals of macromolecules, and the final resolution is contingent on crystal quality. While effective for many stable proteins, growing suitable crystals of large, dynamic assemblies or integral membrane proteins is often a major challenge [5]. NMR spectroscopy, on the other hand, is generally restricted to smaller proteins and demands high concentrations of isotopically labeled samples.

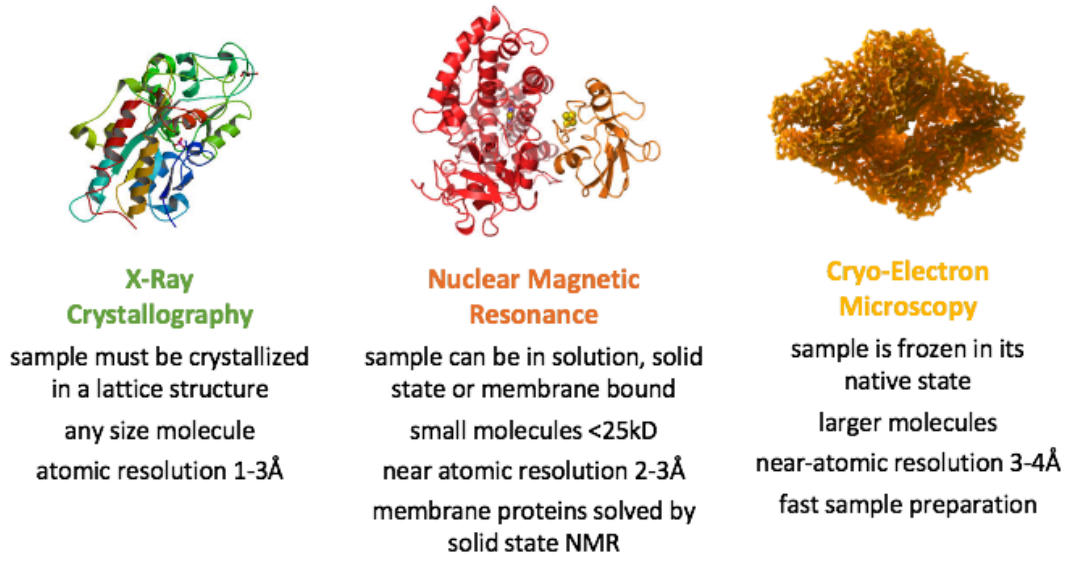


Figure 1. Overview of the three main techniques used for protein structure determination. **Macromolecular crystallography** is a well-established method that enables rapid structure determination once crystals are obtained, with results that can be complemented by **small-angle X-ray scattering (SAXS)** in solution. **Nuclear magnetic resonance (NMR)** spectroscopy is specially useful for studying macromolecular dynamics in solution and membrane proteins in the solid state. **Cryo-electron microscopy (cryo-EM)** has undergone a major transformation with the development of direct electron detectors, enabling near-atomic resolution of macromolecular complexes [6].

These limitations lead back to Feynman's proposition: "*can atomic structures of biological macromolecules be determined without crystallization, simply by "looking" at them with a powerful microscope?*" [5] In the 1970s, pioneers began exploring this possibility, developing the method now known as single-particle cryogenic electron microscopy (cryo-EM). Although early reconstructions were of low resolution, the technique's potential to study macromolecules in a near-native, non-crystalline state attracted significant interest [7], [8].

Over the subsequent decades, persistent efforts within the cryo-EM community led to steady advances in both methodology and resolution. Cryo-EM evolved from a complementary technique to a transformative method in its own right. Recent technological breakthroughs have made atomic-resolution structure determination using cryo-EM a routine process, thereby solidifying its position as a dominant technique in the field [5]. Its profound impact on structural biology was recognized with the "*2017 Nobel Prize in Chemistry, awarded for the development of cryo-EM for high-resolution structure determination of biomolecules in solution*" [9].

1.1.2 Historical Evolution of CryoEM

The direct visualization of the biological microcosm began in 1677, when Antonie van Leeuwenhoek used a single-lens optical microscope to observe unicellular organisms for the first time [10]. However, the wave nature of light imposes a fundamental resolution limit on light microscopy. To surpass this barrier and image subcellular structures, "*radiation with a much shorter wavelength, such as that of accelerated electrons*" [11], was required [12].

Ernst Ruska invented the electron microscope in 1931, and Max Knoll and Ruska built the first working prototype in 1933. This was a big step forward because it made it possible to take pictures at the nanometer and, eventually, atomic scale [2],[13],[14]. However, electron microscopy was initially ill-suited for biological specimens due to two primary challenges: the high-vacuum environment of the microscope, which is incompatible with hydrated biological material, and the high-energy electron beam, which can severely damage delicate structures [2]. Consequently, it took nearly seven decades for EM to mature into a viable method for high-resolution biological imaging [5], [11]. Figure 2 provides a visual overview of this technological evolution.

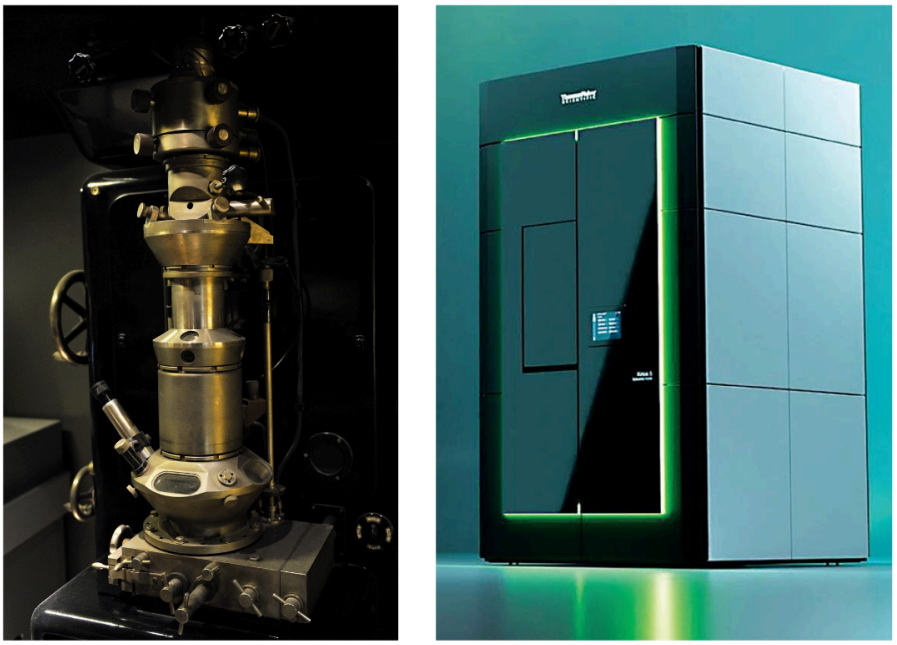


Figure 2. CryoEM Evolution. On the left, the first serially produced electron microscope by Siemens (1938), featuring three magnetic lenses, a resolution of approximately 3 nm, and magnification of about 3,000× [15]. On the right, the Thermo Scientific Krios 5 Cryo-TEM (2025), representing the state of the art in cryo-electron microscopy. The Krios 5 rapidly produces high-quality structures with exceptional data fidelity. Enhanced automation, improved optical precision, and AI-assisted operation make it both highly productive and user-friendly [16].

Improving Biological Contrast and Image Formation (1960–1980s)

A major early challenge was the low intrinsic contrast of unstained biological samples. The introduction of negative staining, using heavy-metal salts like uranyl acetate, enhanced visibility and enabled the first images of large complexes and viruses [17], [18]. Although this technique only revealed low-resolution molecular envelopes, it was a critical step. This era also saw the development of pioneering 2D and 3D reconstruction algorithms for symmetric assemblies and tilted specimens [19], [20], [21]. These methods, applied to negatively stained samples, were foundational precursors to modern cryo-EM workflows. However, this specimen preparation method suffered from dehydration-induced structural degradation and often forced particles into a preferred orientation on the support film [11].

Maintaining Structural Integrity: Frozen Hydration and Cryo-Preparation (1980–1990s)

To preserve structural integrity, biological samples must remain hydrated within the high vacuum of the EM column. Early work by Taylor and Glaeser demonstrated that vitrification, the process of rapid freezing to form non-crystalline, amorphous ice, could protect frozen-hydrated protein crystals from dehydration while preserving their ability to yield near-atomic diffraction patterns [22], [23]. The practical implementation of this concept arrived with the development of plunge-freezing techniques by Jacques Dubochet and colleagues, which enabled the vitrification of single particles suspended in thin aqueous films [7], [24]. This marked the birth of cryo-EM as we know it today, as vitrification preserves samples in a near-native state and mitigates electron beam damage. The approach is still used today with minimal modifications, although the process is now largely automated for improved reproducibility.

Emergence of Single Particle Analysis (1990–2000s)

While EM images are 2D projections, it was shown that a 3D structure could be computationally reconstructed from multiple 2D views [19], [20], [21]. Cryo-EM offered an elegant solution: rather than physically tilting the specimen, one could exploit the random orientations of particles naturally embedded in the vitreous ice. This "zero-tilt" approach, pioneered by Van Heel and collaborators, used angular reconstitution techniques to determine the orientation of particle images, making specimen tilting unnecessary [25], [26]. This innovation became a cornerstone of single-particle cryo-EM.

Another major limitation was radiation damage from the electron beam. Consequently, low electron doses were needed, resulting in images with very low signal-to-noise ratios (SNRs).

Henderson and Unwin overcame the problem by averaging images of many identical proteins in 2D crystals [27]. However, "*the difficulty of growing well-ordered 2D crystals hindered the method's broad application*"[5]. A conceptual breakthrough came from Joachim Frank, who proposed computationally combining noisy images of many individual, non-crystalline particles to determine their structure [8]. When this computational strategy was paired with plunge-freezing, *single-particle analysis* (SPA) within cryo-EM emerged as a versatile technique for resolving 3D structures from macromolecules in solution.

1.1.3 Modern Cryo-EM Era (2000s–Present): Advances in Hardware and Software

For decades, the theoretical potential of cryo-EM to reach atomic resolution remained unrealized due to technical limitations in hardware and software. A series of converging breakthroughs in the early 2010s, particularly the development of direct electron detection (DED) cameras and sophisticated Bayesian computational methods, finally turned this prediction into reality, ushering in the "resolution revolution" [5], [14].

The Direct Electron Detection Camera

The development of DED cameras was a critical turning point. Earlier detectors, like photographic film or scintillator-based Charge Coupled Device (CCD) cameras, either failed to preserve low-frequency signals essential for alignment or suffered from low detective quantum efficiency (DQE), which degraded high-resolution information. DED cameras, introduced in the early 2010s, detect electrons directly, dramatically improving the DQE across all spatial frequencies. Their high frame rates enable "movie-mode" acquisition, where a dose-fractionated series of images is recorded. These frames can be computationally aligned to correct for beam-induced sample motion, a major resolution-limiting factor. This combination of motion correction and improved signal preservation transformed cryo-EM from a qualitative tool into a quantitative method capable of routine sub-2 Å resolution [2], [11].

The most recent DED models offer improved temporal resolution, smaller pixel sizes, and higher frame rates, enabling ultrafast data collection. These advances have also facilitated new data acquisition strategies, such as beam-image shift, beam-tilt, and aberration correction [5]. Together, have significantly expanded the capabilities of cryo-EM beyond static structure determination, enabling both high-throughput and high-resolution imaging.

New Image Processing Algorithms

Parallel to hardware innovations, image processing algorithms have undergone a similar revolution. Early computational methods laid the groundwork, but it was the introduction of probabilistic and Bayesian approaches that dramatically improved the accuracy of particle alignment and classification, particularly for heterogeneous samples. The release of software packages like *RELION*, which implemented a regularized likelihood optimization algorithm, provided a statistically robust and user-friendly platform that became a gold standard in the field [28].

In parallel, software like *CryoSPARC* [29], *Warp* [30], and *Scipion* [31] introduced modern interfaces, GPU acceleration, and automated workflows, lowering the barrier to entry for non-specialists. Pre-processing steps, such as contrast transfer function (CTF) estimation and image denoising further enhanced the quality of final reconstructions [32]. More recently, the integration of deep learning for tasks like particle picking and quality assessment promises to further streamline and improve the cryo-EM pipeline [33], [34], [35], [36], [37].

Automation in Electron Microscopy

Historically, cryo-EM data collection was an arduous, manual process requiring highly skilled operators and considerable time at the microscope. The development of automated data collection software [38], [39], [40] combined with increasingly stable electron microscopes, has transformed this workflow. Today, many cryo-EM facilities operate like synchrotron beamlines, offering user-friendly platforms for unattended, high-throughput data acquisition [41]. This automation has democratized access to cryo-EM, enabling a broader community of scientists to collect high-quality data.

Together, these hardware and software advances have redefined what is achievable, bringing near-atomic resolution within reach for a wide array of biological targets.

1.1.4 Current Relevance and Impact in Structural Biology

Cryo-EM has revolutionized structural biology by providing high-resolution views of macromolecules without the need for crystallization. This has unlocked new frontiers of biological inquiry, particularly for systems that were intractable by traditional methods.

Overcoming Crystallization Barriers

One of cryo-EM's most significant contributions is its ability to solve the structures of membrane proteins and large, dynamic complexes. A prime example is the TRPV1 ion channel,

whose structure remained elusive by crystallography. The determination of its atomic structure in multiple functional states by cryo-EM demonstrated that the method could rival X-ray crystallography for challenging targets. Within five years, this success was replicated across all seven subfamilies of TRP channels [42], [43]. Similarly, cryo-EM has provided unprecedented, atomic-level snapshots of the entire spliceosome machinery through its functional cycle, a feat previously unimaginable [44], [45].

Broad Applicability Across Structural Biology

Cryo-EM now complements, and often surpasses, classical techniques. Its flexibility allows researchers to study a diverse range of systems, including:

- Flexible, multi-domain proteins in multiple conformations.
- Membrane proteins in native-like lipid environments.
- Viruses and their surface glycoproteins, which are critical for vaccine design.
- Macromolecular assemblies captured in distinct functional states.

This capacity to analyze structural heterogeneity is a key advantage for understanding the dynamic behavior of biological systems [2].

Cross-Disciplinary Impact and Pharma Adoption

The ability of cryo-EM to resolve protein–ligand interactions has attracted significant interest from the pharmaceutical industry, where it is now being integrated into structure-based drug discovery pipelines targeting previously inaccessible proteins like GPCRs and ion channels (Cheng, 2018). Furthermore, the rapid growth of the field has attracted expertise from computer science, engineering, and mathematics, fostering an innovative, interdisciplinary environment. High-throughput data collection has also led to the creation of large public datasets (e.g., EMPIAR), which not only improve reproducibility but also fuel the development of next-generation machine learning algorithms for image analysis [46].

In summary, cryo-EM has evolved from a niche, low-resolution technique into a mainstream method for atomic-level structure determination. Its versatility and power have fundamentally reshaped how we investigate complex biological questions, making it one of the most vital tools in modern life sciences [47].

1.2 Single Particles Analysis Workflow

Single-particle analysis (SPA) is the computational method used to reconstruct three-dimensional (3D) density maps from a large collection of 2D projection images of individual, vitrified macromolecules. The transformative nature of this process was aptly described by Sigworth (2016), who highlighted its capacity to derive structural order from immense datasets of noisy images: "*One starts with a set of perhaps 100,000 hopelessly noisy-looking images of single macromolecular 'particles', and by a seemingly magical process... the end result can be one or more 3D density maps from which atomic structures can be determined*" [48].

This "magic" is grounded in a set of core principles. The principal goal of SPA is to reconstruct a 3D density map by computationally averaging thousands or millions of low-dose images. This process relies on two fundamental assumptions:

1. **Structural Homogeneity:** The particles in the sample are assumed to be identical copies of the same macromolecule, differing only in their 3D orientation. This allows the signal from many noisy images to be combined coherently. In practice, perfect homogeneity is rare, and managing sample heterogeneity remains a central challenge.
2. **Projection Assumption:** Despite the extremely low signal-to-noise ratio (SNR), each 2D image is a projection of the 3D particle and contains sufficient information to determine its relative orientation. By correctly assigning an orientation to each particle image, a 3D reconstruction can be assembled.

This chapter provides a conceptual overview of the conventional SPA workflow (Figure 3), which is broadly divided into three stages: sample and grid preparation, data collection, and image processing. Although various software packages and algorithmic approaches exist for each stage, this section focuses on providing a high-level understanding of the process rather than detailed mathematical descriptions.

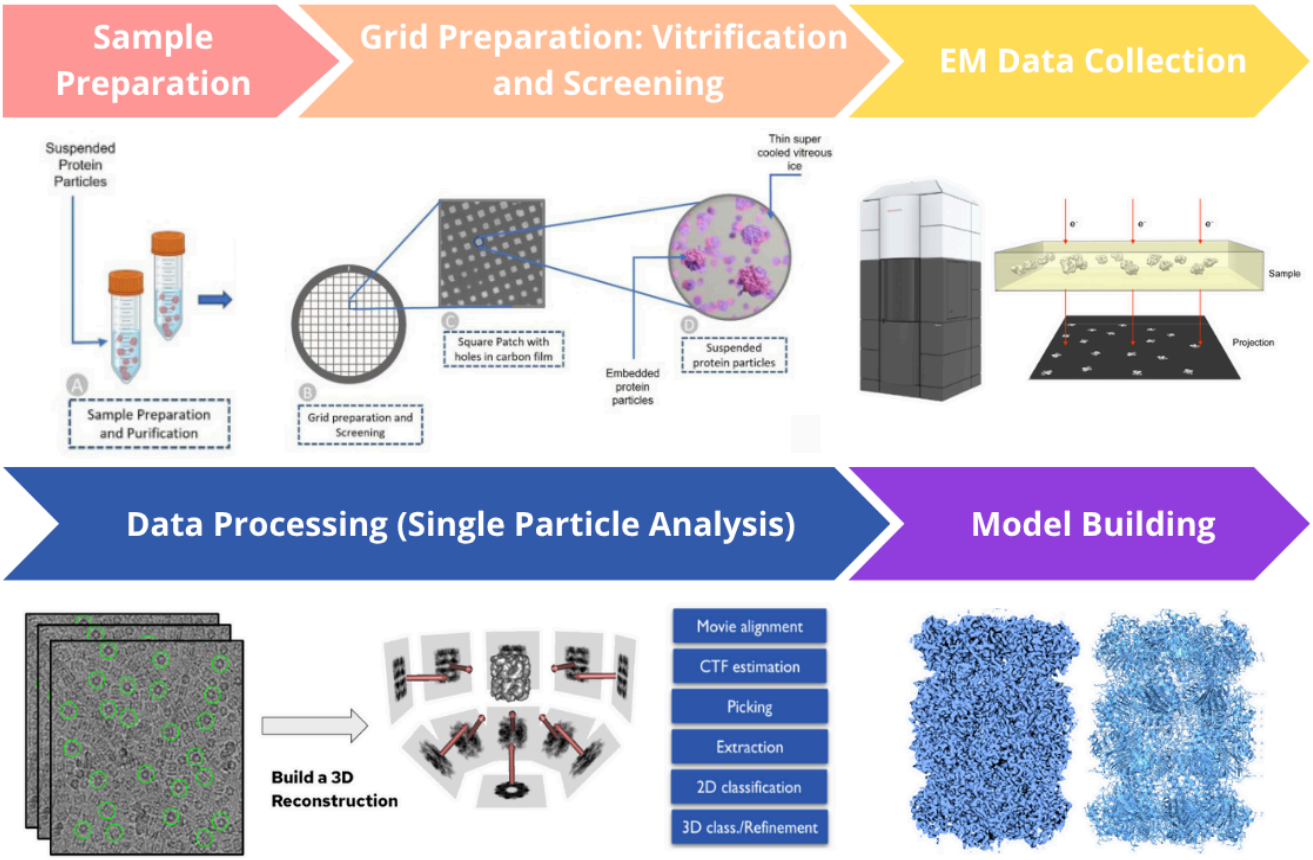


Figure 3. Overview of the SPA Workflow. The workflow begins with an aqueous, purified sample containing dispersed protein particles. The cryo-EM grid, containing holes, is then filled with the suspended particles and vitrified. After vitrification, the grid is loaded into the cryo-electron microscope, where the best regions are first identified during a screening step, followed by high-quality data collection. Once the data have been acquired, single-particle analysis begins. When an interpretable 3D reconstruction is obtained, the workflow proceeds to model building, where structural features of the reconstruction are identified and refined. This figure was created using elements adapted from the following research papers: [1], [3].

1.2.1 Sample and Grid Preparation

The first and arguably most critical step in SPA is the preparation of a high-quality biological sample. The macromolecule of interest must be purified to a high degree and then rapidly frozen (vitrified) in a thin layer of amorphous ice to preserve its native, hydrated conformation.

The process begins by applying a small volume of the purified sample to an electron microscopy (EM) grid, typically a small copper mesh coated with a perforated carbon film. Excess liquid is blotted away, leaving a thin aqueous film suspended across the holes in the carbon support. The grid is then plunged into a cryogen like liquid ethane or propane, which is maintained at cryogenic temperatures (approx. -180 °C) by a surrounding bath of liquid

nitrogen. This ultra-fast freezing prevents the formation of damaging ice crystals and instead traps the particles in random orientations within a layer of non-crystalline vitreous ice. The frozen grid is finally loaded into an electron microscope, where it is maintained at near–liquid nitrogen temperature for imaging [5]. A schematic depiction of this cryo-EM grid is shown in Figure 4.

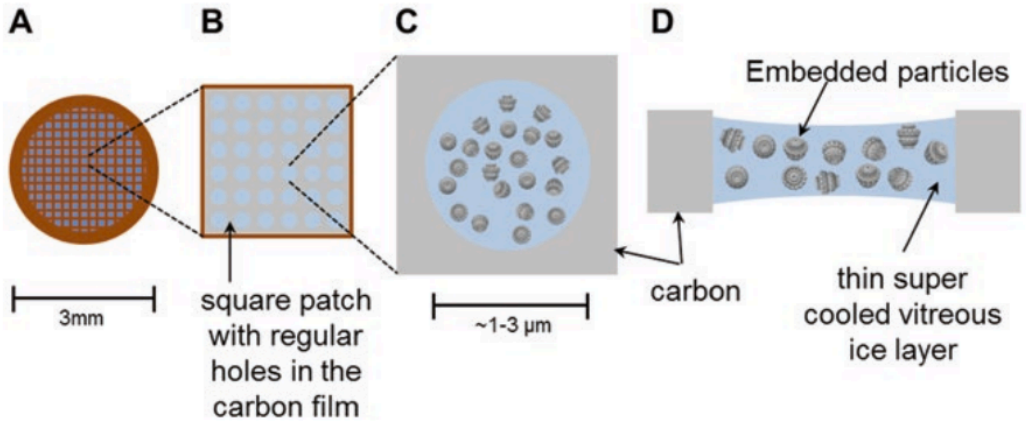


Figure 4. Representation of a CryoEM grid after vitrification."*(A) 3 mm copper mesh grid covered with a film of holey carbon. (B) A magnified image of a square patch reveals microscopic holes in the carbon. (C) Enlarged image of a single hole containing a layer of vitrified ice with protein molecules. (D) Cross section of a hole with particles embedded in ice*" [49].

The quality of the vitrified sample is a primary determinant of success. An ideal sample has an ice layer that is thin and uniform, containing a high density of well-dispersed, intact particles. If the ice is too thick, image contrast is reduced, and overlapping particles can complicate subsequent processing steps [5]. Conversely, if the ice is too thin, particles may be forced into the air-water interface and denature [2]. Achieving optimal ice quality often requires careful, iterative optimization of vitrification parameters and grid preparation protocols.

1.2.2 Data Collection

After vitrification, thousands of 2D projection images are acquired using a transmission electron microscope (TEM). This process involves two distinct phases: screening and high-resolution data collection. While both are essential for obtaining high-quality datasets, they serve different purposes and are typically performed using different instruments.

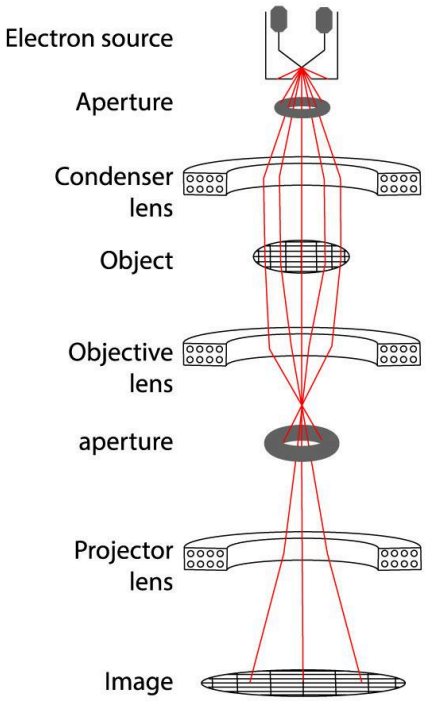


Figure 5. Schematic representation of a Transmission Electron Microscope [2] (modified).

As shown in Figure 5, a transmission electron microscope functions by emitting a high-energy beam of electrons from an electron gun (typically a field-emission gun, or FEG, for high-resolution applications). This beam is shaped and focused by a series of electromagnetic condenser lenses before it passes through the ultra-thin, vitrified specimen. As electrons interact with the sample, they are scattered, creating a projection image based on the sample's density and composition.

The scattered electrons then pass through an objective lens, which provides the primary magnification and is critical for determining the final resolution. Subsequent projector lenses further magnify the image onto a detector, the Direct Electron Detection (DED) camera. The entire system is held under an ultra-high vacuum to prevent electrons from scattering off air molecules, allowing them to travel unimpeded from the source to the detector.

Screening (Grid Evaluation)

Screening involves the initial evaluation of vitrified grids to identify regions suitable for high-quality data acquisition (Figure 6). This step is usually performed using 200 kV microscopes, which offer sufficient resolution for assessing sample quality and allow preservation of 300 kV microscopes for high-resolution imaging.

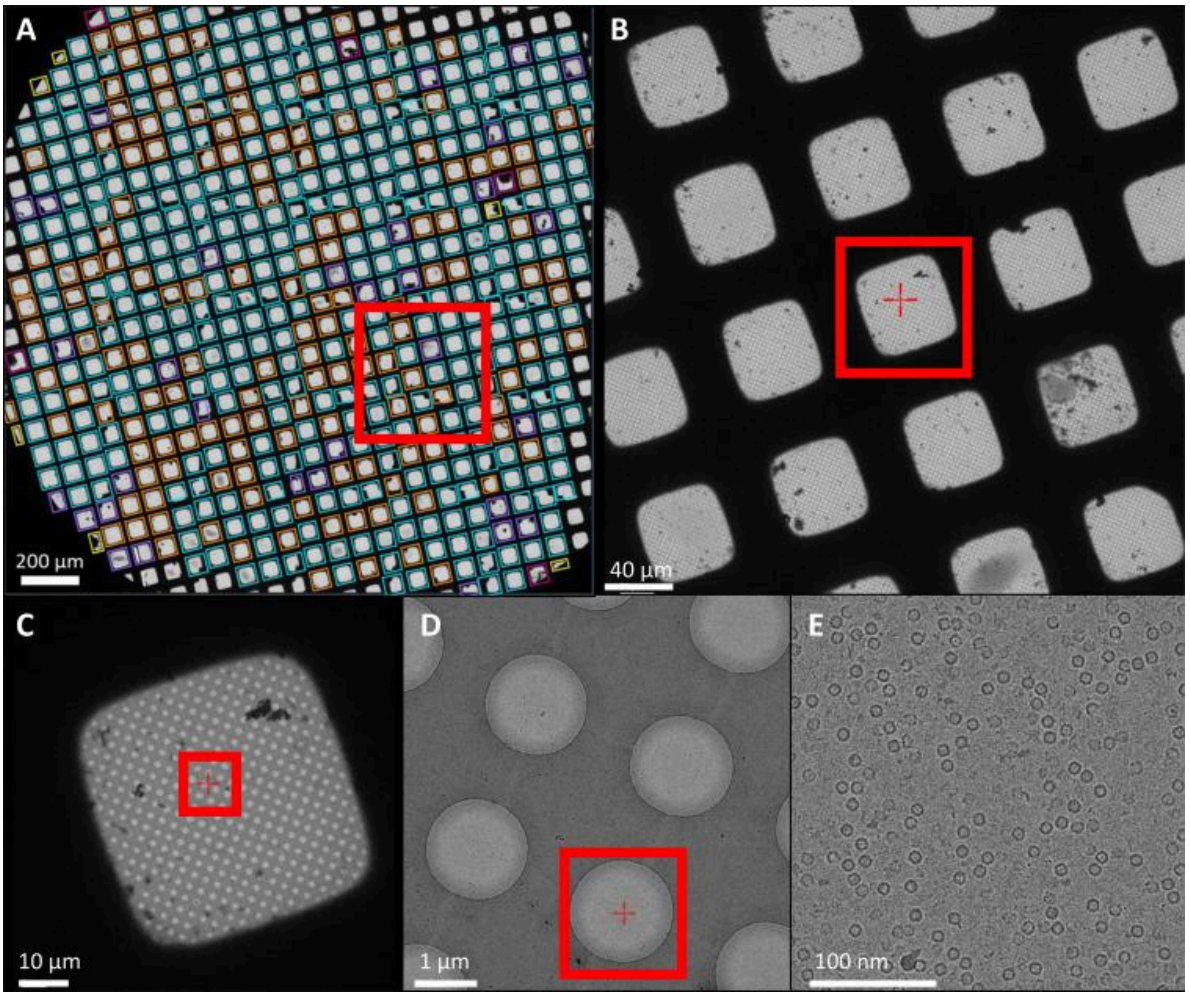


Figure 6. Screening grid evaluation at increasing magnifications. The low-magnification atlas image (A) provides an overview of the grid, enabling assessment of overall quality, ice thickness, and uniformity across grid squares (B–C). At higher magnifications, individual holes can be inspected (D), revealing the presence of a thin, vitrified ice layer suitable for high-resolution imaging (E) [50].

During screening, users assess:

- **Ice thickness**: Ensuring that the ice is thin enough for good contrast but not so thin as to damage the sample.
- **Particle distribution**: Particles should be evenly dispersed and non-overlapping.
- **Particle integrity**: Particles must show consistent morphology and random orientations, without aggregation.
- **Artifacts**: Regions with crystalline ice, carbon edge artifacts, or contamination should be avoided.

Screening is conducted at low magnification and under low-dose conditions to minimize beam damage. This preliminary step is a vital precursor to efficient data acquisition, ensuring that valuable microscope time is dedicated only to optimal grid areas [51].

High-Resolution Data Collection

Once promising grid squares are identified, high-resolution data collection commences, typically on high-end 300 kV instruments that provide a superior SNR for resolving fine structural details. Modern data collection is a highly automated process managed by software like EPU, Leginon, or SerialEM. These platforms enable unattended, high-throughput acquisition, systematically imaging hundreds or thousands of positions on the grid. Key components of high-resolution acquisition include:

- **Microscope calibration**: Ensuring accurate alignment of optics and stage movement.
- **Low-dose imaging**: Balancing electron exposure to preserve the sample while obtaining high-quality signal.
- **Automated acquisition software**: Tools like **EPU** (Thermo Fisher), **SerialEM**, **SmartScope**, and **Leginon** automate large-scale data collection across multiple grid regions with fine control over imaging parameters such as defocus range, exposure time, and beam/image shift.

The scale of modern cryo-EM experiments, which can generate terabytes of raw data in a single session, underscores the importance to automate pipelines for both data management and real-time quality assessment [39], [51].

1.2.3 Image Processing

The final stage of SPA transforms the raw movie data into a high-resolution 3D structure through a multi-step computational pipeline. The theoretical basis for this transformation is the Central Slice Theorem (or projection-slice theorem). This theorem stipulates that the 2D Fourier transform of a projection image is mathematically equivalent to a central slice through the 3D Fourier transform of the original object. Therefore, by collecting many projection images from different angles, one can computationally "fill in" the 3D Fourier space of the object (Figure 7). Assuming the particles adopt random orientations on the grid (i.e., no preferred orientation), a higher number of particles leads to better angular coverage of this volume, increasing the information content and the signal-to-noise ratio (SNR) of the final reconstruction. An inverse Fourier transform then yields the 3D structure in real space. A prerequisite for this is an accurate determination of each particle's orientation, which allows each "slice" to be placed correctly.

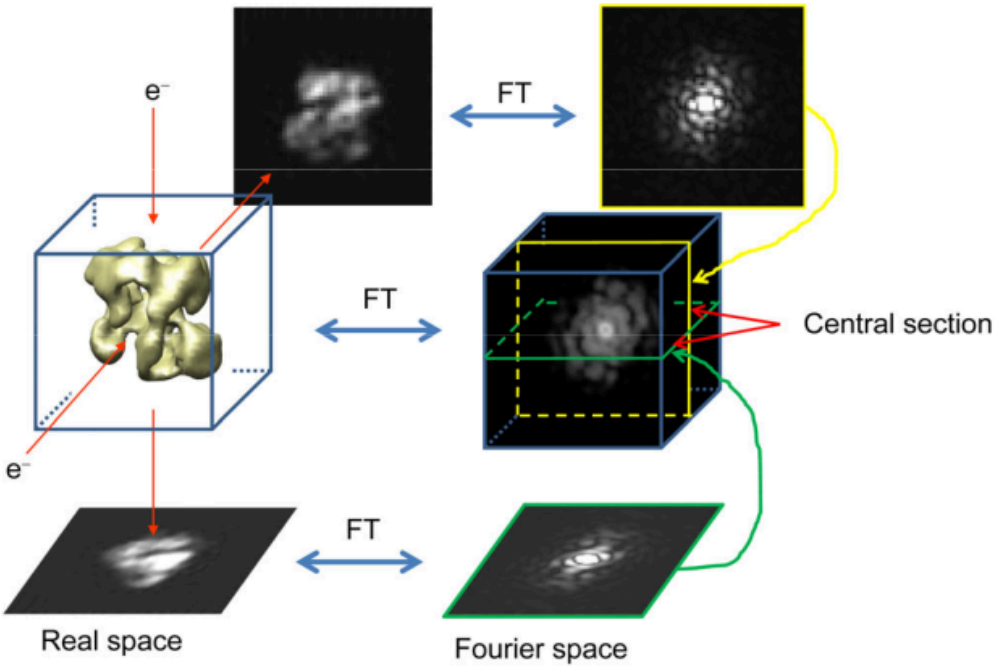


Figure 7. Explanation of the central slice theorem [1].

If certain orientations are missing (e.g., due to preferred orientation on the grid), the corresponding regions of Fourier space will be under-sampled, leading to information gaps (often called a "missing wedge" or "missing cone") that manifest as elongation or distortion in the reconstructed volume.

The standard SPA image processing pipeline includes the following key steps (Figure 8):

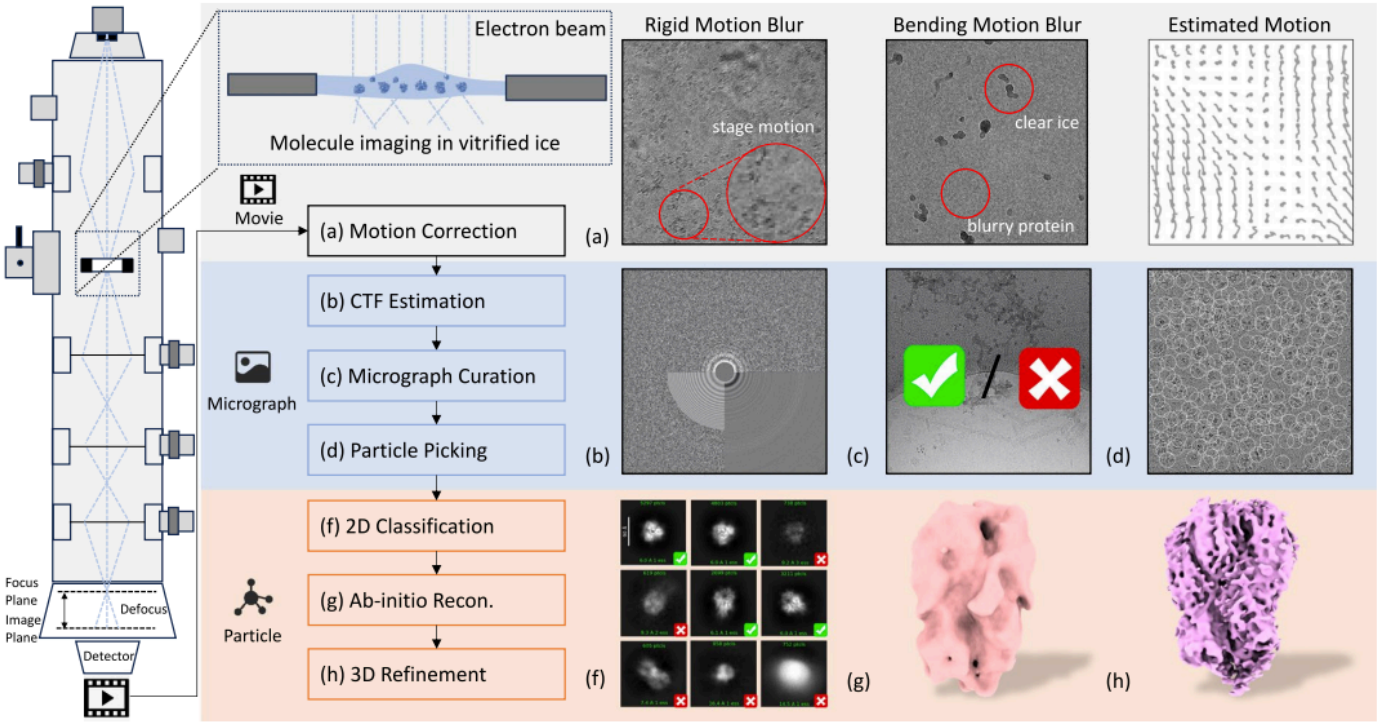


Figure 8. Overview of the SPA data processing pipeline. It covers several key steps: (a) estimation of rigid motion, bending motion and patch-wise local motion during movie alignment; (b) visualization of the 2D fitting results obtained during micrographs CTF estimation; (c) evaluation of micrograph quality and rejection of those unsuitable for downstream analysis; (d) particle picking across the micrograph; (f) 2D classification and averaging of the picked particles, with removal of low-quality subsets; (g) estimation of particle orientations and ab initio 3D reconstruction without prior pose information; and (h) 3D refinement of the reconstructed volume to achieve high-resolution results [52].

1. Movie alignment (Motion Correction)

Images are recorded as movies composed of multiple frames. During this exposure, the high-energy electron beam causes the specimen to move, which if left uncorrected, would blur the final image and destroy high-resolution detail. Motion correction algorithms [53], [54], [55] align these frames to a common reference, correcting for this stage and beam-induced motion. The process is challenging due to the extremely low SNR of individual frames. It is typically performed in two stages: a global correction for whole-frame drift, followed by a more localized correction that accounts for complex, non-rigid movements of individual particles (Figure 8.a). Local motion correction generally follows a divide-and-conquer strategy: the movie is split into small patches, and each region is aligned separately. The global and local motion shifts are then corrected by interpolating and summing values from all movie frames, producing the final motion-corrected micrograph [56].

This step is critical for retrieving high-frequency information that would otherwise be lost due to the motion-induced blurring [57]. For a deeper explanation of all these methods, a movie alignment review can be found in [58].

2. Contrast Transfer Function (CTF) Estimation and Correction

Due to the wave nature of electrons and the aberrations inherent to the electron optical system, cryo-electron microscopy (cryo-EM) images are not perfect projections of the specimen. Instead, they are modulated by the microscope's *contrast transfer function* (CTF), which describes how the microscope transmits contrast as a function of spatial frequency. The CTF introduces oscillations in the Fourier domain, attenuating or inverting certain frequency components of the image and thereby distorting both amplitude and phase information.

Biological macromolecules embedded in vitreous ice behave as weak-phase objects, since they are primarily composed of light atoms that weakly scatter the electron beam. As a result, images recorded in focus exhibit extremely low contrast. To enhance particle visibility, a controlled amount of underfocus is intentionally introduced during imaging to generate phase contrast. However, this defocus also modulates the image signal according to the CTF, reducing the reliability of high-resolution information due to the oscillatory nature of the transfer function. Therefore, cryo-EM imaging represents an intrinsic trade-off between contrast and resolution (Figure 9). Modern data collection strategies typically employ a range of defocus values to ensure complementary sampling of different spatial frequency bands across the dataset [59].

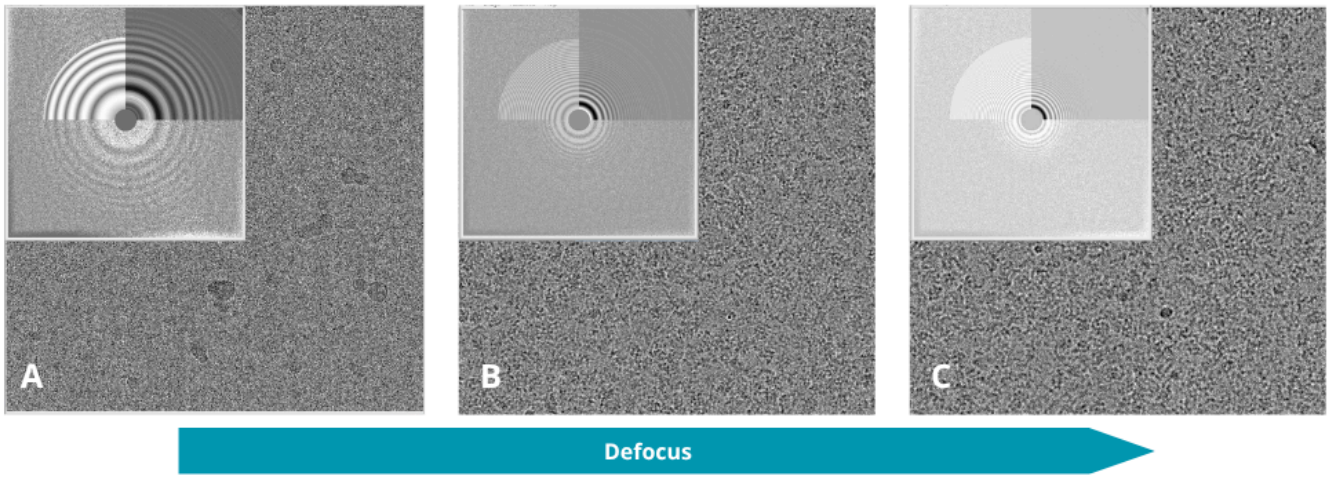


Figure 9. Defocus variation effect. Micrographs are from EMPIAR-11051. (A) Micrograph with an average defocus of 0.47 µm (B) Micrograph with an average defocus of 1.81 µm. (C) Micrograph with an average defocus of 2.80 µm.

The CTF depends on several instrumental and imaging parameters, including the accelerating voltage, spherical aberration coefficient, amplitude contrast ratio, and the defocus value

intentionally applied during data acquisition. Accurate estimation of these parameters is essential for computational correction of the CTF and subsequent recovery of high-resolution information.

CTF estimation algorithms such as ***CTFFIND4*** and ***Gctf*** determine these parameters by analyzing the characteristic *Thon rings* present in the power spectrum of each micrograph [60], [61]. These concentric rings arise from the interference between scattered and unscattered electron waves. By fitting a theoretical CTF model to the observed power spectrum, the defocus, astigmatism, and other relevant optical parameters can be extracted with high precision [57].

In practice, local variations in defocus across the field of view—caused by sample tilt, uneven ice thickness, or stage curvature—can significantly affect the accuracy of global CTF estimates. To address this, modern approaches employ **local or patch-based CTF estimation**, in which the micrograph is divided into smaller regions and the defocus is estimated independently for each patch. This refinement produces a spatially resolved CTF model that improves phase correction accuracy and overall reconstruction quality.

The potential resolution content of a micrograph is often assessed by computing the cross-correlation between the experimental power spectrum and the fitted theoretical CTF (Figure 10). This correlation serves as a confidence metric that typically decreases with increasing spatial frequency, reflecting the diminishing reliability of CTF fitting at higher resolutions. It is common practice to filter micrographs using this metric. For instance, we retain only those micrographs with a CTF confidence extending to at least 5 Å, thereby ensuring that subsequent image processing steps operate on data of sufficient quality [57].

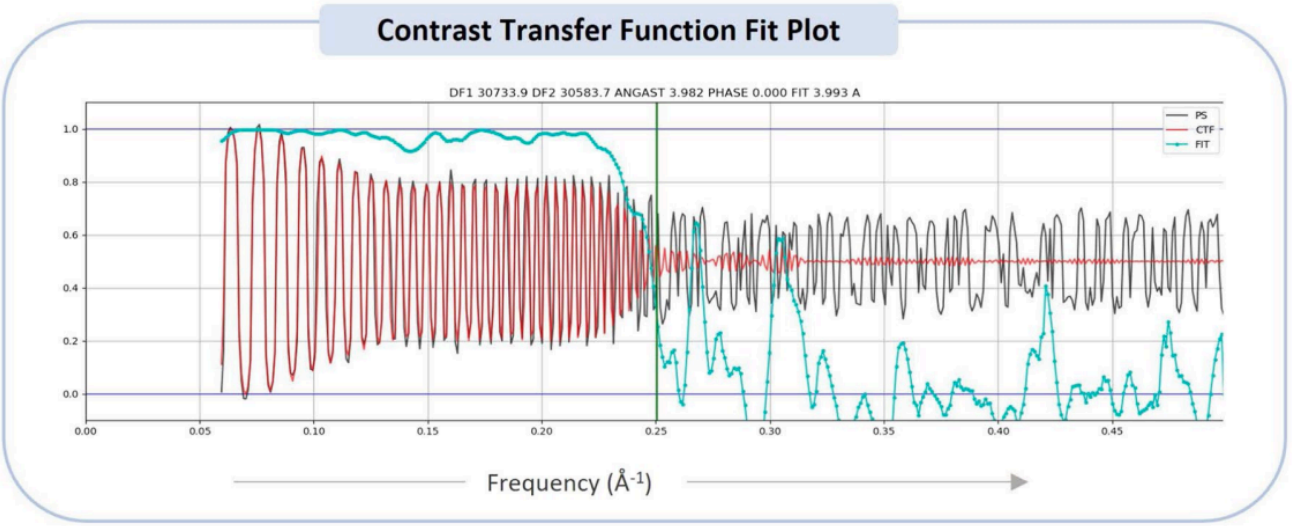


Figure 10. The CTF fit plot. "*X-axis displays frequency, in units in inverse angstroms (Å−1) and Y-axis shows correlation metric between power spectrum (PS) and CTF value. Black: observed experimental power spectrum. Red: calculated CTF. Cyan: cross-correlation (fit)*" [3].

Once CTF parameters have been estimated, computational correction is applied during the refinement and reconstruction stages to mitigate their effects. This correction, implemented mathematically in the frequency domain, partially restores the phase relationships and amplitude modulations imposed by the CTF. Accurate CTF correction is essential for preserving high-resolution information and ensuring the fidelity of the final 3D reconstruction. Nevertheless, challenges remain in modeling complex experimental factors such as beam-induced motion, specimen charging, and local variations in ice thickness. These limitations continue to motivate ongoing development of more robust and adaptive CTF estimation and correction algorithms, such as patch-based refinement methods, which aim to provide more reliable performance under diverse imaging conditions.

3. Particle Picking

Once the micrographs have been corrected for beam-induced motion and the microscope's contrast transfer function (CTF) is estimated, the next critical step is the identification and extraction of individual particle projections (Figure 11). This process, known as *particle picking*, directly determines the quality and reliability of all subsequent image processing steps, as inaccurate selection may introduce non-particle images such as noisy background, ice contamination, carbon edges, or broken particles that degrade reconstruction quality.

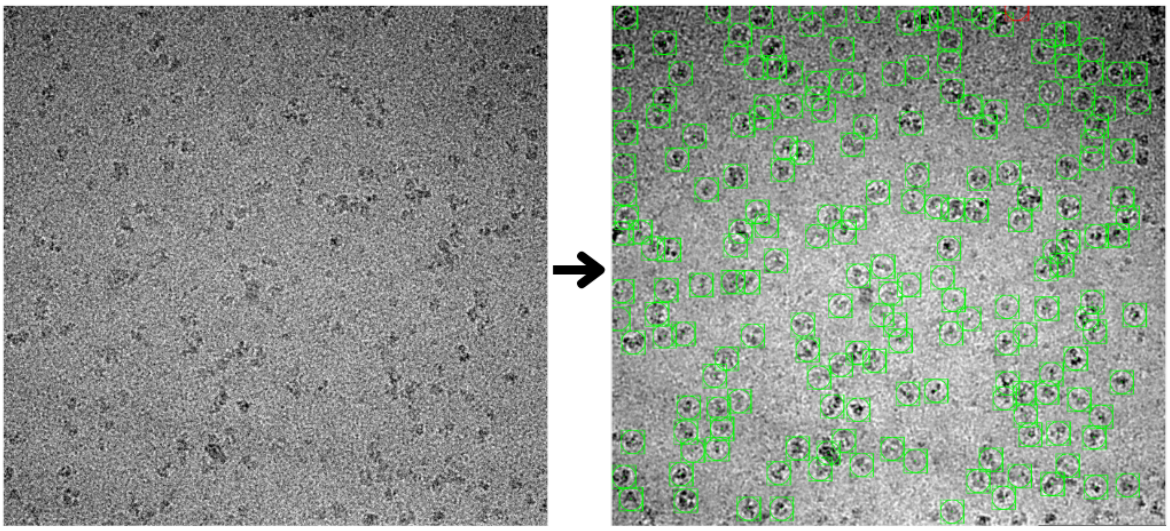


Figure 11. Particle picking process. The micrograph is from EMPIAR-10816.

Particle picking strategies can be broadly divided into three categories:

- **Manual picking:** performed interactively by the user on representative micrographs (e.g., in *Xmipp* [62]), ensuring accurate selection at the cost of being highly time-consuming and subjective.
- **Template-based or semi-automated picking:** uses a set of user-defined reference templates to locate similar features in other micrographs (e.g., *Gautomatch* [63], *Relion* [28]). This method speeds up particle identification but can introduce template bias.
- **Fully automated picking:** leverages algorithms that rely on image statistics, blob detection, or modern deep learning models. Recent tools such as *crYOLO* [33], *Topaz* [34], and *DeepPicker [64]* use convolutional neural networks trained on large datasets to accurately discriminate particles from noise with minimal user supervision.

Automated and deep-learning-based methods have significantly accelerated data processing while improving accuracy, particularly in large datasets with thousands of micrographs. However, the training and parameterization of these models still require careful validation. Generalization remains a persistent challenge. Models trained on standard proteins may perform poorly on new, unseen samples with different characteristics, highlighting the need for a strategy that can adapt to the unique properties of each dataset without requiring manual intervention [3], [52].

4. 2D Classification and Selection

The extracted particle images are typically noisy and represent a mixture of different particle orientations, conformations, or damaged molecules. The goal of **2D classification** is to group

similar particle projections together to improve the signal-to-noise ratio and identify well-defined structural views.

In **reference-free 2D classification**, as implemented in programs such as *RELION* [28], *cryoSPARC* [29], or *Xmipp* [62], particle images are iteratively aligned and averaged to produce *class averages* that are far more interpretable than individual particle images (Figure 12). Although the statistical approaches vary, RELION relies on a robust Expectation–Maximization (EM) algorithm, whereas cryoSPARC employs a much faster stochastic gradient descent (SGD) optimization, the overall strategy is similar and based on multi-reference alignment (MRA).

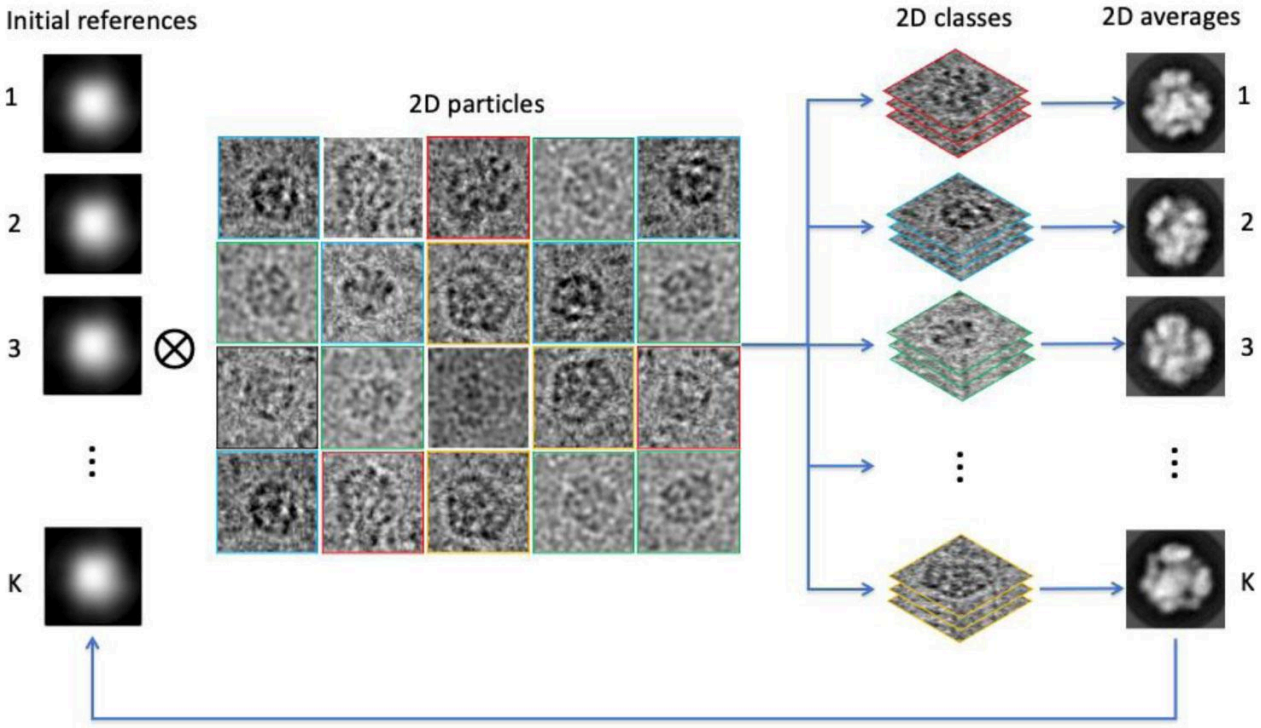


Figure 12. ***"Schematic representation of a 2D classification iteration"*** [2]. Data is extracted from [65].

The resulting averages reveal important information such as high-resolution features, symmetry, and the accuracy of the particle alignment. Poor or featureless classes, often corresponding to damaged particles, aggregates, or non-particle picks, are discarded, effectively purifying the dataset. Beyond removing particles, 2D classification also provides valuable insight into particle heterogeneity and preferred orientation issues. The final particle set is typically selected by visual inspection of 2D class averages. The curated subset of "good" particles emerging from this step is then used as input for generating an initial 3D model.

5. Initial 3D Reconstruction

From the selected particle stack, an initial three-dimensional (3D) model is estimated. This **ab initio reconstruction** is essential to obtain a first, coarse representation of the macromolecule, which will later be refined to higher resolution.

If no structural template is available, *de novo* reconstruction algorithms exploit the **central slice theorem**. As previously mentioned, this theorem states that each 2D projection corresponds to a central section of the 3D Fourier transform of the object. Consequently, a 3D reconstruction can be generated by assembling these central sections and applying a 3D inverse Fourier transform [66], [67].

Early methods relied on *common-lines* techniques, where intersections between Fourier transforms of 2D projections were used to infer relative angular orientations [26]. While effective at low noise levels, these methods were sensitive to false correlations under realistic cryo-EM conditions. Modern *statistical* approaches instead use iterative optimization of alignment parameters (translations and Euler angles), starting from a random initial "ball" model that progressively evolves toward a consistent 3D shape, as shown in Figure 13 [48]. These iterative procedures form the basis of projection-matching algorithms employed in modern software such as *RELION* [28], *cryoSPARC* [29], and *EMAN2* [68].

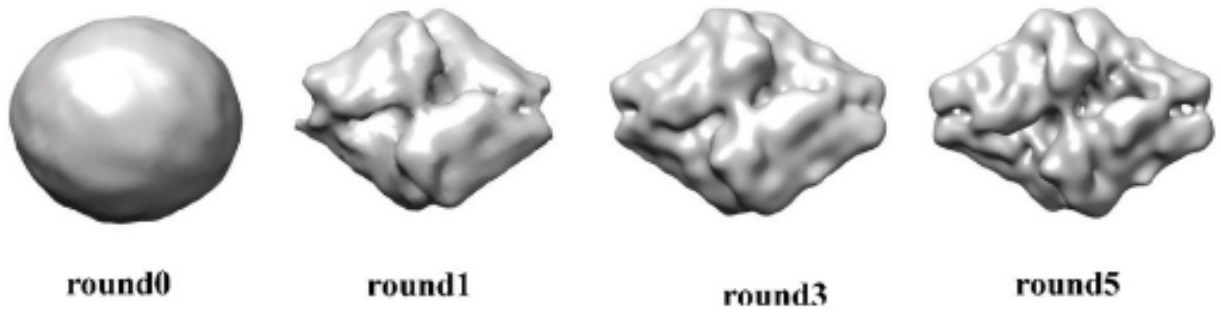


Figure 13. **Initial model generation.** The example shows the initial model of β-galactosidase. The process starts with a random model. The total cycle is five rounds. Each round includes one global search which is followed by refinement (modified) [69].

The accuracy of this first map is crucial: a poor or incorrect initial volume can bias the entire reconstruction process or hinder convergence, whereas a satisfactory initial model facilitates accurate angular assignment and high-resolution refinement.

6. 3D Classification and Refinement

To handle **structural heterogeneity** arising from conformational flexibility, compositional variability, or radiation damage, particles are further analyzed using **3D classification**. This step separates the dataset into structurally homogeneous subsets, each representing a distinct conformational or compositional state of the macromolecule (Figure 14).

In *RELION*, this is achieved through a **maximum-likelihood (ML3D)** framework [28], [70], [71], which estimates orientation and class membership probabilities rather than single deterministic assignments. This probabilistic approach accounts for uncertainty in angular assignment and naturally separates particles into consistent structural groups. Other statistical methods, such as **principal component analysis (PCA)** [72], covariance analysis [73], or manifold embedding [29], have been applied to capture continuous conformational changes.

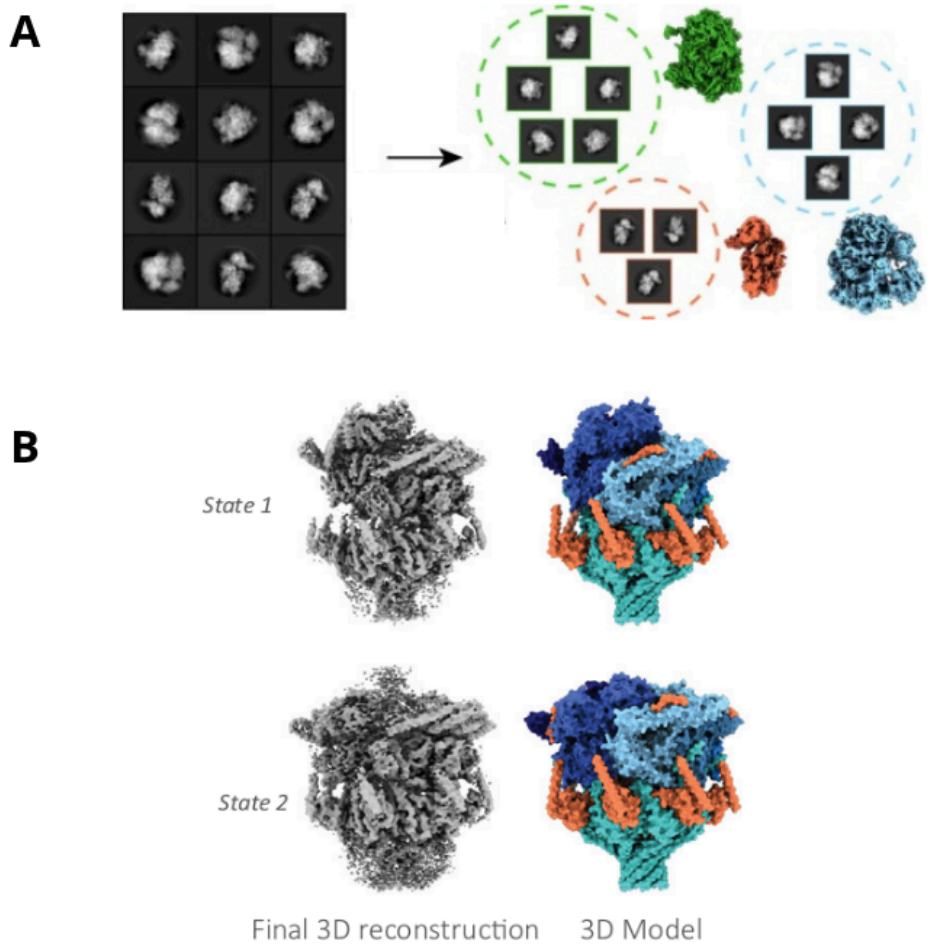


Figure 14. 3D classification examples under various scenarios. A) Biochemical heterogeneity: distinct structures of unrelated protein complexes obtained from a single cryo-EM dataset (modified) [74]. **B) Conformational heterogeneity:** high-resolution structures of the same protein complex captured in different conformational states (modified) [75].

After selecting the most homogeneous and well-populated 3D class, **3D refinement** is performed. This iterative process alternates between alignment of particle images to the current 3D reference and reconstruction of an updated map using the newly estimated orientations. Each iteration increases map resolution and sharpness (Figure 15.A). Advanced refinement schemes also incorporate **per-particle CTF refinement**, **beam-tilt and higher-order aberration correction**, and **Bayesian polishing** to correct for residual motion and improve high-resolution detail [76].

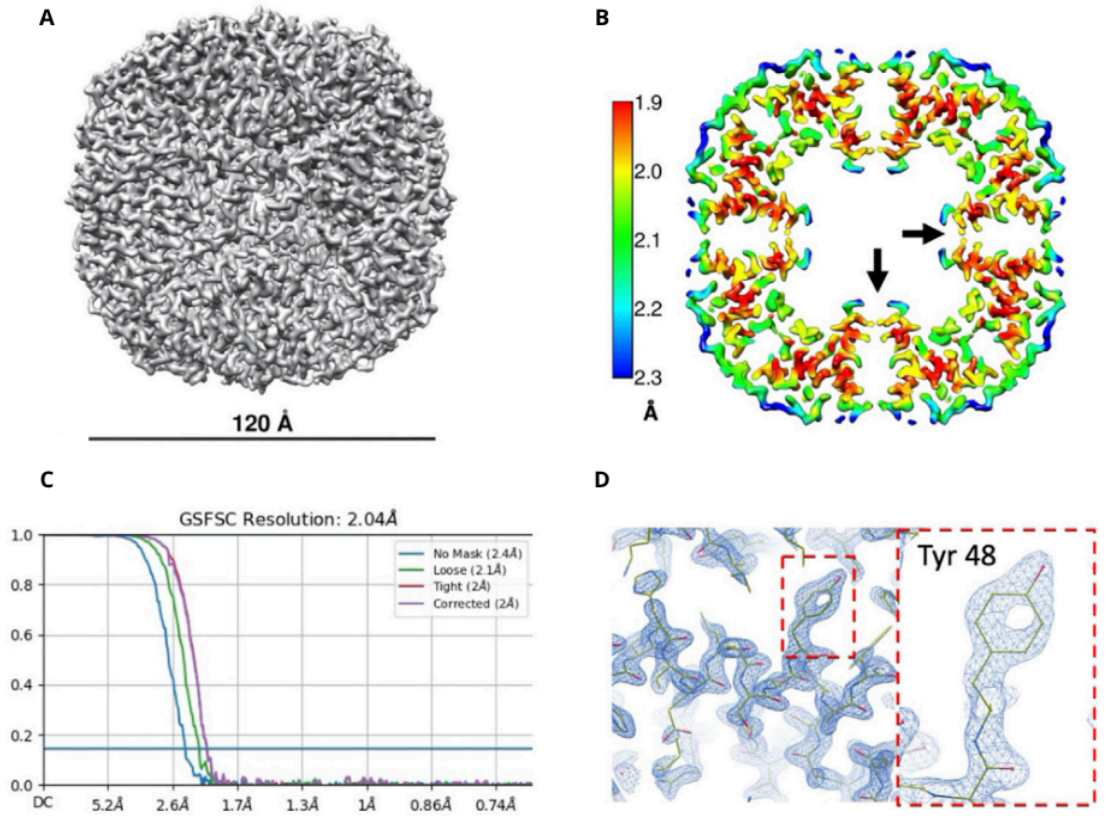


Figure 15. **An overview of 3D refinement and model fitting.** An example of cryo-EM data processing results for a protein complex (Flag-artemin). (A) Reconstructed density map. (B) Local resolution estimation highlights structural details within the map. (C) Resolution assessment using the Fourier Shell Correlation (FSC). (D) Atomic model fitting into the refined map in Coot (modified) [77].

The global map resolution is evaluated using the **gold-standard Fourier Shell Correlation (FSC)** criterion (Figure 15.C), which measures the correlation between two independently refined half-maps across spatial frequencies. These half-maps are generated from two separate half-datasets that are processed independently and later compared, ensuring an unbiased resolution estimate and preventing overfitting [78].

7. Validation and Atomic Modeling

When the final density map reaches near-atomic resolution (typically better than ~4 Å), it can be interpreted in terms of an atomic model. This process involves either **fitting known atomic structures** (rigid-body fitting) or **building a new model *de novo*** directly into the density map using programs such as *Coot [79]*, *Phenix* [80], or *ISOLDE [81]* (Figure 15.D).

The model is then refined against the experimental density while preserving stereochemical constraints. Structural validation includes:

- **Cross-validation** between the model and independent half-maps to assess overfitting.
- **Geometric validation**, checking bond lengths, angles, and torsions.
- **Local resolution estimation**, to ensure atomic interpretation is justified by local map quality.

Ultimately, the validated atomic model offers molecular insights into the macromolecule's function, dynamics, and interactions—fulfilling the primary objective of single-particle cryo-EM.

1.3 On-the-Fly Processing in SPA: Rationale and Challenges

The convergence of automated data acquisition systems (e.g., *Leginon* [82], *EPU* [39], *SerialEM* [38]) and hardware improvements, including automatic nitrogen filling, more stable lens and stage systems, has enabled continuous, unattended cryo-EM data collection over extended sessions (24-72h or more) [41]. This high level of automation has turned cryo-EM into a high-throughput technique, making large datasets more accessible to the structural biology community [57]. However, this success has created a new challenge: a "data deluge" that makes post-acquisition quality control inefficient and risky. Waiting until a multi-day session concludes to analyze the data can lead to the waste of invaluable (and scarce) microscope time and storage on a suboptimal sample or flawed collection strategy. "*The running costs and depreciation costs of a modern cryo-EM instrument is of the order of 5000 US$ per day*" [11]. This scenario has made on-the-fly, or real-time, data processing an essential tool for maximizing the efficiency and scientific output of modern cryo-EM facilities [57].

1.3.1 Motivation for Real-Time Feedback

To achieve high-resolution reconstructions in SPA, good-quality data must be collected. On-the-fly processing offers immediate feedback during the data collection session, enabling informed, data-driven decisions. This feedback is often organized in a **tiered approach**: basic pre-processing metrics followed by deeper image-processing assessment (Figure 16).

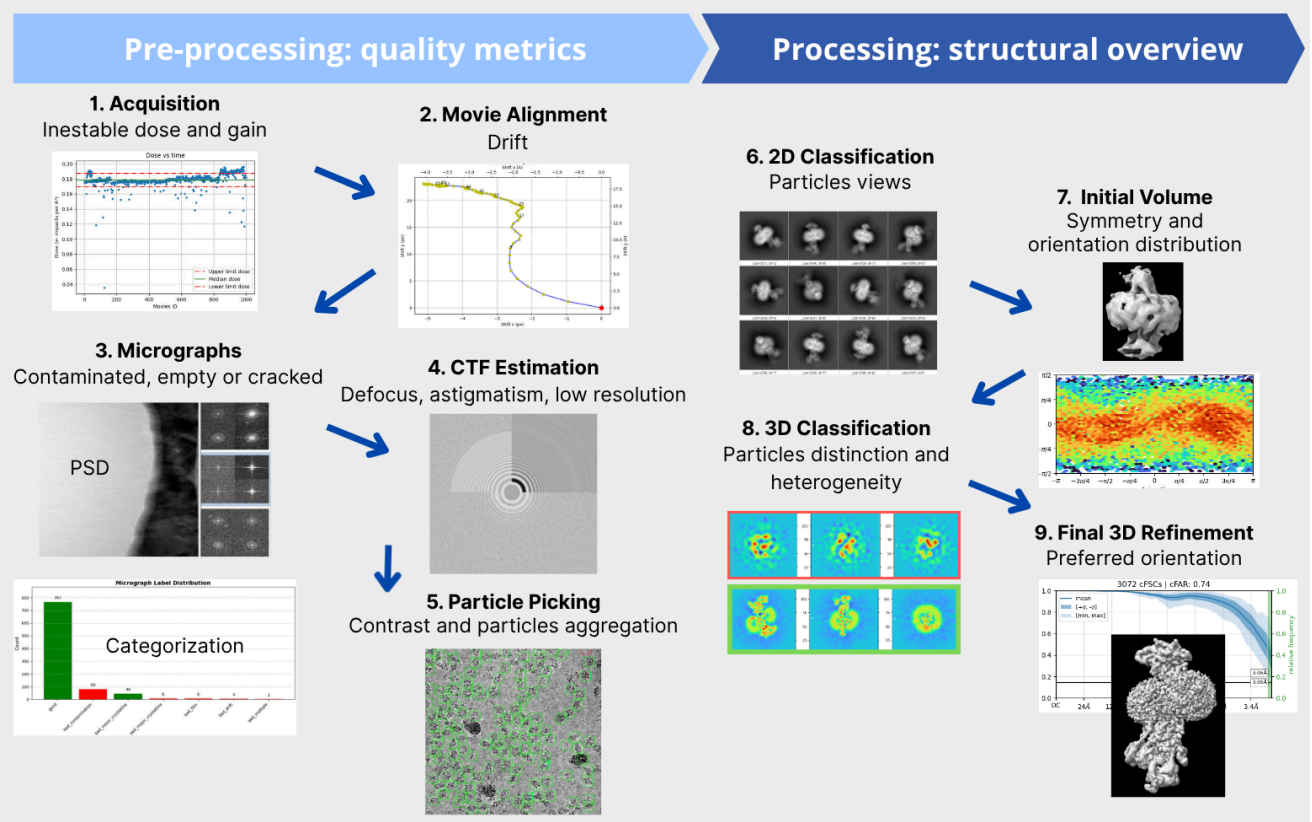


Figure 16. An Overview of the on-the-fly processing feedback. The image processing pipeline is divided into two main stages: the *Pre-processing stage* (quality metrics, left) and the *Processing stage* (structural overview, right). Each image processing step provides valuable information derived from the analyzed data. As the pipeline progresses, the feedback becomes increasingly informative, offering more profound insights into the sample's quality and structural features.

The first tier comprises motion correction, CTF estimation, and particle picking/extraction. These steps produce immediate metrics: drift and beam-induced motion once it is corrected [56]; CTF estimation reports defocus, astigmatism, ice thickness, and a confidence value for the resolution fit; micrographs are often excluded if the resolution fit does not extend to a satisfactory resolution (e.g., ~5 Å) [57]. Although metrics such as frame and global drift, CTF fit quality, or ice thickness provide a quick overview, they often are insufficient to diagnose complex sample issues. A dataset can appear acceptable based on these metrics yet fail to yield a high-resolution reconstruction [47].

Consequently, a second tier of feedback, which requires moving into 2D classification and preliminary 3D analysis, is often necessary. At this level, one can directly inspect the particles themselves: problems such as preferred particle orientation, biochemical heterogeneity, dissociation of complexes, or sample instability at the air-water interface often become evident only in 2D class averages or early 3D maps [57]. The most definitive assessment comes from a

quick **initial 3D reconstruction**. For example, a preferred-orientation problem that may be hinted at in 2D classes manifests clearly in a 3D map as anisotropic resolution, flattening, or missing view directions. This technique enables corrective actions, such as switching to tilted data collection, to be taken mid-session [83]. Aside from preferred orientation, some of the other factors mentioned often require further optimization of sample preparation, and a decision can be made as to whether to abort or continue with the data collection.

Conversely, real-time feedback can also signal when a dataset has reached sufficient quality/quantity, allowing the session to end early. Increasing data collection and instrumentation efficiency not only improves data quality and accelerates the rate at which samples are collected but also improves the utilization of other limited resources crucial for data processing including data storage, computational power, and human time spent on data processing [57].

1.3.2 Technical and Computational Challenges

Despite its clear benefits, implementing a robust on-the-fly processing workflow presents significant logistical and technical challenges related to data scale, computational infrastructure, and workflow complexity.

Data Volume and Velocity

Cryo-EM has firmly entered the era of "big data" [41]. Modern DED cameras, combined with high-speed acquisition strategies, can generate data at rates of more than 700 micrographs per hour, equivalent to roughly 3-6 terabytes of raw data per microscope per day [41], [57]. This data velocity places immense strain on the supporting IT infrastructure, as data transfer, storage, and computation must all operate in concert to avoid creating a system-wide bottleneck. The data generated throughout the SPA pipeline varies significantly in size:

- **Raw Movies:** The initial data product, typically ~2 GB per movie, represents the largest data component.
- **Micrographs:** After motion correction, each movie is converted to a single micrograph of around 50 MB.
- **Particle Stacks:** Collections of extracted particle images are the next major data component, and a stack of one million particles in 256x256 pixel boxes can exceed 250 GB. Datasets of several million particles are increasingly common [41].
- **2D and 3D processing**: After particle extraction, the usual workflow proceeds through 2D classification, 3D classification, and finally 3D refinement. The resulting 2D class

averages occupy very little storage, whereas each 3D volume typically requires on the order of tens of megabytes, around 50 MB per model [41].

- **Metadata (.xml, .star):** Smaller files containing essential information about acquisition parameters (defocus, astigmatism) and processing results are also stored.

Infrastructure and Computational Demands

On-the-fly processing is computationally demanding, requiring a dedicated high-performance computing (HPC) environment to keep pace with data acquisition. A typical pipeline requires a cluster with rapid networking, a parallel file system, and, most critically, significant CPU and GPU resources [84]. While early pre-processing steps like motion correction, CTF estimation and particle picking are intensive and must be executed for every incoming micrograph, the subsequent stages of 2D and 3D processing are highly parallel tasks that rely heavily on Graphics Processing Units (GPUs) [41]. The significant performance gains from GPU acceleration, originally driven by the gaming and machine learning industries, have been critical for the speed of modern cryo-EM software [30]. While improved algorithms have also substantially reduced the overall computational load, the widespread reliance on GPUs introduces new complexities.

Resource management poses a challenge: in multi-user facilities, GPU resources are shared, scheduling becomes complex, multiple stages needing GPU access may conflict, and memory overflow becomes a risk. Scaling hardware linearly with acquisition speed is cost-prohibitive for many centers. Thus, intelligent pipeline design and strategic processing choices become essential. For example, processing only a subset of micrographs for initial assessment, binning data during pre-processing, or delaying heavy tasks until after the session can reduce load without sacrificing actionable feedback [57].

Facility Constraints and Workflow Integration

In a multi-user facility environment, infrastructure must serve multiple microscopes and potentially dozens of users simultaneously. This creates several constraints:

- **Resource allocation**: Computational resources must be shared, and a single user's intensive on-the-fly job cannot be allowed to monopolize the cluster at the expense of others.
- **Data management**: Facilities must implement robust data management policies for transferring, archiving, and deleting the vast quantities of data generated daily.
- **Software heterogeneity**: Users may prefer different processing packages (e.g., *cryoSPARC Live, RELION, Warp, Xmipp*), and the facility's infrastructure must be flexible enough to support these diverse workflows and software integration.

Workflow Complexity and Human Factors

Beyond the hardware, cryo-EM data processing remains notoriously complex, often involving multiple rounds of human trial-and-error and dependency on human expertise. Operators make many subjective decisions–what parameters to set, which micrographs or 2D classes to retain–that can significantly affect outcomes [51], [85]. While an expert may navigate these choices efficiently, new users often find the process daunting. The sheer variety of biological samples and grid conditions makes it nearly impossible to create a universal, one-size-fits-all processing protocol [36].

Addressing these challenges requires the development of intelligent and automated systems that are reliable, flexible, and capable of minimizing the need for human intervention while still allowing expert oversight when necessary. This overarching problem—managing the interplay between infrastructure, software, and human workflows to ensure data quality and maximize the scientific return of high-throughput cryo-EM—defines the primary need for the next generation of on-the-fly processing pipelines.

1.4 State of the Art in Automated Image Processing Pipelines

The challenges posed by growing data volumes and the demand for rapid feedback have driven the development of sophisticated software pipelines designed to automate the SPA workflow. These systems seek to minimize manual intervention, provide real-time quality assessment, and efficiently manage the complex flow of data from the microscope to the final 3D reconstruction. Although many stand-alone programs exist for individual processing steps, the focus here is on integrated platforms that provide comprehensive, end-to-end solutions for on-the-fly processing. While these pipelines differ in design philosophy, flexibility, and degree of automation, they share a common objective: to make cryo-EM processing faster, more reliable, and more accessible.

1.4.1 Integrated Workflow Management Systems

A key development in the field has been the creation of workflow management systems that act as a "wrapper" or "glue" for various established cryo-EM programs. These platforms provide a unified environment for executing, tracking, and visualizing complex processing jobs while handling different file formats.

Scipion

Scipion is a flexible, plugin-based framework designed to integrate a wide array of cryo-EM software packages into reproducible workflows [84]. Its design philosophy is not to replace existing, well-validated algorithms but to provide a common platform where they can be executed and tracked, abstracting the user from the technical details of file formats and program execution. This makes it particularly well-suited for multi-user environments where portability, standardization, and reproducibility are paramount [86].

For on-the-fly processing, *Scipion* implements a "streaming" mode. It actively monitors acquisition directories for incoming data and triggers a user-defined workflow for each new movie or micrograph. A key feature is its execution engine, which can be configured to launch jobs in batches, preventing the computational cluster from being overwhelmed by thousands of individual jobs. This is achieved through a "chunking" mechanism, where micrographs are grouped together for processing, balancing the need for rapid feedback with efficient resource utilization [86].

The platform's graphical user interface (GUI) is central to its utility, allowing users to build complex pipelines visually and monitor their progress in real-time. *Scipion* provides a rich set of visualization tools, including plots of CTF parameters, particle distribution, and 2D class averages, which are updated dynamically as new data is processed [87]. This allows for immediate quality assessment. For instance, a user can quickly identify issues like poor CTF fits or particle aggregation and decide whether to adjust acquisition parameters or stop the session.

Scipion integrates a broad ecosystem of software tools—including MotionCor3, CTFFIND4, Gautomatch, Sphire, Miffi, RELION, and cryoSPARC—allowing users to combine the most suitable programs for each project. The framework manages all intermediate data and records extensive metadata, ensuring full traceability and reproducibility across the workflow [84]. While earlier implementations primarily focused on pre-processing and preliminary analyses [86], [87], Scipion's modular design enables the construction of complete, end-to-end pipelines capable of delivering a preliminary 3D reconstruction during the data-collection session. Despite these advances, achieving a fully automated, high-resolution pipeline that requires minimal user intervention remains an ongoing challenge.

1.4.2 All-in-One Processing Suites

In contrast to wrapper-based systems, several popular software packages offer a more self-contained, "all-in-one" approach, providing a complete suite of tools from pre-processing to refinement.

RELION

RELION (Regularised LIkelihood Optimisation) has become a benchmark in the field, particularly for its powerful Bayesian approach to 3D classification and refinement [28]. To facilitate real-time feedback, *RELION-4.0* introduced a dedicated on-the-fly processing pipeline built around a sophisticated scheduler that automates the workflow based on a user-defined configuration [37].

The pipeline operates through a dependency-based execution model. Instead of continuously streaming particles into running jobs, the scheduler launches new jobs as their required inputs become available. For example, once a set number of micrographs have been pre-processed, a subsequent particle picking job is triggered. This job-based approach is organized through predefined Schemes. The *prep Scheme* handles movie import, motion correction, and CTF estimation. The *proc Scheme* then takes over, selecting micrographs based on CTF quality, performing automated particle picking, 2D classification, automated 2D class selection, initial 3D model generation, and finally, 3D refinement [37].

This system offers a high degree of automation for key decision-making steps. For particle picking, it can use either its internal Laplacian-of-Gaussian (LoG) picker or an integrated version of the deep-learning tool *Topaz*. A significant innovation is the *relion_class_ranker* tool, which programmatically evaluates and selects the best 2D classes to generate a high-quality particle set for 3D analysis, a task that traditionally requires careful manual curation [37].

However, this automation is not entirely "hands-off" and has notable limitations. The entire process is guided by the *relion_it.py* script, which requires a user to provide several crucial parameters upfront. This includes an accurate estimate of the particle diameter for picking and extraction, as well as setting appropriate thresholds for 2D class selection (e.g., the minimum score for *relion_class_ranker*). Incorrectly setting these parameters can result in suboptimal outcomes, including improper particle picking when the particle diameter is poorly defined, the inclusion of junk particles, or the exclusion of certain particle orientations based on the 2D averages used in the final reconstruction. Therefore, while *RELION's* on-the-fly pipeline is powerful, its successful implementation still relies on a considerable degree of user expertise to guide the initial setup and ensure the automated decisions are based on sound parameters. Furthermore, the user can only modify the automated scheme programmatically in a nontrivial manner.

CryoSPARC

CryoSPARC (Cryo-EM Single-Particle Ab-initio Reconstruction and Classification) was developed with a strong emphasis on speed and user-friendliness, pioneering novel algorithms for fast *ab initio* 3D reconstruction from scratch [29]. Its on-the-fly implementation, "*cryoSPARC Live*", is designed for high-throughput feedback, providing a streamlined interface that displays key quality metrics, 2D class averages, and a continuously updating 3D reconstruction as data streams from the microscope.

The workflow begins with pre-processing but quickly requires user input to proceed. The user must define crucial thresholds for curating micrographs and for the blob-based particle picker, such as the minimum and maximum particle diameter, which can be difficult to estimate accurately for novel specimens. The user must also adjust blob picker thresholds to obtain effective particle picking. After an initial set of particles is picked, the user must manually inspect the results, curate 2D templates from these picks, and then launch a more accurate template-based picking job. Further progress is also gated by manual intervention: 2D classification, *ab initio* model generation, and 3D refinement jobs do not begin automatically and must be initiated by the user after they have manually selected the best classes or initial models [57]. While newly processed particles are incorporated into subsequent rounds of classification and refinement, this iterative process is supervised rather than fully automated.

This semi-automated approach has several limitations. The reliance on manual curation at multiple checkpoints creates bottlenecks and makes the process susceptible to user bias and error. Furthermore, CryoSPARC is a proprietary, closed-source product, which prevents its modification or integration with external tools. The processing pipeline is rigid; users cannot customize workflows or substitute alternative algorithms, which limits its flexibility for non-standard projects. While fast and powerful, its design trades the adaptability of open-source frameworks for a more controlled, but less flexible, user experience.

NextPYP

NextPYP is an open-source, end-to-end platform developed by the Bartesaghi Lab at Duke University that integrates single-particle cryo-EM (SPA) and cryo-electron tomography (cryo-ET) workflows within a unified environment [88]. Designed for large-scale, high-throughput processing, it combines some established community tools with in-house algorithms to achieve state-of-the-art results while maintaining user-friendliness and scalability. Its architecture features a modular workflow engine, a web-based graphical interface, and on-the-fly processing capabilities that perform motion correction, CTF estimation, and automated particle or sub-volume picking as data is acquired. The platform supports multiple particle-picking algorithms, refinements, and classification for both SPA and subtomogram

averaging and includes native integration with SLURM clusters. Distributed as a containerized package, it simplifies installation and ensures reproducibility across computational infrastructures.

NextPYP's main strengths lie in its all-in-one design, robust cluster integration, and simultaneous support for SPA and ET workflows. Its streaming capabilities and intuitive interface make it well suited for facility-level and real-time feedback pipelines. However, as a relatively new tool, it still has a smaller user base, and several SPA-specific processing steps lack the variety of well-established community algorithms, understandably, given its primary focus on cryo-ET pipelines. Furthermore, its containerized setup can limit flexibility for integrating external tools or customizing workflows. Overall, *NextPYP* complements established cryo-EM platforms such as *RELION*, *CryoSPARC*, and *Scipion* by offering a modern, scalable framework that aligns with ongoing efforts toward efficient, high-throughput image processing.

1.4.3 Specialized and Feedback-Oriented Pipelines

Other pipelines have been developed with a specific focus on optimizing certain aspects of the on-the-fly workflow, such as pre-processing speed or automated decision-making.

Warp

Warp is a software package that focuses exclusively on ultra-fast, GPU-accelerated pre-processing, designed to keep pace with, or even outpace, the highest data acquisition rates [30]. It performs motion correction, CTF estimation, particle picking, and micrograph denoising in a highly automated fashion. Key technical innovations include a motion correction algorithm that accurately accounts for specimen-tilting-induced blurring and a highly sensitive deep learning-based particle picker, BoxNet, which can be trained on-the-fly.

During a session, *Warp* provides real-time feedback on essential pre-processing metrics, such as defocus, astigmatism, and particle counts, allowing for immediate assessment of microscope performance and sample quality. However, its scope is intentionally limited. Warp does not perform any downstream processing steps like 2D classification or 3D reconstruction. Its primary output is a continuously updated stack of high-quality particle coordinates and extracted images.

This specialization makes Warp an exceptionally powerful front-end tool but not an end-to-end solution. To proceed with structure determination, the results must be exported to a full SPA suite like *RELION* or *CryoSPARC*. This modular approach is common in many facilities, where Warp's speed and accuracy are leveraged for pre-processing before handing off to

other programs for classification and refinement. Its main operational constraints are its requirements for NVIDIA GPUs [30], [57].

1.4.4 Other Automated Tools and Methods

Beyond fully integrated software suites, several specific tools and routines have been developed to address key bottlenecks in the on-the-fly processing pipeline, particularly in automated decision-making. These approaches often focus on replacing specific manual curation steps with robust, algorithm-driven methods.

TranSPHIRE

TranSPHIRE was designed with the ambitious goal of creating a highly automated, "hands-off" processing pipeline that incorporates feedback loops to optimize its own parameters and minimize human intervention [89]. Its core philosophy is to address the "human factor" challenge by automating decisions that are typically made by expert users. For example, after an initial particle picking with the integrated deep learning tool *crYOLO*, *TranSPHIRE* performs 2D classification and then uses a machine learning-based *Cinderella* model to automatically select *good* classes.

A distinctive feature of the TranSPHIRE workflow is its "feedback-optimized picking" strategy, which provides an internal feedback mechanism to improve particle picking over the course of a session. After an initial round of processing, a user can manually identify underrepresented views from the 2D class averages. *TranSPHIRE* then uses these selected views to automatically retrain the *crYOLO* picker, biasing it to discover more particles that resemble the desired, less-common orientations [89]. This represents a step towards more intelligent data processing, although it still relies on manual input to guide the feedback.

Despite its innovative design, *TranSPHIRE* has several practical limitations. The pipeline's reliance on generating a large number of 2D classes for its automated selection process can be time-consuming, with initial 3D reconstructions taking up to 15 hours, which may be too slow for rapid on-the-fly decision-making. The software is built around a fixed set of programs and is not designed for easy modification or the integration of new tools, which limits its flexibility. Furthermore, like Warp, it is not intended as an environment for final, high-resolution processing; results must be exported to other software suites for completion. Finally, as a less widely adopted platform, it has not been as extensively tested across a diverse range of biological samples as more established packages, and its success relies heavily on the performance of its automated 2D class assessment.

Others

A prominent example is the development of user-free routines that automate the selection of high-quality micrographs and particles. These systems replace subjective manual curation with quantitative, data-driven decisions. For micrograph selection, some systems employ sophisticated scoring functions to evaluate images based on multiple criteria, including the resolution of the CTF fit, estimated ice thickness, and particle density, and then automatically discard those unlikely to contribute to a high-resolution reconstruction. Similar strategies have been introduced for automatically scoring and filtering 2D class averages to provide a cleaned particle set without manual intervention. Furthermore, additional methods employ convolutional neural networks (CNNs) trained to classify both micrographs and 2D averages as 'good' or 'bad' based on their overall visual characteristics. These models effectively emulate the initial expert assessment traditionally required during manual curation [36].

1.4.5 Summary and Open Challenges

The pipelines and tools described in this chapter represent a significant leap forward in managing the complexity and scale of modern cryo-EM data processing. They have successfully automated many of the most repetitive and time-consuming tasks, enabling the real-time feedback that is critical for efficient use of high-end microscopes. However, a closer examination of the state of the art reveals several persistent gaps and challenges.

A primary issue is the **expertise bottleneck**. Despite the "automated" label, nearly all current pipelines require significant *a priori* knowledge and expert intervention to function optimally. Users must often provide critical parameters, such as particle size, symmetry, and various quality-control thresholds, before processing can even begin. Incorrect initial parameters can lead to suboptimal or even failed reconstructions, meaning the success of these pipelines still relies heavily on the user's experience with their specific sample. **Reliable parameter estimation** further complicates this issue: in practice, achieving robust and reproducible results often requires running multiple algorithms or repeating key processing steps to reach consensus. This approach, while increasing confidence in the results, also multiplies computational demands and slows down overall throughput.

This leads to a second challenge: the **flexibility-versus-automation trade-off**. Current solutions present a stark choice. Highly flexible, wrapper-based systems like Scipion provide immense power for customization but can be complex to configure for a truly automated, end-to-end workflow. Conversely, streamlined all-in-one suites like *CryoSPARC* and *RELION* offer greater ease of use and a high degree of internal automation, but their rigid, monolithic

design makes it difficult to integrate novel algorithms or customize workflows for non-standard samples.

Finally, very few systems provide truly **unattended, end-to-end processing**. Most on-the-fly pipelines are designed to deliver preliminary results—a set of curated micrographs or cleaned particles—and stop short of producing a high-resolution structure without additional manual intervention. The hand-off from real-time analysis to final refinement is often not seamless, and the initial automated steps may need to be repeated or adjusted before a final reconstruction can be completed.

Therefore, a critical gap remains in the field: the need for a processing pipeline that is not only automated but also **intelligent and adaptive**. Such a system should be capable of inferring key parameters directly from the data, dynamically adjusting its strategy based on real-time quality assessment, and integrating the best available tools within a flexible yet robust framework. The development of a pipeline that moves toward this ideal of a fully unattended, "*sample-to-structure*" workflow constitutes the central motivation for the work presented in this thesis.

CHAPTER 2 – MOTIVATION, OBJECTIVES, AND CONTRIBUTIONS

2.1 Motivation

As established in the previous chapter, single-particle analysis (SPA) by cryogenic electron microscopy (cryo-EM) has become a central technique in structural biology. The automation of data acquisition has transformed high-end electron microscopes into powerful, high-throughput instruments capable of generating terabytes of data in a single session. However, this success has shifted the primary bottleneck from data collection to data processing and, crucially, to data quality assessment.

Despite the existence of several on-the-fly processing pipelines, a significant operational gap remains. Researchers frequently invest days of valuable microscope time collecting massive datasets with little to no effective real-time feedback on the ultimate feasibility of achieving a high-resolution reconstruction. The feedback from existing pipelines is often limited to pre-processing metrics or requires significant manual intervention to proceed to later stages, failing to provide a conclusive verdict on sample quality. This delay between data collection and comprehensive quality assessment leads to the inefficient use of institutional resources, as entire datasets may be collected from suboptimal samples, only to be discarded days or weeks later.

The core motivation for this thesis is to address this important inefficiency. There is a clear and pressing need for a truly automated, end-to-end processing workflow designed specifically for on-the-fly analysis. Such a pipeline should function as a powerful diagnostic tool, providing researchers with rapid, actionable insights into their data during the acquisition session. This would empower them to make informed, data-driven decisions: to continue with confidence, to adjust collection strategy to overcome issues like preferred orientation, or to terminate a failing experiment early and return to sample optimization.

2.2 Objectives

To address the challenges outlined above, the central objective of this thesis is to design, implement, and validate a fully automated, unattended, and intelligent image processing pipeline for on-the-fly quality assessment and 3D exploration in cryo-EM.

This primary objective is composed of the following specific aims:

1. **Develop a comprehensive, end-to-end workflow** that proceeds from raw movies to a preliminary 3D reconstruction without requiring user intervention during the run.
2. **Build the pipeline within a flexible and extensible framework (Scipion)** to allow for the integration of best-in-class software packages, ensuring the pipeline remains state-of-the-art.
3. **Implement robust, multi-stage quality control** by integrating a cascade of automated filters at every major processing step and complement this system with a centralized Quality Dashboard that provides real-time, visual feedback on data quality, filtering decisions, and processing outcomes.
4. **Create a novel, data-driven particle-picking strategy** that automates box size estimation and trains a high-confidence picking model using a consensus approach.
5. **Design a hierarchical classification strategy** that robustly filters high-quality particles and automatically selects the most stable 3D class.
6. **Incorporate a self-consistency check** that validates the final particle set by correlating 2D class averages with projections from the refined 3D maps.
7. **Ensure the pipeline is scalable and deployable** in diverse computational environments, from local workstations to large, facility-scale HPC clusters.

2.3 Contributions

This thesis makes several significant contributions to the field of cryo-EM image processing, centered on the development of a novel solution that directly addresses the limitations of current pipelines.

The core achievement is **an intelligent and unattended diagnostic pipeline**. This robust, fully automated on-the-fly workflow functions as a powerful diagnostic tool, providing the timely, actionable feedback that has been largely missing from cryo-EM. The pipeline's intelligence is distinguished by key features, including a cascade of automated filters, a novel strategy for data-driven particle picker training that removes the need for manual parameter tuning, and a unique self-validating 3D analysis workflow that ensures the reliability of the results. The value of this central contribution is demonstrated through three key outcomes:

First, the pipeline exhibits **demonstrated generality and robustness**. Its effectiveness is not limited to a single sample type; its successful application to the majority of the 34 diverse datasets in the benchmark CryoPPP dataset, without any manual parameter tuning, proves its broad applicability and resilience to variations in particle size, shape, and sample quality.

Second, the work has undergone **real-world implementation and validation**. Beyond benchmark tests, the pipeline has been successfully deployed as the standard on-the-fly processing solution at high-throughput national cryo-EM facilities (ESRF, France). This implementation provides extensive evidence of its reliability, utility, and positive impact in a real-world scientific environment.

Finally, the pipeline provides users with **a seamless start for high-resolution refinement**. By delivering a complete and curated set of results, including high-quality micrographs, a trained picking model, cleaned particle stacks, and validated initial 3D maps, the workflow gives researchers a significant head start for subsequent expert-driven refinement, thereby accelerating the entire process of structure determination.

CHAPTER 3 – METHODOLOGY

This chapter details a novel, fully automated pipeline for on-the-fly cryo-electron microscopy (cryo-EM) data processing, designed to transform raw movies into preliminary 3D reconstructions without user intervention, specifically useful in the context of CryoEM facility environments. The core algorithms implementing the methods described were developed in Python and are integrated into the open-source Scipion framework, ensuring their broad accessibility and integration into the wider scientific community framework.

3.1 Scipion-based Automated Processing Pipeline

Scipion is a Python-based workflow engine that integrates a wide variety of structural biology software into a cohesive environment. Its plugin-based architecture wraps external software (*e.g.*, *CryoSPARC, RELION, MotionCor3, SPHIRE, Xmipp, CISTEM*, ...) into standardized plugins that contain protocols that will call specific image processing steps. Every protocol oversees all inputs, outputs, and parameter conversions, allowing them to be connected into a logical workflow.

Orchestrated within this environment, our pipeline integrates 37 protocols from 11 different software packages, all of which are integrated via *Scipion Plugins*. A key feature is its ability to initiate image processing upon data collection and to manage computational resources via job scheduling systems, such as SLURM, ensuring scalability from small workstations to large HPC clusters. The pipeline is divided into four principal stages (Figure 17):

1. **Data Curation and Quality Control:** Initial assessment and filtering of raw data.
2. **Automated Particle Picking and Model Training:** Identification and extraction of particles from curated micrographs.
3. **Initial 2D and 3D Analysis:** 2D classification of particles and parallel generation of *de novo* 3D models.
4. **Refinement and Parallel Validation:** Refinement of the initial 3D models, coupled with robust validation procedures.

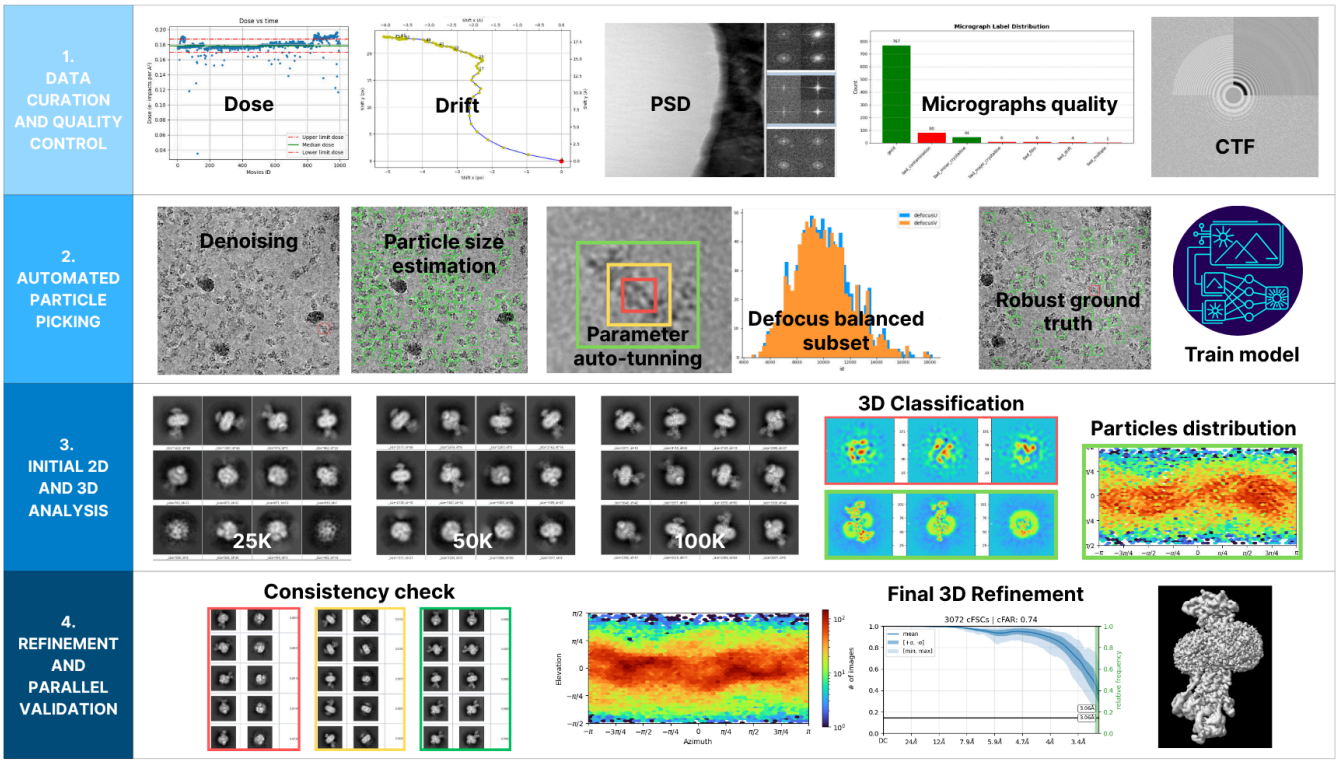


Figure 17. Graphical summary of the four main stages of the automated image processing pipeline.

3.2 Data Curation and Quality Control

The first stage of the pipeline is critical for the success of any high-resolution reconstruction. It comprises a cascade of automated filtering steps to ensure that only the highest quality data proceeds to downstream analysis (Figure 18). This is achieved by assessing raw movies and aligned micrographs using information from both real and Fourier space.

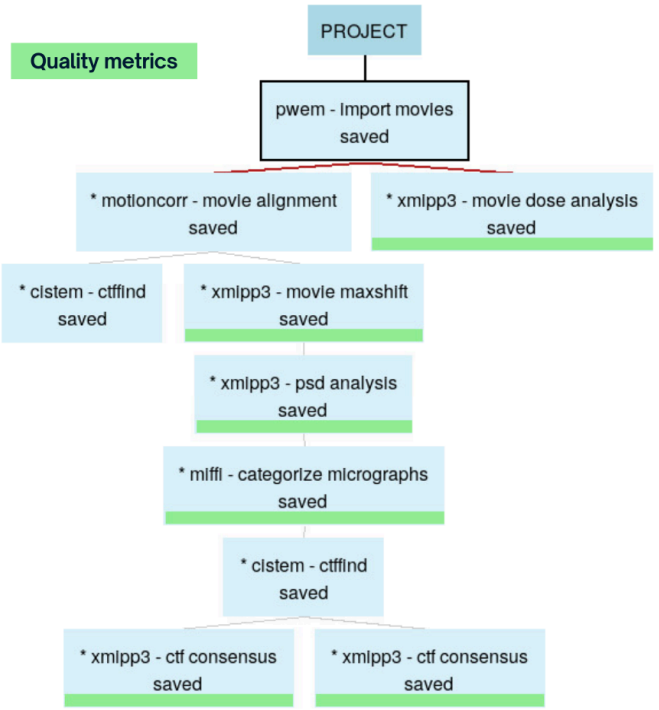


Figure 18. A detailed Scipion workflow diagram for the Data Curation stage. Each box represents a protocol in Scipion, where each protocol corresponds to a specific image processing step in the pipeline. Quality filters used during data curation are highlighted with green labels.

3.2.1 Movies-level curation

The process of generating a cryo-EM image begins at the detector. Modern direct electron detectors (DEDs) are sensitive enough to capture the impact of individual electron events. Rather than a single, long exposure that would be blurred by sample movement, data is collected as a "movie", a rapid series of short-exposure frames. The raw movie data itself, before alignment, contains valuable information about the data collection process.

Movie Dose Analysis Program

A critical parameter that can be extracted at this early stage is the electron dose rate. While the total accumulated dose determines the overall signal-to-noise ratio and the extent of radiation damage, the rate at which this dose is delivered across the movie frames gives clues about the stability of the system. An ideal acquisition involves a stable and consistent electron dose delivered to the sample over the entire exposure time. Monitoring the average dose per frame is therefore a powerful diagnostic tool. Significant fluctuations can indicate instability in the electron gun or the microscope's illumination system, allowing for early intervention and preventing the collection of suboptimal data (Figure 19).

Furthermore, this measurement serves as a valuable proxy for the thickness of the vitrified ice. When the electron beam traverses the sample, some electrons are scattered by both the ice and the biological macromolecules embedded within it. Thicker ice will scatter more electrons, resulting in a lower average number of electrons reaching the detector for a given incident beam intensity. By measuring the mean dose per frame for each movie and plotting this value over the course of a data collection session, one can immediately identify trends or outliers. This allows for the detection of areas on the grid with inconsistent ice thickness, a common factor that can compromise the final resolution of the 3D reconstruction.

To automate the assessment of dose stability and ice thickness, we developed the *movie dose analysis protocol*. The program calculates a representative dose value for each incoming movie and uses a statistical approach to flag outliers in real-time. The core of the algorithm is the calculation of the mean dose per square Ångström ($e^-/Å^2$) for each movie. Let M_i be the *i-th* movie in a dataset, and *s* be the sampling rate (pixel size) in Å/pixel. To optimize processing time, reducing the analysis from approximately 8 seconds to 2 seconds per movie, the program avoids reading all frames. Instead, it samples a small subset of frames (specifically the first, middle, and last). For each sampled frame, the program reads the image data, represented as an array of pixel values, and calculates its mean pixel intensity, denoted as $\overline{p}$

$$\overline{p} = mean(image\ array)$$

This value is then converted to the mean dose per unit area, *d,* using the pixel size, *s*:

$$d = \frac{\overline{p}}{s^2}$$

The representative mean dose for the entire movie, $\overline{d}_i$, is then calculated as the average of these dose values from the three sampled frames. To establish a stable baseline for comparison, the program first processes an initial, tunable batch of *n_samples* movies (typically 20). It then calculates a robust global reference dose, μ, calculated as the median of the representative doses, $\overline{d}_i$. The median is used specifically for its robustness against anomalous outliers:

$$\mu = median(\{\overline{d}_1, \overline{d}_2, ..., \overline{d}_{n_{samples}}\})$$

Once this global median μ is established, the program defines an acceptance interval based on a user-defined percentage difference thresholds, τ (e.g., 5%). A subsequent movie M_k with a mean dose $\overline{d}_k$ is accepted if it falls within this interval, and rejected otherwise:

$$Accept\ if\ \ \mu \bullet \left(1 - \frac{\tau}{100}\right) \leq \bar{d}_k \leq \mu \bullet \left(1 + \frac{\tau}{100}\right)$$

To adapt to gradual changes in experimental conditions, a sliding window mechanism periodically recalculates μ using all movies processed thus far. This ensures that the reference value can adapt to slow, legitimate drifts in experimental conditions over a long data collection session, preventing an initial, potentially unrepresentative, estimate from skewing the entire analysis. Simultaneously, the program assesses the quality of the data within this most recent window. It calculates the percentage of movies that were rejected. If this 'faulty percentage' exceeds a predefined threshold (*i.e.,*30%), a warning is logged. This serves as a critical alert for the microscope operator, indicating a potential acute issue, such as a sudden change in ice quality or beam instability, that requires immediate attention. Since the optimal ice thickness is protein-dependent, this tool is used to monitor data collection quality rather than to discard movies outright. The program is integrated into our workflow via *Xmipp* software plugin [90], scipion-em-xmipp, more specifically, as a movie *dose analysis protocol*.

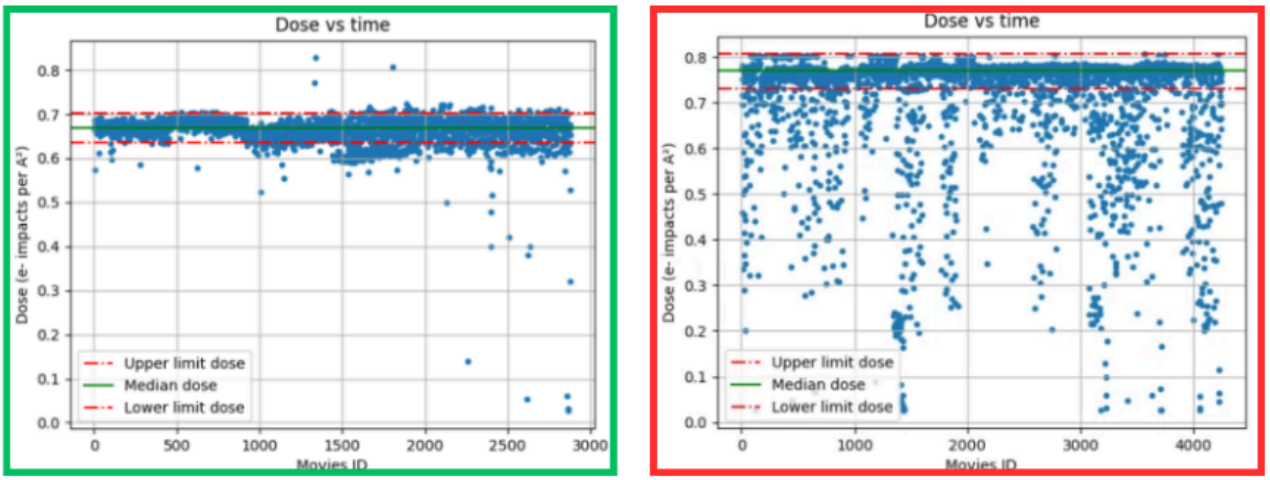


Figure 19. Examples of dose analysis plots. Highlighted in green is an optimal acquisition example showing a stable and consistent electron dose. Highlighted in red is an unstable acquisition, displaying clear dose inconsistencies during data collection. Such erratic behaviour can indicate issues with the electron source or variations in ice thickness. The global median (μ) is shown as a green line, while user-defined difference thresholds are shown in red, marking the movies that do not meet the filtering criteria.

3.2.2 Motion Correction and Drift Monitoring

During an electron exposure, the sample suffers from continuous movement. This motion, if uncorrected, blurs the final image and destroys high-resolution details [56]. Motion can be categorized as rigid mechanical drift, which often refers to the mechanical drift of the sample

stage during imaging [55] and non-rigid bending, which is associated with the beam-induced deformation of the sample [59]. Motion correction algorithms are adequate, but excessive or rapid drift can cause irretrievable loss of high-resolution information. Such a situation can degrade the quality of the final 3D reconstruction [56]. Moreover, a sustained high level of drift over time may also be an indicator of more severe problems related to the mechanical stability of the microscope itself. It is therefore crucial to monitor the magnitude of motion and discard movies with excessive drift.

Movie Max Shift Protocol

To quantify drift, we implemented a calculation that runs immediately after frame alignment (*MotionCor3 [53]*). It computes two key metrics from the absolute shifts (x_j, y_j) of each frame j relative to a reference.

A. **Maximum Frame-to-Frame Shift (S_{frame}):** This metric identifies sudden movements by quantifying the largest displacement between any two consecutive frames. Let $l_j = \sqrt{(x_j - x_{j-1})^2 + (y_j - y_{j-1})^2}$ be the Euclidean distance (scalar shift) between consecutive frames j and $j - 1$. S_{frame} is defined as the maximum of these shifts over the entire movie, converted to Ångströms using the pixel size s.

$$S_{frame} = max_{j=2,\dots,N}(l_j) \bullet s$$

B. **Accumulated Movie Shift (S_{movie})**: This metric quantifies the total path length traveled by the specimen during the exposure, capturing non-linear and continuous drift more accurately than a simple start-to-end vector. It is calculated as the sum of all incremental shifts l_j:

$$S_{movie} = (\sum_{j=2}^{N} l_j) \bullet s$$

The protocol allows for flexible rejection criteria based on these metrics: movies can be discarded if *either* threshold is exceeded (OR logic), if *both* are exceeded (AND logic), or based on a single metric. To ensure high-quality data input for downstream processing, we employed an "OR" logic: movies are discarded if either S_{frame} or S_{movie} surpasses its respective limit (Figure 20).

Both the rejection criteria and the specific thresholds are user-configurable parameters, allowing the protocol to be tailored to different experimental setups. The default thresholds were determined empirically by observing the filter's performance across multiple datasets and

acquisition strategies. For standard data collection, we set S_{frame} = 10 Å and S_{movie} = 45 Å. These values are not intended to be hypersensitive cutoffs but rather robust safety nets designed to identify and remove micrographs with gross drift issues that preclude high-resolution information recovery. Consequently, the pipeline is relatively insensitive to minor variations in these parameters; small adjustments to the thresholds do not significantly alter the final reconstruction quality, as the primary goal is to eliminate statistical outliers rather than borderline cases.

For experiments involving tilted data collection, where drift is inherently higher and resolution expectations are often lower, we propose relaxed thresholds: a maximum frame-to-frame shift of S_{frame} = 30 Å and a global accumulated drift of S_{movie} = 120 Å.

This program is also integrated into the *Xmipp* software plugin as *movie max shift protocol*.

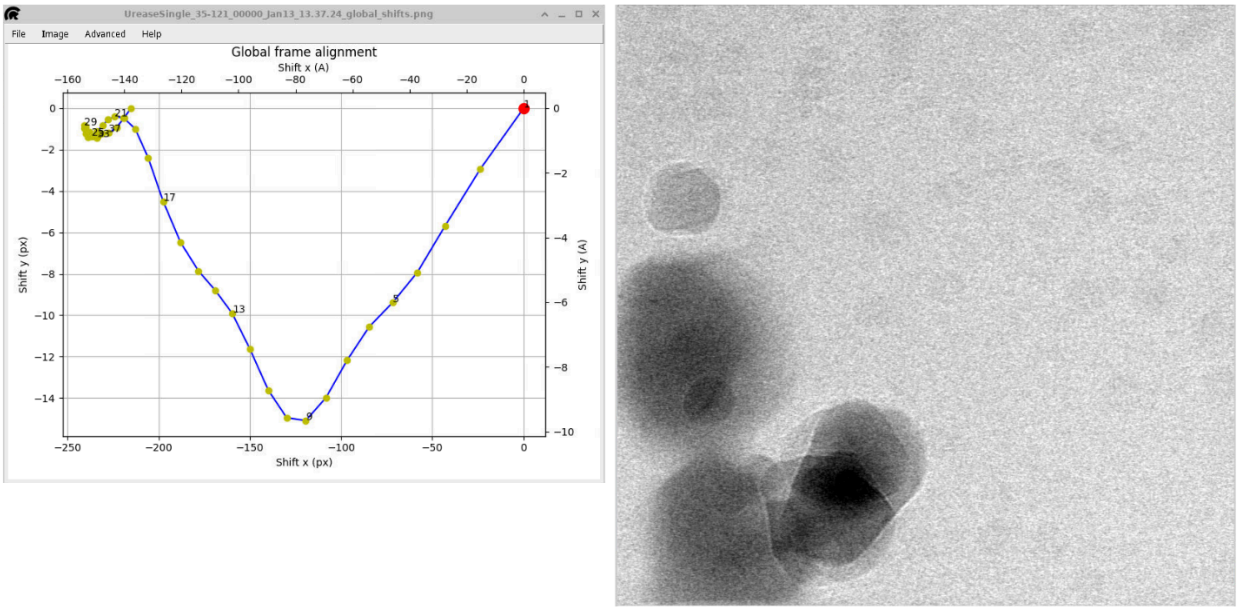


Figure 20. Example of a micrograph discarded by the Max Shift Filter. On the left, the trajectory plot shows the global alignment path in both the X and Y directions. Each yellow dot represents an individual frame, while the blue line indicates the movement between consecutive frames. On the right, the corresponding faulty micrograph is shown, exhibiting a large maximum drift per frame (15 Å) and a substantial overall global drift (154 Å). Micrograph is from EMPIAR- 10389.

Movie alignment algorithms complete study

Movie alignment is often the most computationally demanding step in the initial stages of the Single Particle Analysis (SPA) workflow, frequently representing the primary processing bottleneck. For this reason, this thesis included a comprehensive study to characterize the

performance of leading alignment algorithms and identify optimal processing strategies as a crucial component. The results of this work were published in *"Performance and Quality Comparison of Movie Alignment Software for Cryogenic Electron Microscopy"* [56]. The study aimed to understand not only the quality of the results produced by different algorithms but also their computational performance, a critical factor for designing an efficient on-the-fly processing pipeline. The main recommendations derived from this study directly influenced our automated workflow.

The performance analysis revealed several key factors that are paramount for achieving the throughput necessary for on-the-fly processing. Firstly, **GPU scaling** is essential. While all tested programs benefit from GPU acceleration, their ability to scale across multiple GPUs is crucial. The study revealed that for every algorithm tested, by simply adding more GPUs it does not guarantee a linear increase in performance. Beyond a certain point, the system becomes limited by I/O (Input/Output) throughput, where the speed of reading movie files from storage cannot keep up with the processing speed of the GPUs. Understanding this interplay between computational power and I/O capacity was vital for designing a balanced, cost-effective system and for optimizing the movie alignment process on large-scale cryo-EM facilities.

Secondly, the study highlighted the dramatic impact of **storage speed**. The read/write operations for large movie files can easily become the rate-limiting step, a fact that is exacerbated when using multiple GPUs. Our findings strongly recommend the use of high-speed Solid-State Drives (SSDs), for temporary storage during on-the-fly processing. The performance gains observed when moving from traditional hard drive storage (HDDs) to local SSDs were substantial, directly addressing the I/O bottleneck.

Thirdly, **batch processing** emerged as a vital optimization strategy. Submitting movies for alignment in batches rather than individually allows for more efficient resource utilization, particularly for GPU memory and initialization overhead. This approach minimizes idle time and maximizes throughput, a principle that has been integrated into our pipeline's job scheduling system.

Combining these insights, the study provides a clear roadmap for building a high-performance on-the-fly processing system. The core recommendation is that to keep pace with modern detector acquisition rates, a combination of efficient multi-GPU scaling, fast SSD storage, and optimal batch processing is not just beneficial but essential. These findings provided the evidence-based foundation for the hardware and software configuration of the automated pipeline described in this thesis, ensuring it is capable of handling the data deluge from cryo-EM facilities.

3.2.3 Micrographs-level curation

Once we have our aligned images, micrograph curation is a crucial subsequent step for any high-resolution 3D reconstruction. Due to persistent challenges such as uneven ice thickness, low signal-to-noise ratio (SNR), and sample impurities, a significant fraction of micrographs in a typical dataset are of insufficient quality for downstream processing. Traditionally, curation relied heavily on metrics derived from the Contrast Transfer Function (CTF) estimation, using thresholds for defocus, astigmatism, and goodness-of-fit to filter the data. While useful, these methods can be inconsistent. More recently, deep learning (DL) approaches have demonstrated superior efficiency and accuracy in this task [35], [91]. As demonstrated by tools like Miffi, DL-based methods can significantly outperform traditional CTF-based filtering, "*setting a new standard for automated micrograph curation*" [52].

To create a robust filtering system, our pipeline integrates a hybrid approach that leverages the complementary strengths of information from both Fourier space and real space. A micrograph and its Fourier transform are two mathematically linked representations of the same data. Analyzing both provides a more complete assessment of quality than either view alone, a foundational concept in image processing [66].

- **Real Space:** Visual inspection of the micrograph directly reveals the spatial distribution of particles and the presence of gross contaminants, such as large ice crystals or carbon film edges. It provides the most direct and intuitive assessment.
- **Fourier Space (PSD):** The Power Spectral Density (PSD) is the squared magnitude of the micrograph's Fourier transform, visualizing the image's spatial frequency components. This view is exceptionally sensitive to periodic signals and global image properties. The most critical feature in the PSD of a good micrograph is the presence of Thon rings, which are the concentric rings corresponding to the zeros of the microscope's Contrast Transfer Function (CTF). The visibility and extent of these rings are a direct indicator of the micrograph's potential resolution; the further the rings extend from the center, the higher the resolution of the information contained within the micrograph. Furthermore, the shape and integrity of the Thon rings serve as powerful diagnostic tools: elliptical rings reveal astigmatism. At the same time, directional streaking or blurring in the PSD can indicate uncorrected sample drift. A complete absence of Thon rings implies a severe lack of high-resolution signal, often due to empty foil holes, very thick ice, poor contrast, or significant drift [60].

Micrograph Power Spectrum Density (PSD) Analysis Protocol

To assess signal quality in Fourier space, we developed a program that analyzes a micrograph's PSD. This tool quantifies the consistency and quality of the signal across a micrograph by analyzing its Power Spectrum Density (PSD). The algorithm partitioned the micrograph into four corner quadrants. For each quadrant, a series of image processing steps are applied: the quadrant is first normalized to have a zero mean and a standard deviation of one. Then, its PSD is computed and subsequently cropped to a user-defined resolution limit (typically 4Å) to focus on the most informative frequency range. Finally, a low-pass filter is applied to reduce high-frequency noise and enhance the visibility of the Thon rings (Figure 21).

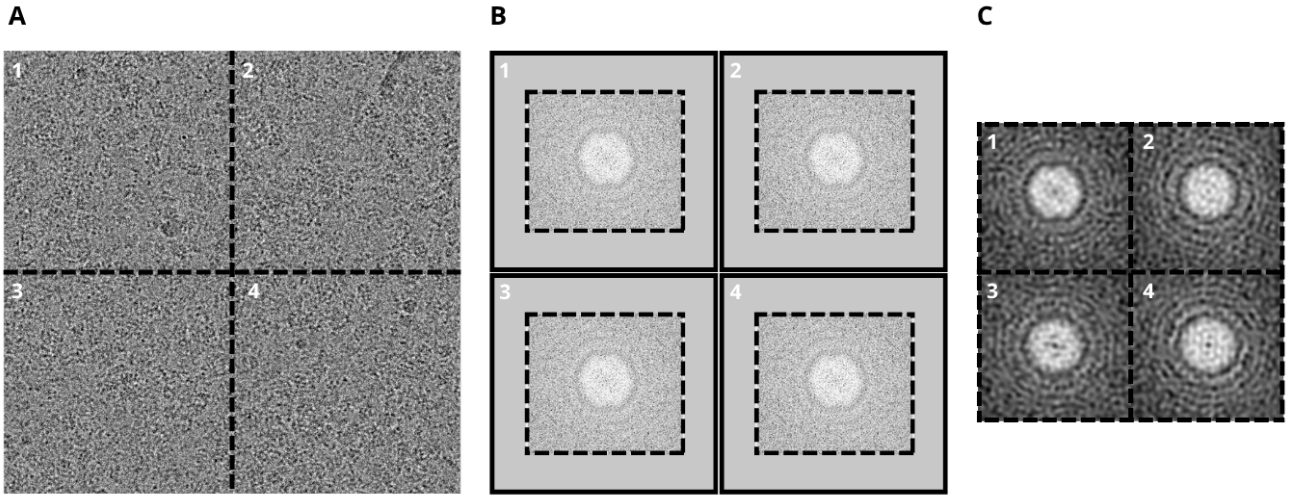


Figure 21. Image processing used for the PSD analysis. (A) Micrographs are divided into four corner quadrants. (B) Power Spectral Density (PSD) computation, cropped to a user-defined resolution limit. (C) Low-pass filtering applied to the PSD images. The micrograph is from EMPIAR-11051.

After this process, the algorithm performs two types of correlation analyses on the filtered quadrant PSDs to derive quality metrics:

1. **Cross-Quadrant Correlation:** The cross-correlation coefficient is calculated for all six unique pairs of the four quadrant PSDs. This measures the consistency of the Thon ring pattern across the entire micrograph. A high average correlation indicates a uniform CTF across the micrograph, characteristic of a flat, high-quality area.
2. **Rotational Autocorrelation:** Each quadrant's PSD is correlated with a 90-degree rotated version of itself. High correlation value indicates circular Thon rings, signifying low astigmatism. Conversely, a low correlation indicates elliptical rings, a hallmark of high astigmatism.

The full set of these ten correlation values (six cross-quadrant and four rotational) is collected. The **mean** and **standard deviation** of this set are then used as final quality scores (Figure 22).

- A low **mean** correlation suggests inconsistent or poor signal, which can arise from tilted specimens, thick ice, severe astigmatism, or empty micrographs (resulting in correlating noise against noise).
- A high **standard deviation** indicates regional variability, which is also a strong indicator of problems like specimen tilt, large contamination or variable ice thickness.

Based on empirical testing, default acceptance thresholds were set to a mean correlation > 0.35 and a standard deviation < 0.15. This program is integrated into the *Xmipp* software plugin as *micrographs psd analysis protocol*.

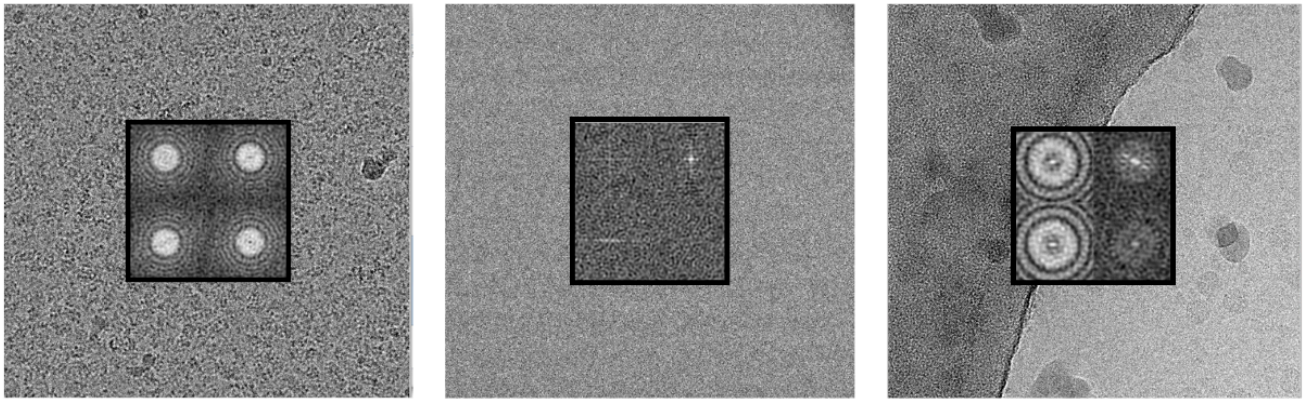


Figure 22. Examples of PSD Analysis Plots. The left micrograph shows an ideal image with a uniform signal across the field (high correlation: 0.8; low standard deviation: 0.02). The center micrograph shows a low-contrast, empty image (low correlation: 0.1; low standard deviation: 0.03). The right micrograph illustrates a two-texture case acquired at the edge of a hole, with half of the image over the foil (medium correlation: 0.56; high standard deviation: 0.2).

Miffi Categorize Micrographs Program

To leverage the power of deep learning for micrograph curation, the Miffi software package [35] was integrated into our workflow. Miffi follows a sophisticated hybrid approach, using a dual-branch convolutional neural network (CNN) that processes information from both real and Fourier space simultaneously to make a more informed classification. One branch of the network analyzes the real-space micrograph, learning to identify visual features such as particle distribution, aggregation, and large contaminants. The second branch analyzes the micrograph's Power Spectrum Density (PSD), allowing it to learn features related to the quality of the CTF, such as the presence and resolution of Thon rings, astigmatism, and drift. By combining the

outputs from these two branches, Miffi makes a holistic assessment that captures a wider range of quality indicators than methods relying on a single domain.

The network is trained to classify micrographs into one of several distinct categories, providing clear, actionable labels for data curation. These labels differentiate between high-quality data and various common defects (Figure 23):

- **Good:** High-quality micrographs suitable for further processing.
- **Bad Film:** Areas with support film artifacts.
- **Bad Drift:** Micrographs exhibiting excessive, uncorrected drift or cracked micrographs.
- **Bad Minor Crystalline Ice:** The presence of low ice crystallinity.
- **Bad Major Crystalline Ice:** Large, obstructive crystalline ice formations.
- **Bad Contamination:** Micrographs with other forms of contamination, such as ethane or ice.

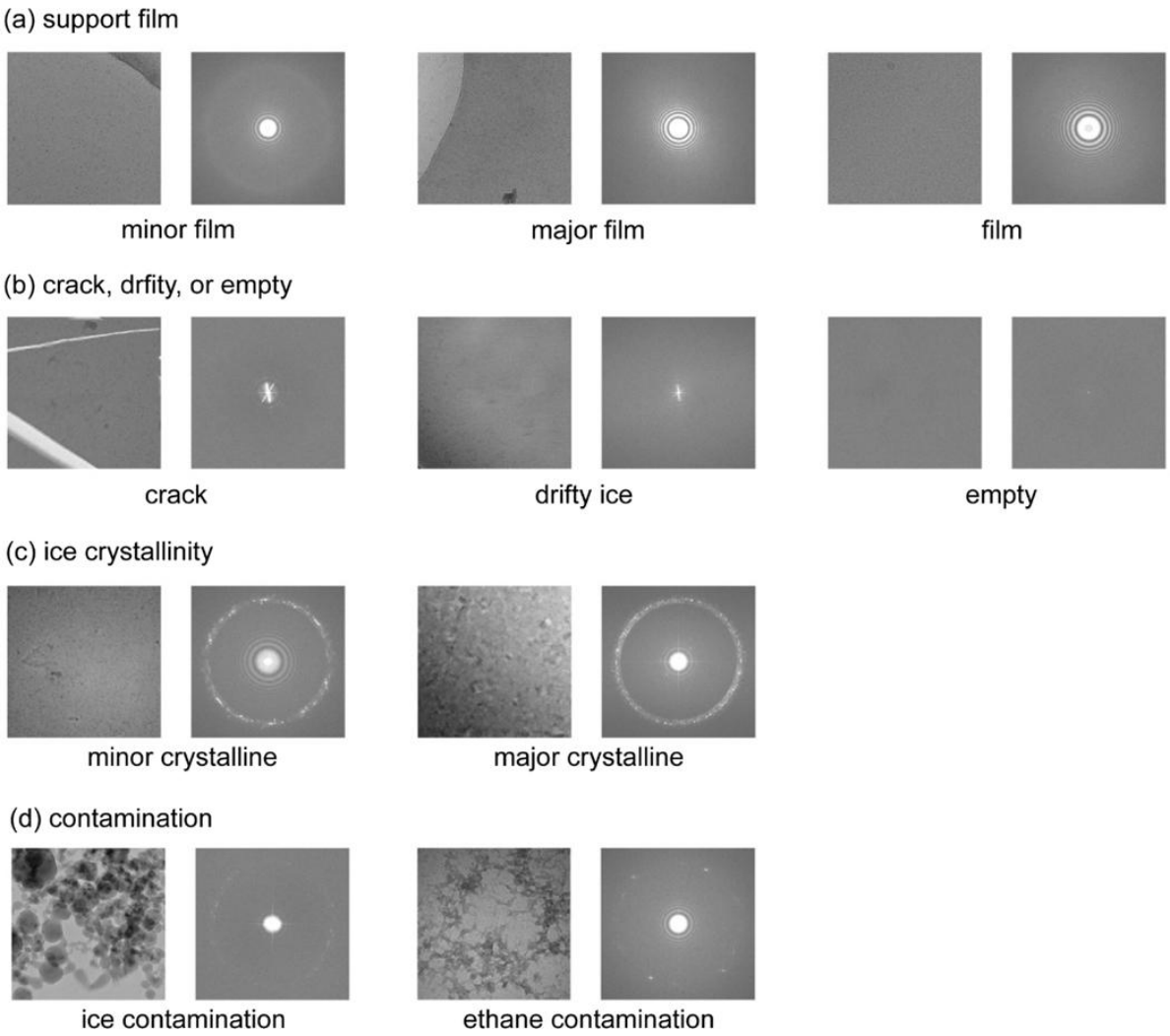


Figure 23. Examples of problematic micrographs for each label category in the miffi training set. Each real-space micrograph on the left is paired with its corresponding power spectrum on the right. (a) Support film: the support carbon film appears as a darker region in the real-space view, while the vitrified ice occupying the hole is noticeably lighter. (b) Sample displacement: micrographs affected by stage drift or cracking display characteristic signs such as asymmetric resolution or streak-like features in the Fourier domain. In contrast, empty holes typically produce a flat, featureless power spectrum dominated by strong intensity near the origin. (c) Ice crystallinity: crystalline ice is identified by a sharp ring at approximately 1/3.7 $Å^{-1}$ in the power spectrum. When crystallinity is pronounced, alternating bright–dark bands are visible directly in the micrograph (right); when subtle, the ring in Fourier space is the more reliable indicator (left). (d) Contamination: representative cases include ice crystals adhered to the hole (left) and small ethane contaminants embedded in the vitreous ice (right) [35].

This automated classification provides a level of consistency and accuracy that surpasses traditional CTF-based methods, achieving over 90% accuracy in identifying high-quality micrographs [35].

Miffi was integrated into the Scipion framework via a dedicated plugin. The plugin handles the download and installation of the Miffi software and its associated pre-trained models. The plugin provides a single protocol called *miffi - categorize micrographs*, which is specially designed for high-throughput use. The protocol can operate in streaming mode, process micrographs in user-defined batches to optimize resource usage, and take full advantage of available GPUs for acceleration. In this filter, we only stay with *Good* and *Minor crystalline ice* - labeled micrographs.

3.2.4 CTF Estimation Filter Protocol

As a final quality examination, a traditional filter is applied based on the parameters derived from CTF estimation. The CTF describes how aberrations in the microscope's objective lens, primarily defocus, modulate the phase and amplitude of the electron wave, which in turn affects image contrast at different spatial frequencies. Accurate estimation and correction of the CTF are fundamental to recovering high-resolution information, as uncorrected phase reversals can lead to signal cancellation during particle averaging.

For this step, our pipeline utilizes the robust and widely adopted *CTFFIND5* algorithm [92]. The core of this program is to determine the defocus values (defocus U, defocus V, and astigmatism angle) that best describe the observed micrograph. It achieves these objectives by dividing the micrograph into a series of smaller, overlapping tiles, computing the power spectrum for each, and comparing it to a library of theoretical CTF models to find the best fit. *CTFFIND5* improves upon previous versions with enhanced features for fitting data from tilted specimens, handling anisotropic magnification distortion, and more robustly scoring fits in the presence of crystalline ice, making it highly suitable for both SPA and cryo-electron tomography [92].

Once the CTF parameters are estimated, they are evaluated against a set of quality thresholds to decide whether the micrograph should be kept for further processing. The key parameters used for filtering are:

- **CTF Fit Resolution:** This metric quantifies the highest spatial frequency at which the experimental power spectrum shows a significant correlation (typically >0.5) with the fitted theoretical CTF model. This serves as a proxy for the quality of the high-resolution signal present in the micrograph. For instance, very high-resolution fits may indicate the presence of strong signal from a carbon support film, while unusually low-resolution fits can be indicative of crystalline ice, motion correction failure, or severe radiation damage (Figure 24). A good resolution limit is around < 5~6Å.

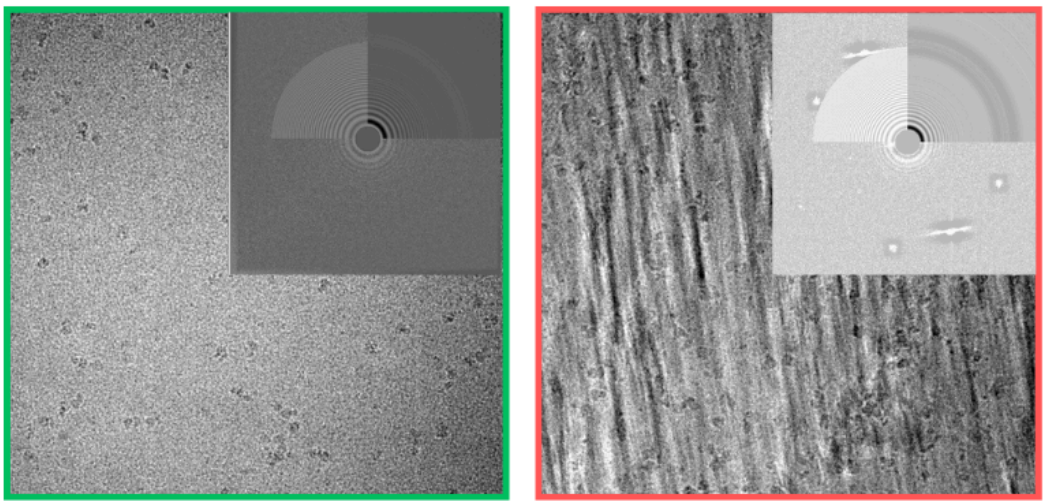


Figure 24. Examples of CTF estimation. Highlighted in green is an ideal CTF fit with a resolution of 4 Å, and in red a damaged micrograph with a resolution fit of 8.2 Å. Micrographs are from EMPIAR-10816.

- **Defocus Range**: This range ensures that the estimated mean defocus falls within a reasonable, user-defined range for the entire dataset (1000-40,000 Å). It is a critical parameter for maintaining dataset consistency. Micrographs with very high defocus values, while exhibiting strong contrast, may lack high-resolution information. Conversely, those with very low defocus values may have recoverable high-resolution detail but suffer from such low contrast that particles are difficult to detect or even undetectable. This relation can be observed in Figure 9.
- **Astigmatism Ratio:** Astigmatism, caused by lens asymmetry, results in different defocus values along two perpendicular axes, causing a directional blurring of the image (Figure 25). While often represented as the absolute difference between the two defocus values, this metric is not ideal for filtering as its significance depends on the magnitude of the overall defocus. A 0.1 μm difference is much more severe at a mean defocus of 0.5 μm than at 3.0 μm. To overcome this, we use a normalized **astigmatism ratio**:

$$Astigmatism\ Ratio\ = \frac{\left|defocus_U - defocus_V\right|}{mean(defocus_U, defocus_V)}$$

This metric provides a more reliable filter than the absolute difference between the two defocus values, as its significance is independent of the overall defocus. Micrographs with a ratio exceeding a user-defined threshold (*i.e.,* 0.1) are discarded.

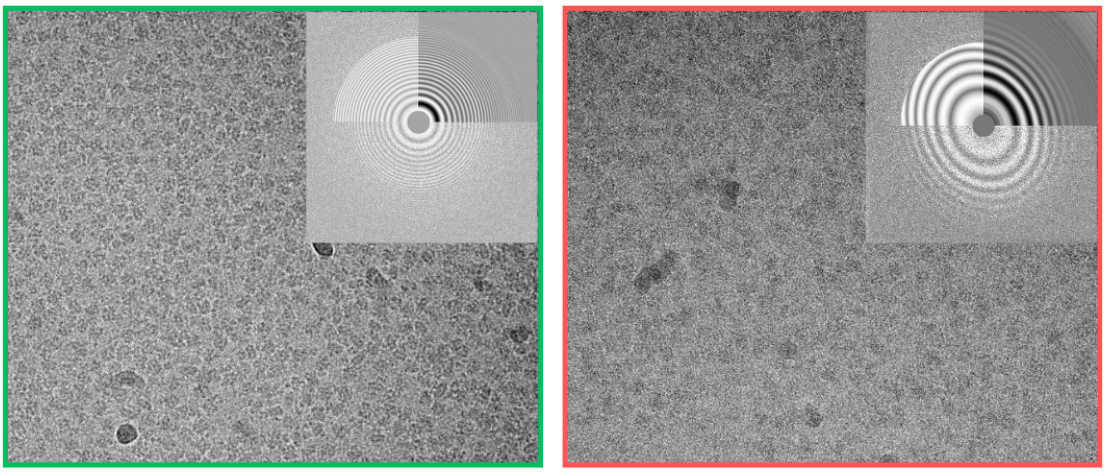


Figure 25. **Examples of CTF estimation and astigmatism.** Highlighted in green is an ideal CTF with an astigmatism ratio of 0.004, and in red an astigmatic micrograph with an astigmatism ratio of 0.32, displaying the characteristic oval Thon ring pattern typical of astigmatic micrographs. Micrographs are from EMPIAR-11057.

This program is integrated into the *Xmipp* software plugin as the *CTF consensus protocol*. Although used in this pipeline as a CTF filter, it can also serve to compare different CTF estimations and calculate a consensus resolution. The algorithm assumes that two CTFs are consistent if the phase (wave aberration function) of both remains within 90 degrees of each other. The reported consensus resolution corresponds to that at which the phase difference between the two CTFs reaches 90 degrees.

3.2.5 Quality monitor (Dashboard): Centralized visualization tool for high-throughput cryoEM facilities

During the international stay at the ESRF (The European Synchrotron) Cryo-EM Facility, a key limitation was identified: although *Scipion* provided access to all image-processing information through its graphical interface, the absence of a centralized, user-friendly visualization tool for monitoring and integrating relevant processing data across high-throughput acquisitions became evident.

While Scipion internally stores all metadata from the processing pipeline, its visualization can be cumbersome due to the complexity of workflows that often involve more than thirty distinct processing protocols, each with its own specialized viewers for displaying quality metrics. During active data collection, manually opening each quality filter to inspect its outputs is both inefficient and impractical.

To address this limitation, a **two-component solution** was designed and implemented:

1. **Quality metrics protocol**: A new *Scipion* protocol (*quality metrics protocol*) was developed to connect with key stages of the processing pipeline, particularly those involved in data curation. This protocol extracts relevant metadata in real time and stores it in a structured .csv file, allowing continuous monitoring and analysis of data collection parameters.
2. **Interactive Web Dashboard**: Using *Streamlit*, a Python framework for building interactive web applications, an online dashboard was developed to visualize the collected information in real time. The dashboard integrates quality control data from *Scipion* filters, allowing facility operators to monitor acquisition progress and assess data quality through a clear, centralized, and easily accessible web interface.

Quality metrics protocol

As introduced in Section 3.2 (Data Curation and Quality Control), a series of processing steps within the *Scipion* workflow are used to monitor and assess image quality at multiple stages, including movies, micrographs, and CTF estimations. These steps generate quality scores and statistics used as filters to identify and discard suboptimal data. The filters employed in this workflow include: **Movie Dose Analysis, Movie MaxShift, Micrographs PSD Analysis, Miffi Categorize Micrographs**, and **CTF Consensus**.

As shown in Figure 26, these protocols are connected to the ***Quality Metrics Protocol*** through the *Input Protocols* field. The protocol retrieves all metadata associated with each filter and operates in a streaming mode. At every *Sampling Interval*, it monitors the status of the input protocols and continues until all have reached either *Finished* or *Failed* status, indicating that no further updates are required.

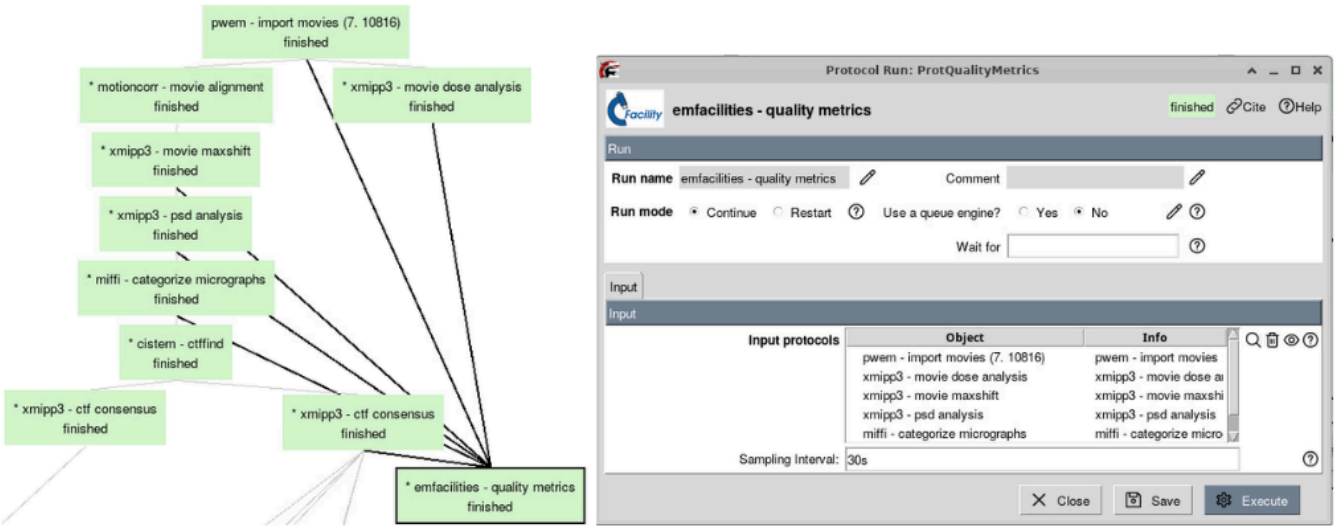


Figure 26. Overview of the Quality Metrics Protocol. On the left, the *Scipion* workflow displays the Data Curation Stage with the *Quality Metrics Protocol* attached to the quality filters. On the right, the protocol's GUI form shows the configurable parameters for execution (Input Protocols and Sampling Interval).

This custom-developed data collector is integrated via the *emfacilities Scipion* plugin and leverages its dedicated environment to load all necessary Python modules, including those required to launch the Streamlit dashboard.

Interactive Web-Dashboard

The dashboard, named ***Live Quality Metrics Monitor***, provides an overview of ongoing microscope data collection, offering real-time visualization of metadata and quality control metrics. It consists of three main sections: **Main View**, **Filters Views**, and **Scores View**.

Built using ***Streamlit***, this technology was chosen for its ability to transform Python scripts into interactive web applications with minimal overhead, automatic UI generation, and native support for dynamic data refresh, features that make it ideal for facility environments where accessibility and responsiveness are critical.

Main View

The **Main View** is the landing page of the dashboard, launched directly from the *Scipion* GUI via the *"Analyze Results"* button in the Quality Metrics Protocol.

As shown in Figure 27, the left panel (blue) provides interactive controls such as:

- **Object Views Selector**: Switch between Main, Filters, or Scores Views.
- **Reload Button**: Manual data refresh, although the dashboard auto-refreshes every 60 seconds.
- **Entries Range Filter**: Two-sided slider to limit displayed entries, defaulting to the last 100 micrographs.

The central (white) area includes four main panels (Figure 27):

- **Project Information Panel**: Displays the Scipion project name, start time, total processing duration, last update timestamp, and overall status (running or finished).
- **Acquisition Information Panel**: Summarizes acquisition parameters such as pixel size (Å/px), voltage (kV), magnification, spherical aberration (Cs), and dose per frame ($e^-/Å^2$).
- **Filters Statistics Panel**: Shows the number and percentage of accepted versus total images per filter. Percentages below 60% appear in red, signaling potential acquisition quality issues.
- **General Table Panel**: Presents all collected metrics per image in an interactive table. Rows corresponding to accepted micrographs are highlighted in green, while rejected ones appear in red. Columns can be dynamically sorted by any field.

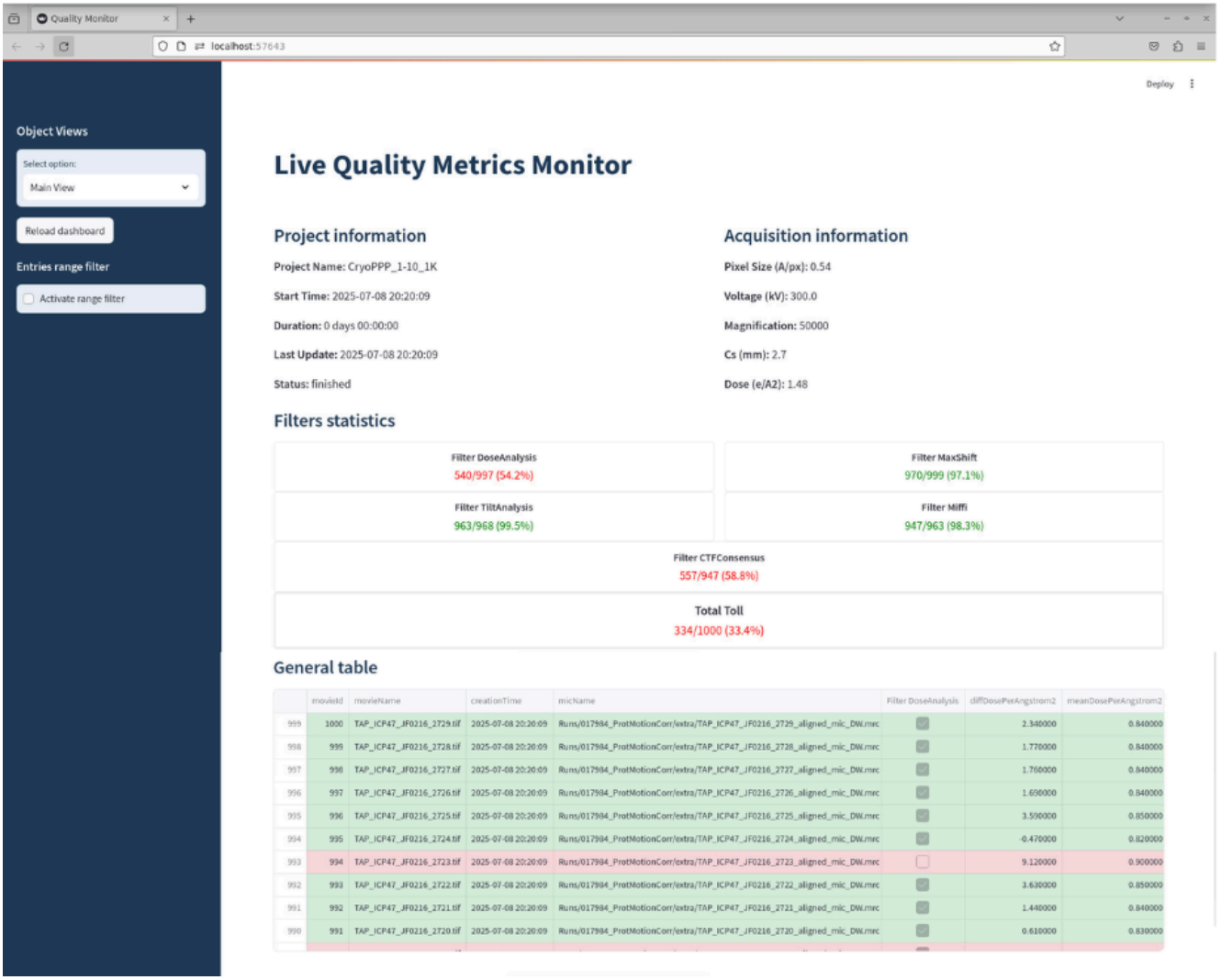


Figure 27. Main View of the Live Quality Metrics Monitor.

The **Object Views** and **Entries Range Filter** features (Figure 28) are shared across all views. The range filter is particularly useful during acquisition, allowing operators to focus on the most recent data to monitor quality evolution over time. By default, when activating the range filter it will display and calculate the statistics for the last 100 images.

Figure 28. Interactive panels of the Main View. The *Object Views* and *Entries Range Filter* features are highlighted on the left with red arrows. The central white panel dynamically updates according to the user's interaction with these two features, displaying only the data or visual elements corresponding to the selected filter and entry range.

Filters View

The **Filters Views** include **Dose**, **Drift**, **Tilt**, **Miffi**, and **CTF**, each corresponding to a specific quality control step in the image processing pipeline. These views share a common base structure, ensuring consistency and extendability, while allowing individual customization for each filter.

In addition to the common features, the Filters View layout includes (Figure 29):

- **Interactive Plots Panel:** Enables selection of variables for plotting through a dropdown menu. This dropdown menu shows all the quality metrics offered by each Filter. For each variable, a two-sided slider defines upper and lower limits, a very useful feature to exclude outliers that affect the representation. All interactions here dynamically update the remaining panels.
- **Filters Distribution Panel**: Displays the proportion of accepted versus rejected images, with percentages and counts available via hover interactions.
- **Filter Statistics & Thresholds Panels**: Summarize acceptance ratios (highlighting in red those below 60%) and list the threshold variables and their limit values extracted from *Scipion* metadata. Users can toggle between viewing all, accepted, or rejected images. This last interaction dynamically updates the rest of the panels.

- **Micrographs Viewer Panel:** Displays the selected micrograph corresponding to a point or entry in the plots and table, enabling users to directly inspect the associated image. Since cryo-EM raw micrographs have inherently low contrast, making structural features barely visible without further processing, a lightweight preprocessing function was implemented to enhance contrast prior to visualization. This function performs a simple contrast stretching operation based on the 2nd and 98th intensity percentiles of the image histogram, effectively improving the visual clarity of the micrographs while preserving their original intensity distribution. This preprocessing step ensures that micrographs are displayed with sufficient contrast to facilitate real-time visual inspection within the dashboard environment.
- **Dataframe Panel:** Shows an interactive table listing micrographs and their corresponding filters parameters, sorted by *movieId*. This is an interactive table, you can order by any of the fields. Metrics used for thresholding are highlighted (green for accepted, red for rejected). Selecting a row (two clicks) triggers image visualization in the *Micrograph Viewer panel* and a green message tells which micrograph was selected to be displayed.

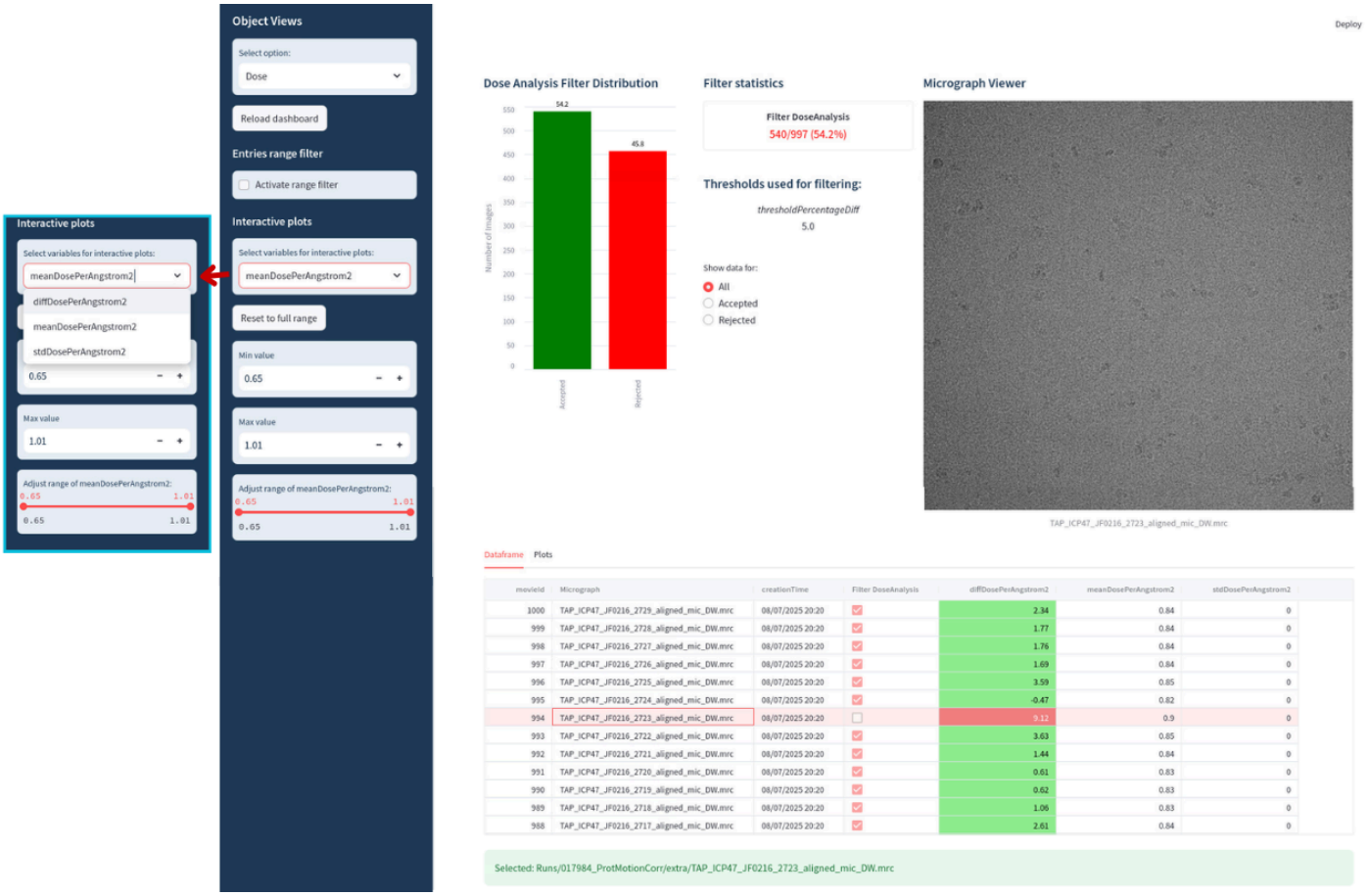


Figure 29. Example of Filters View (Dose). On the left, indicated by a red arrow, is the dropdown menu containing all the available quality metrics for this specific filter. This selection affects the plots displayed in the *Plots Panel*.

- **Plots Panel** (Figure 30): Contains two interactive visualizations that can be selected from a dropdown menu: a scatter plot (responsive to range filters and clickable for micrograph display) and a histogram (with hover interactions showing image counts per bin with dynamic bin selection). For the scatter plot it is important to remark that only the quality metrics that were used as thresholds are the ones that will contain the thresholds in the plots and that will be marked in blue or red based on accepted or rejected data points.

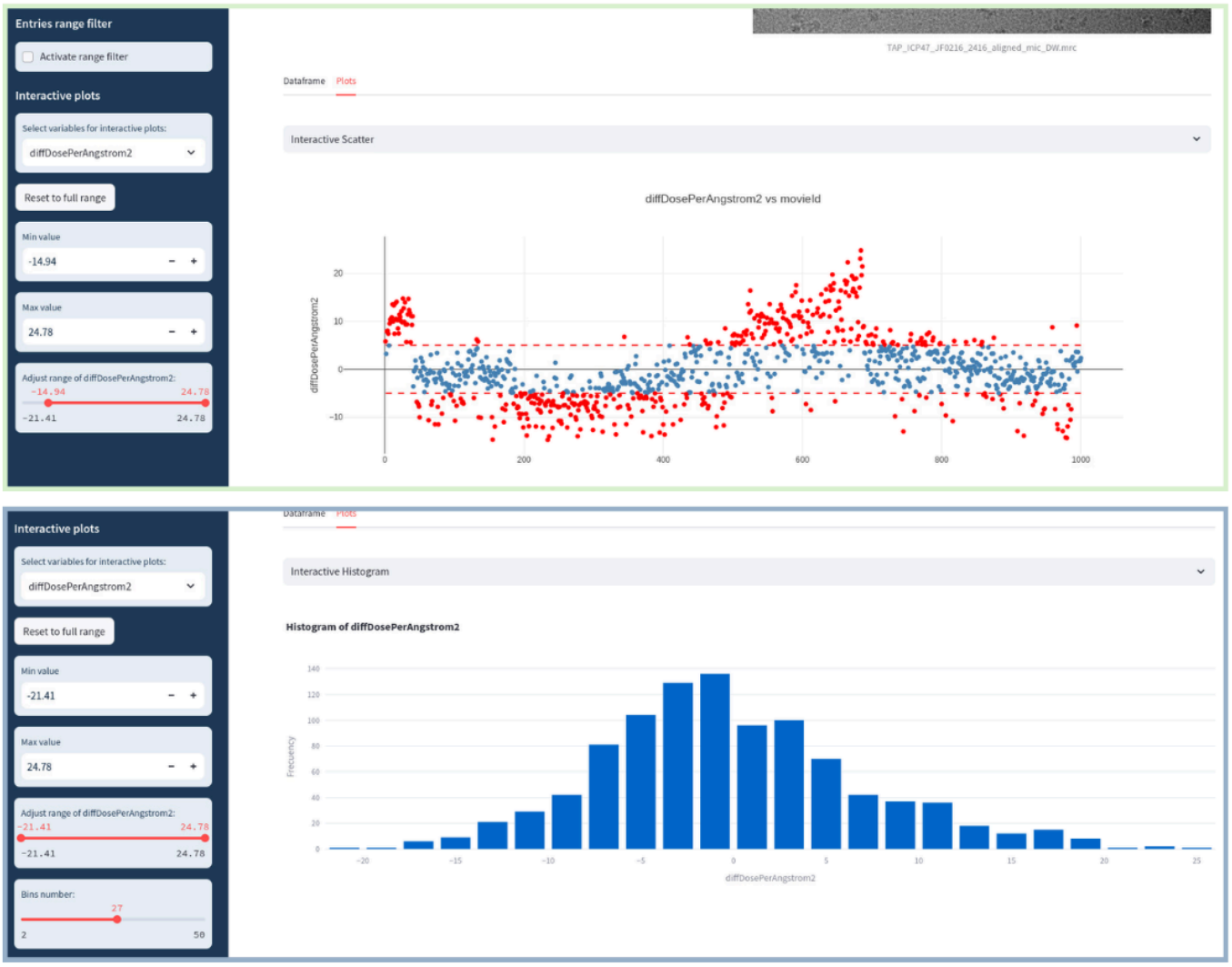


Figure 30. Plots Panel Options in the Filters View (Dose). The top panel shows the *scatter plot view*, while the bottom panel displays the *histogram view*. The control panel on the left provides a set of interactive buttons that allow users to customize and interact with the plots. These tools facilitate the exploration of quality metrics and the identification of outliers or trends in the dataset.

Scores View

The **Scores View** is designed for exploratory data analysis, enabling users to investigate relationships among quality metrics from different filters. In the current sequential filtering approach, only micrographs passing all filters remain for analysis. However, to fully exploit this feature, a **parallel filtering strategy** is proposed (Figure 31), ensuring all filters process all images simultaneously so that all the images contain all the quality metrics and all the relations can be explored.

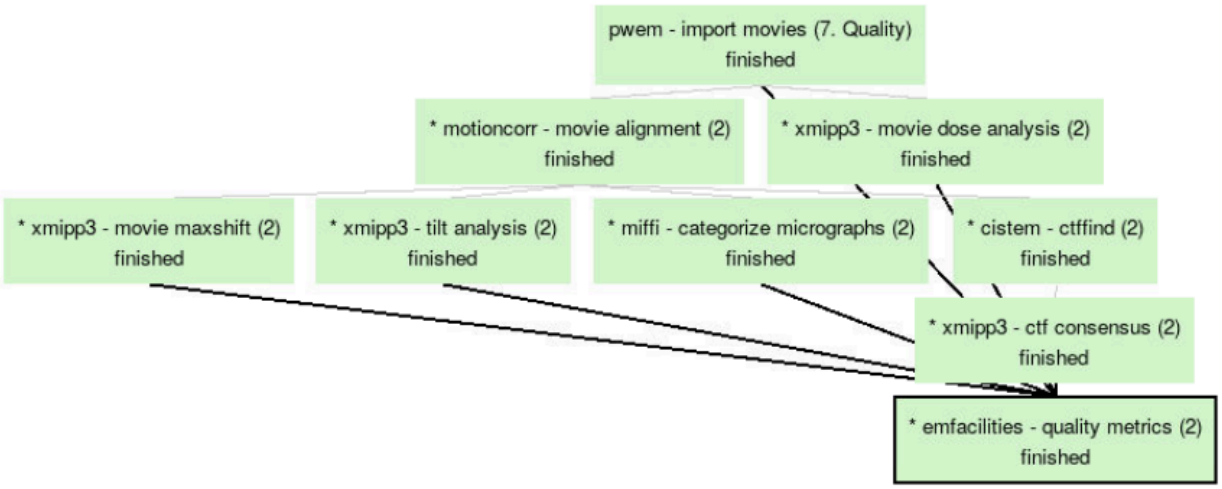


Figure 31. Parallel filtering strategy. Scipion workflow illustrating the Data Curation Stage in a parallel configuration with the quality metrics protocol attached to the quality filters.

The **Scores View** presents a distinct panel layout designed to facilitate the comparative analysis of quality filters. Among its components (Figure 32):

- The **Filter Discrepancy Matrix Panel**: Provides a quantitative overview of how consistently different filters agree or disagree in their classification of micrographs. Users can select from three metrics to compute this agreement:

    A. *Discrepancy:* measures the percentage of samples for which two filters produce different binary outcomes (accepted or rejected), providing an intuitive indication of disagreement.
    B. *Jaccard:* quantifies the ratio between the intersection and the union of two filters' positive classifications, expressing their degree of overlap as a percentage.
    C. *Correlation:* calculates the Pearson correlation coefficient between two binary filter arrays, reflecting the linear relationship between their respective outputs.

    This implementation produces a symmetric matrix that can be visualized as a heatmap within the dashboard, providing an intuitive view of filter agreement patterns and helping users assess the coherence and complementarity of their quality filters.

- **Variable Correlation Matrix Panel**: Provides an overview of the relationships between all available quantitative quality metrics across the different filters. By computing pairwise correlations, this panel helps identify metrics that are highly interdependent or redundant, as well as those providing complementary information. Such analysis is particularly useful for understanding whether multiple quality criteria

are capturing similar image characteristics or evaluating distinct aspects of micrograph quality. To generate this matrix, it filters out non-numeric and identifier columns (e.g., *movieId*, *micName*, *creationTime*) and converts categorical variables, such as *miffiLabel*, into binary encodings when necessary. The remaining numeric data are then used to compute a Pearson correlation matrix, which can be visualized as a heatmap within the dashboard to highlight strong positive or negative relationships among metrics. The resulting correlation matrix serves as a valuable diagnostic tool for evaluating the behavior of different quality metrics and guiding the design of more balanced filtering strategies within the automated processing workflow.

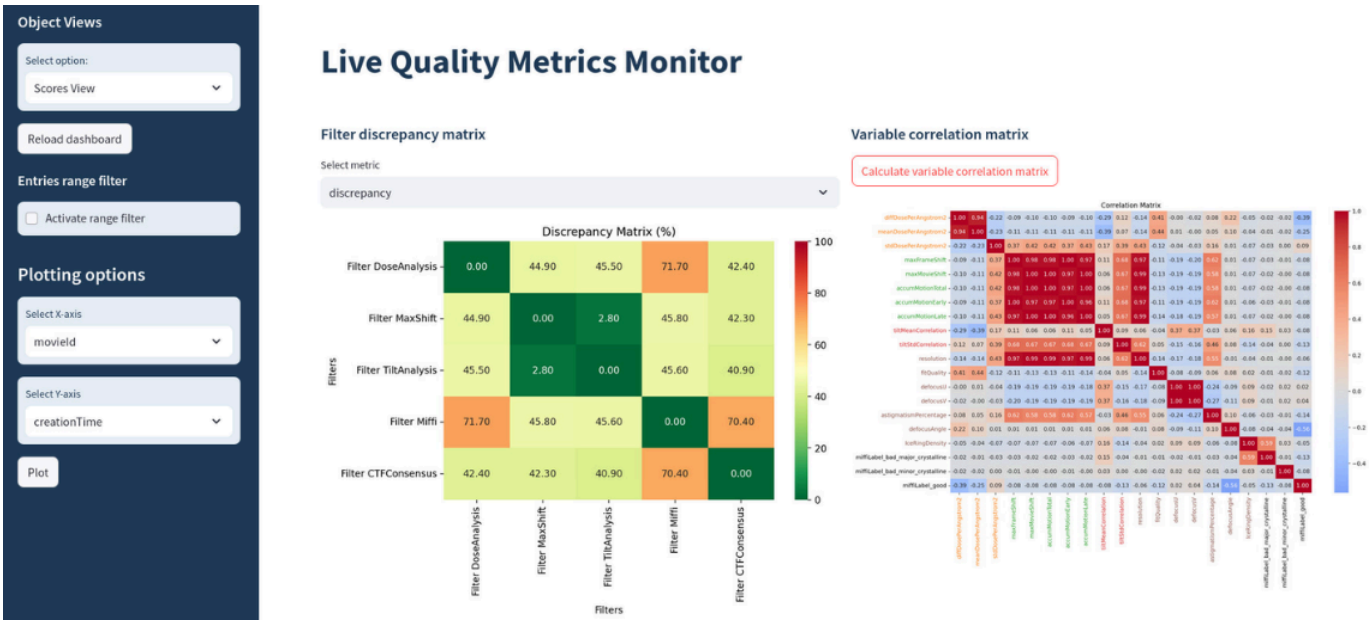


Figure 32. Example of the Scores View. The main white panel displays the two components described above: the *Filter Discrepancy Matrix* and the *Variable Correlation Matrix*.

- **Plotting Options Panel and Correlation Tab Panel:** Enables a deeper exploration of the relationships identified in the **Correlation Matrix**. Here, users can interactively select any pair of quantitative quality metrics (*Plotting Options Panel*) to visualize their correlation and examine potential trends or dependencies (*Correlation Tab Panel*). This visual inspection allows for a more intuitive understanding of how different acquisition and processing parameters influence one another, providing valuable insight into the behavior and quality of the collected data. For example, in Figure 33, the variable *meanDosePerAngstrom²* is plotted against the *CTF fit resolution*, revealing a strong negative correlation. This means that micrographs acquired with higher dose per frame tend to exhibit better (i.e., higher) CTF-estimated resolution. This observation is consistent with the physical principles of cryo-EM imaging, where micrographs with

thicker ice, typically associated with lower dose per frame, tend to have more noise and thus poorer resolution estimates.

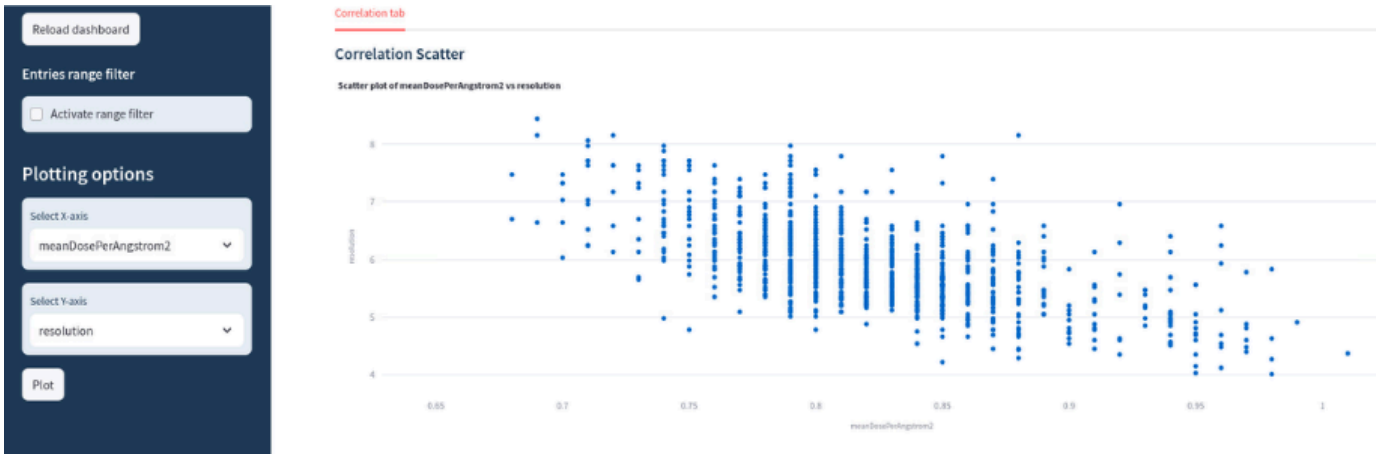


Figure 33. Plotting Options Panel and Correlation Tab Panel Example. The plots display the *meanDosePerAngstrom² (movie)* versus the *resolution (CTF)*. These two metrics represent opposite extremes of the data curation pipeline, from raw movie quality metrics to CTF quality assessment.

This example demonstrates how combining metrics from distinct stages, raw movie metadata (dose) and CTF estimation (resolution), can provide a more comprehensive interpretation of sample and data quality. By integrating these distinct sources of information, the dashboard supports a more informed evaluation of experimental conditions and helps identify underlying causes of quality variation across micrographs.

Main characteristics of the Dashboard

The Dashboard most important features are the following:

A. **Accessibility**: At present, the dashboard runs locally on the workstation where the quality monitoring protocol is executed. However, its design allows for straightforward deployment on a small internal server within the facility. In such a configuration, the dashboard could be made accessible through a secure URL provided by the facility staff. This would enable both external users, who typically lack access to the internal network, and facility members working remotely to conveniently monitor data collection quality.

This could be achieved by hosting the dashboard on an internal small server that acts as a gateway between the microscope network and external users. A reverse proxy (e.g., Nginx or Caddy) could manage encrypted HTTPS connections and user authentication, ensuring that only authorized visitors can access the data. For temporary access, a session-based or tokenized URL could be generated automatically at the start of an acquisition, providing time-limited, read-only access to the dashboard. Alternatively,

a tunneling service such as Cloudflare Tunnel or ngrok could be employed to create a short-lived public endpoint without modifying institutional firewall settings. These solutions would allow users to securely visualize acquisition progress in real time from any web-enabled device, while maintaining complete isolation from the microscope and image processing control systems.

Owing to its web-based architecture, the dashboard is inherently responsive and platform-independent, allowing users to monitor the acquisition seamlessly from a computer, tablet, or smartphone.

B. Partial Interactivity: To ensure data integrity, the dashboard is visualization-only and cannot modify processing results. Nonetheless, it offers full interactivity for exploration, clickable entries, adjustable filters, and dynamic plots.

C. Exploratory Analysis: Our dashboard not only facilitates interactive visualization of these quality parameters but also enables exploration of their relationships through correlation and discrepancy matrices. This metadata analysis offers insight into how different quality filters behave and interact across the pipeline, revealing patterns and potential redundancies.

D. Model-Oriented Data Structure: The structured data collection design lays the foundation to developing machine learning models by providing a structured approach to metadata retrieval and exploration, a critical step before model building and training.

E. Extendability: Both the *Quality Metrics Protocol* and the *Dashboard* are modular and easily extendable. New quality filters can be integrated by defining the metadata to extract in the protocol and adding a corresponding Filter View, either reusing the existing skeleton or designing a custom layout.

This development emerged as a natural evolution of *Scipion's* processing capabilities, addressing the need for a centralized visualization tool in high-throughput Cryo-EM environments. The centralized implementation enhances data traceability, accessibility, and interpretability, while providing a strong foundation for future extensions focused on automated decision support and advanced data visualization.

3.3 Automated Particle Picking Strategy

Once micrographs have been curated and high-quality images are obtained, the next stage of the pipeline is particle picking. The objective of this critical step is to accurately locate the coordinates of every individual protein particle in each micrograph while simultaneously avoiding false positives such as crystalline ice contamination, carbon edges, malformed or aggregated particles, and background noise. The quality of the final particle set is paramount, as

it directly impacts all subsequent processing steps, including 2D classification, 3D reconstruction, and refinement. The presence of false positive particles can complicate downstream classification and, in severe cases, prevent the 3D reconstruction process from converging entirely.

3.3.1 Overview of existing pickers and limitations

The picking task is exceptionally challenging due to several intrinsic properties of cryo-EM data. The extremely low electron dose used to prevent radiation damage results in images with a very low signal-to-noise ratio (SNR). This, combined with the low inherent contrast of biological macromolecules embedded in vitreous ice, makes particles difficult to distinguish from the noisy background, even for a trained human expert. The problem is further compounded by the presence of contaminants, variations in ice thickness and defocus, and the unpredictable appearance of particles due to their different orientations [3].

To address this challenge, numerous automated and semi-automated particle-picking algorithms have been developed. A common traditional technique is **template matching**, which uses user-defined particle images as templates to locate similar patterns in micrographs. While effective in some cases, this approach requires careful manual adjustment of parameters and often selects invalid particles (false positives), necessitating a subsequent, labor-intensive manual curation step.

To overcome these limitations, a new generation of programs based on artificial intelligence (AI) and machine learning (ML) has emerged. Tools such as *Xmipp* [62], *APPLE picker* [93], *DeepPicker* [64], *crYOLO* [33], *WARP* [30], and *Topaz* [34] have significantly enhanced particle localization efficiency and accuracy. However, a major challenge for these methods is **generalization**. Models are often trained on limited datasets of a few standard proteins (e.g., Apoferritin) and perform well on similar data but struggle to generalize to new, unseen proteins, especially those with irregular or complex shapes. They often fail to distinguish effectively between 'good' and 'bad' particles, such as those in aggregates, near ice contamination, or in carbon-rich areas [94].

This generalization gap highlights a key bottleneck that has historically hindered the development of truly automated AI-based pickers: the lack of large, high-quality, and diverse manually labeled datasets for training. To address this issue, community efforts have been undertaken to create large-scale, expertly curated datasets, such as CryoPPP and CryoCrab [3], [52], which are expected to facilitate the development of the next generation of ML-based methods.

A crucial lesson learned from the creation of these large datasets is the importance of accounting for the significant micrograph diversity within a single protein dataset. A key variable is the defocus level, which dramatically alters particle appearance and image contrast. As shown in Figure 34, particles in low-defocus micrographs are rich in high-resolution detail but have very low contrast, making them challenging to detect, whereas particles in high-defocus micrographs have strong contrast but lack high-frequency information. Most particle picking methods do not explicitly account for this variance, making it difficult to define a single set of parameters or a single model that performs well across the entire defocus range of a typical dataset. This insight illustrates the importance of picking strategies that can adapt to the unique properties of each dataset without requiring manual intervention.

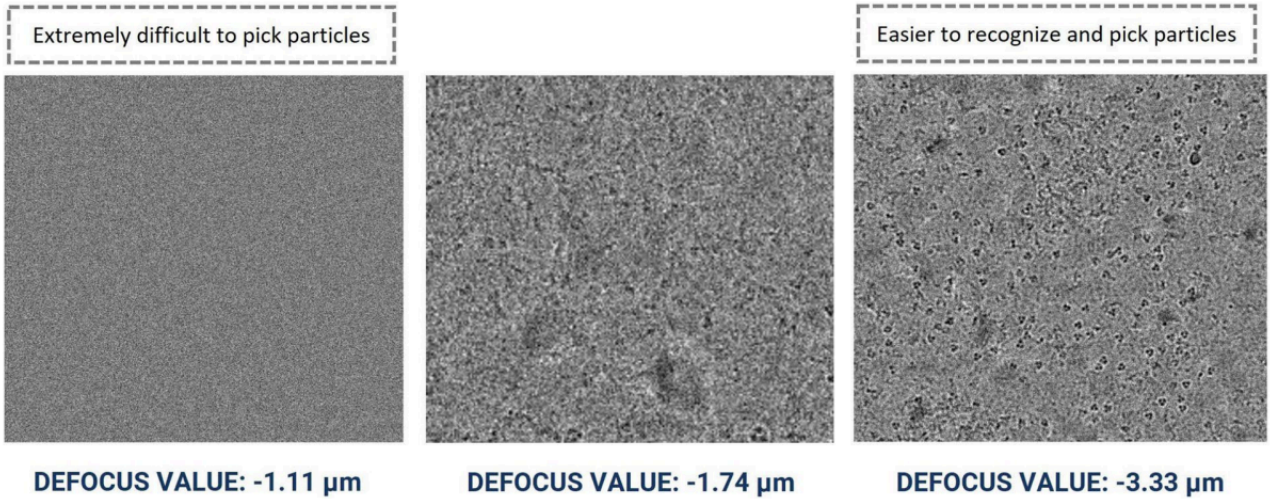


Figure 34. ***"Cryo-EM micrograph images of EMPIAR ID 10532 (Influenza Hemagglutinin) with different defocus values****. Micrographs with smaller defocus values make particle picking difficult and vice-versa"* [3].

3.3.2 Image processing strategy for Automated Model Training

A central innovation of this workflow is the robust, on-the-fly training of a data-specific particle picking model. While conventional pickers can be used "out-of-the-box," they often lack specificity. The traditional alternative, manual picking to create a training set, is a bottleneck for novice and expert users alike. Our principal goal is to generate a well-suited picking model automatically, without prior knowledge of the protein, that adapts to the unique characteristics (size, shape, contrast, etc.) of the sample during data acquisition. This process is divided into three stages (Figure 35):

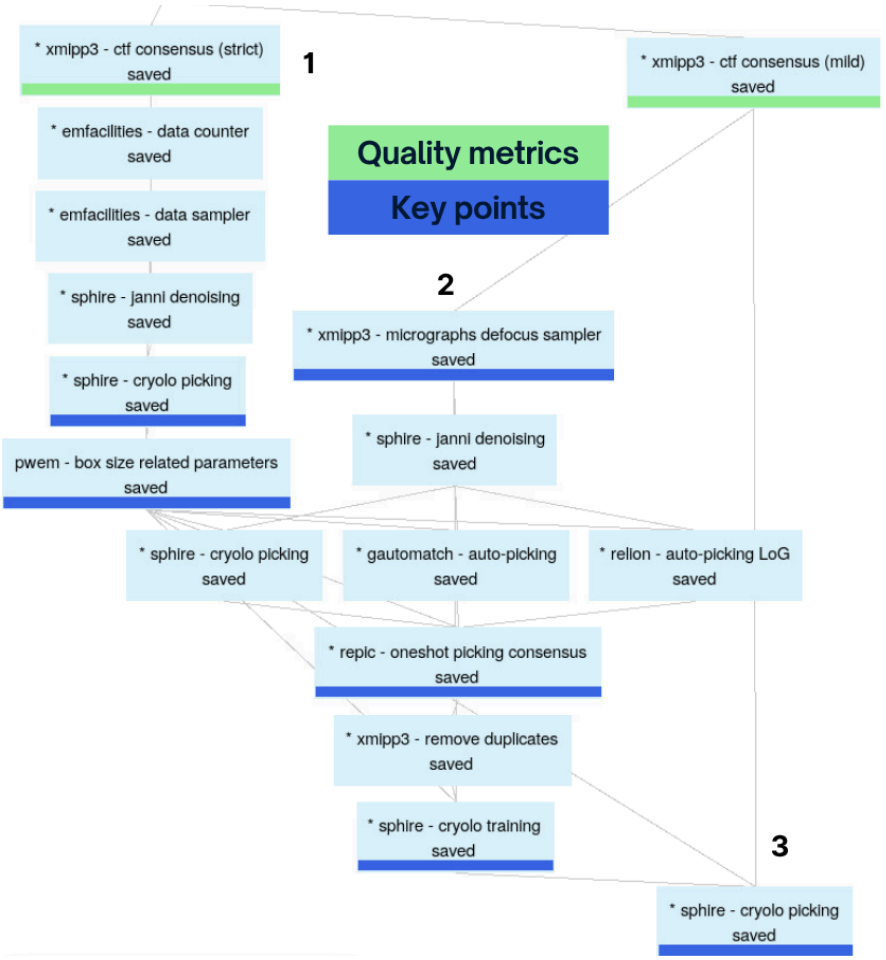


Figure 35. A detailed Scipion workflow diagram for the Automated Particle Picking strategy. Green labels indicate the final protocols of the curation steps, showing the sequential connection between main stages. Dark blue labels indicate the key protocols for the automated picking strategies, which are explained in detail later in this section. Numbers refer to the picking strategy stage.

Stage 1: Automatic Particle Size Estimation

The particle diameter is the most important prerequisite for any particle picker, as it determines the size of the object the algorithm will look for in the micrographs. Since this information is often unknown for a novel protein, our workflow needs to automate its estimation. We leverage the internal box size estimator from the *crYOLO* program, which has proven effective for this task [33]. However, the accuracy of this estimation is highly dependent on the quality of the input micrographs. Presenting empty, contaminated, or low-contrast images can lead to an incorrect estimation, which would cause the entire subsequent picking process to fail. To mitigate this risk, our strategy employs, apart from the multi-level filtering curation methods described previously, the following steps (Figure 36):

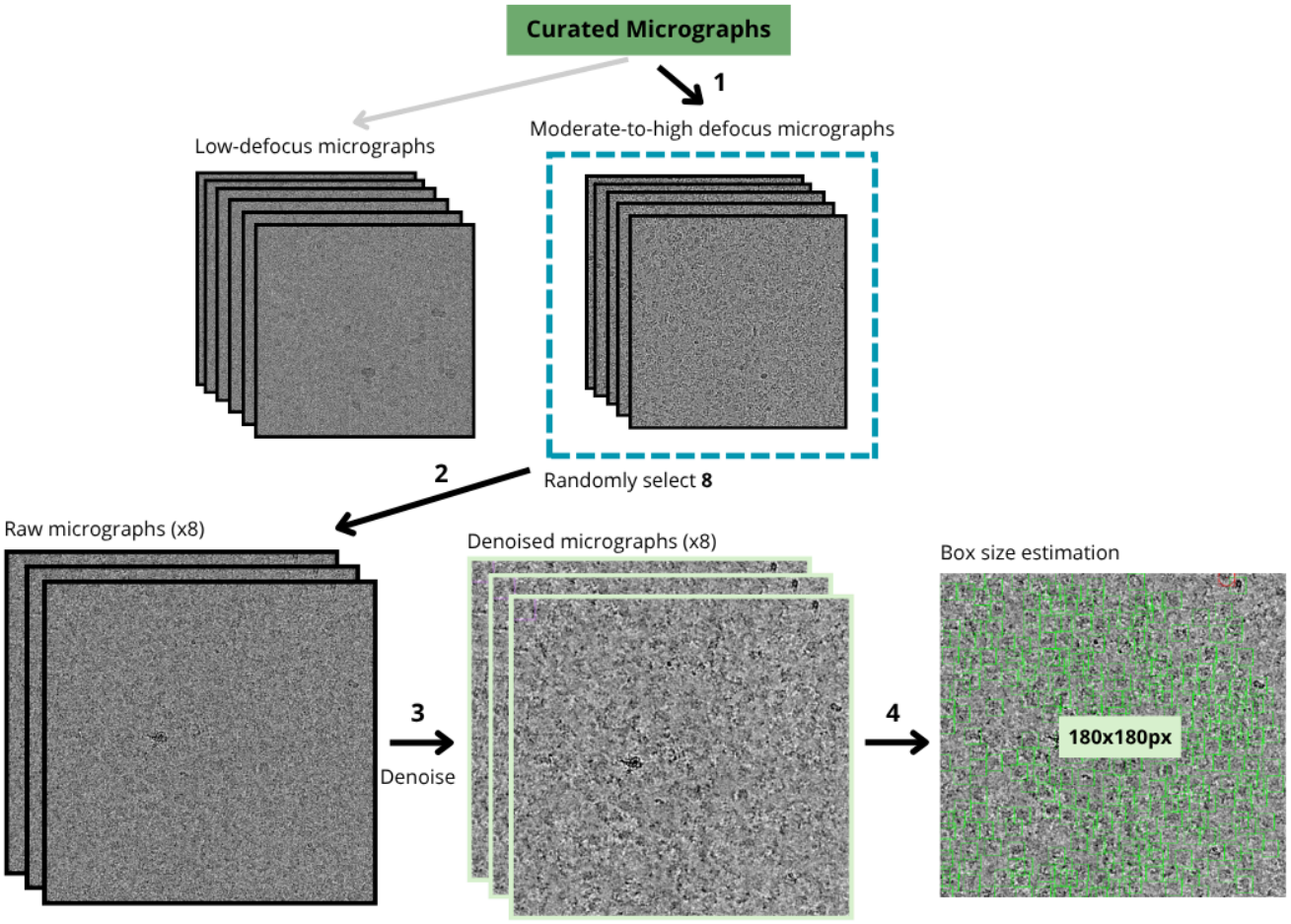


Figure 36. The schematic for automatic particle size estimation. Each number corresponds to a processing step involved in this operation, which is described in detail below. Micrographs are from EMPIAR-11051.

1. **Extract high quality and high contrast micrographs:** an additional stringent CTF filter that selects for higher resolution fits (resolution limit of 5Å) and moderate-to-high defocus values (10,000-30,000 Å), ensuring the presence of high-contrast particles.
2. **Random sampling of micrographs:** From the first 100 micrographs that pass this stringent CTF resolution filter, a random subset of eight is selected using custom-developed *data counter* and *data sampler protocols*. This random sampling ensures that the chosen micrographs are representative of the initial dataset, capturing a variety of particle views and local environments (different grids location), which is crucial for a robust estimation.
3. **Denoising and improving contrast:** To further enhance particle visibility, these eight curated micrographs are then denoised using *JANNI*, a neural-network-based noise2noise implementation. It is used to denoise previously unseen micrographs and is especially effective for low-SNR data [95].

4. **Box size estimation:** The final set of high-quality, high-contrast, denoised images is provided to the internal box size estimator of the crYOLO program [33], which provides a reliable and accurate determination of this crucial parameter.

Stage 2: Training of a Data-Specific Particle Picking Model

With the particle size determined, the next step is to train a new, data-specific model. This process is designed to overcome two main challenges: preparing a micrograph set that accounts for the dataset's variability and building a robust set of ground-truth coordinates for training. These are the steps designed to address these challenges (Figure 37):

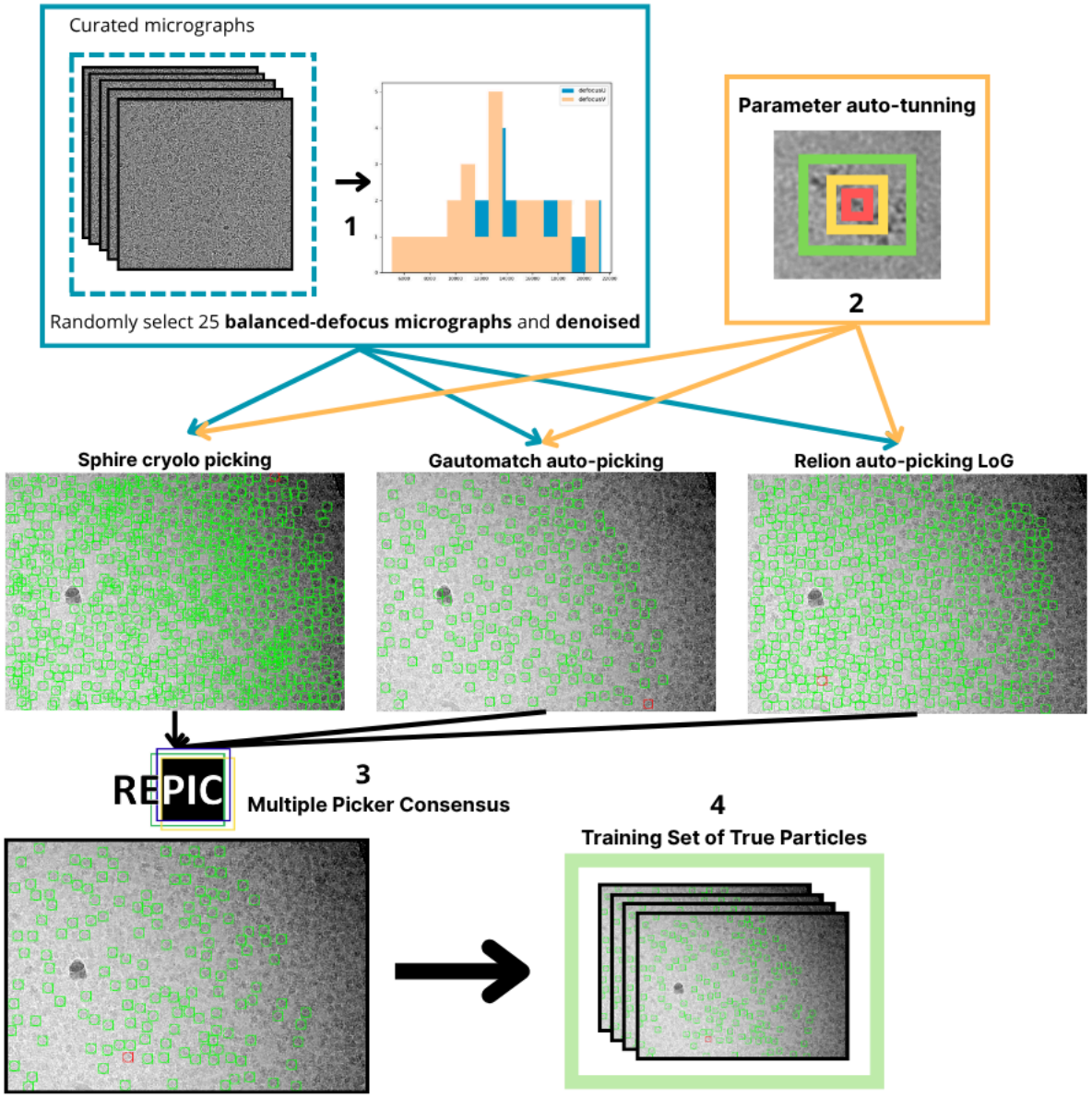


Figure 37. The schematic for training a data-specific particle picking model. Each number corresponds to a processing step involved in this operation, which is described in detail below. Micrographs are from EMPIAR-11057.

1. **Defocus-Balanced Micrograph Sub-sampling and Denoising:** To account for the diversity in particle appearance caused by varying defocus levels, the model is trained on a defocus-balanced subset of micrographs. A custom program, *micrographs defocus sampler protocol*, performs this balanced sampling based on the average defocus of each micrograph. The program waits until an initial set of 250 curated micrographs is available (configurable parameter), providing a representative sample of the defocus range, and then selects a balanced subset of 25 for training (configurable parameter). These 25 micrographs are then denoised using JANNI to improve the SNR and enhance the performance of the initial pickers.

2. **Parameter Auto-Tuning for Multiple Pickers:** The estimated box size is used to automatically configure and run three distinct picking algorithms (template-free picking algorithms), each employing a distinct mathematical approach:

    - **crYOLO:** a fast and accurate cryo-EM particle-picking program built on convolutional neural networks and leverages the widely You Only Look Once (YOLO) object-detection model [33].
    - **Gautomatch:** a GPU-accelerated program designed for rapid, precise, and flexible particle picking, supporting both template-free and template-based approaches in a fully automated manner. The template-free, correlation-based method is the one used [63].
    - **RELION's Laplacian-of-Gaussian (LoG) picker**: template-free auto-picking algorithm that employs a Laplacian-of-Gaussian (LoG) filter to detect particle-like features [37].

    A custom Scipion protocol, *pwem - box size related parameters*, generates the optimal settings for each picker based on the particle's diameter and pixel size. Particle pickers are useful but have practical limitations such as consistency and predictability. State-of-the-art particle-picking algorithms often produce inconsistent results as each method applies its own criteria for separating particles from background, determined by its specific model and/or training data. Consequently, it is difficult to predict in advance which picker will perform best on a challenging specimen. Using multiple pickers with different underlying algorithms to the 25 denoised micrographs generates a broader and more representative set of initial particle candidates.

3. **Multiple Picker Consensus for Ground-Truth Generation:** Since no single picker is optimal for all datasets, a consensus-based approach is used to generate a high-confidence set of ground-truth particles. The picks from the three algorithms are fed into the *REPIC* software [96], which frames the consensus problem as an integer

linear programming task. REPIC identifies particles common to the different pickers, producing a high-quality consensus set that is robust even if one of the pickers performs poorly. Reconstructions using consensus particles without particle filtering are shown to achieve resolutions comparable to those from particles picked by experts [96]. This automated process effectively replaces the need for manual picking to generate a training set.

4. **Training a Data-Specific Model:** The high-confidence particle coordinates from the *REPIC* consensus serve as the ground truth for training a new crYOLO model. The training is performed on the 25 original (non-denoised) micrographs, using the coordinates derived from the denoised images. This strategy leverages the improved picking performance on the denoised images without incurring the computational cost of denoising the entire dataset. We use transfer learning from the general crYOLO model, which accelerates training and leverages the general model's ability to avoid common artifacts. The resulting model is highly specialized for the target particle and the specific conditions of the dataset. We are using crYOLO training for the following reasons:

    - **crYOLO makes training easy:** to train a specialized model, no selection of negative examples and only a small number of micrographs are needed [33].
    - **crYOLO makes training tolerant:** crYOLO is designed to be robust to incomplete annotations, meaning that it can still learn effectively even when many particles are missing from the training set. While an ideal dataset would contain exhaustively picked particles for every micrograph, generating such annotations is often challenging and in densely populated images, extremely time-consuming. Fortunately, crYOLO does not require fully labeled micrographs; it can be trained successfully using sparsely annotated examples. Studies have shown that its performance on partially labeled datasets remains comparable to models trained on completely annotated micrographs [33].

Stage 3. Final Picking with the Data-Specific Model

Once the new model has been trained, it is applied to the entire set of curated micrographs from the data collection session. Because this model has been trained on the actual particles from the dataset, it exhibits significantly higher precision and recall than any general model. It is more effective at distinguishing true particles from noise and contaminants and is better adapted to the specific contrast and orientation distribution of the sample. This final picking step yields a large, high-quality set of particle coordinates that serves as the input for the subsequent stages of 3D reconstruction and refinement.

3.4 Initial 2D and 3D Analysis

Following the successful generation of a high-quality particle picking, the pipeline performs the third stage: initial 2D and 3D analysis (Figure 38). After particles are extracted, the workflow branches into two parallel processing tasks. The first performs consecutive 2D classification to produce time-updated class averages, providing a detailed view of particle views over time. Concurrently, the second generates *de novo* 3D models and uses them for a first 3D classification to separate true particles from noise and contaminants. This parallel design provides a rapid and comprehensive structural overview, which is critical for giving feedback during data collection.

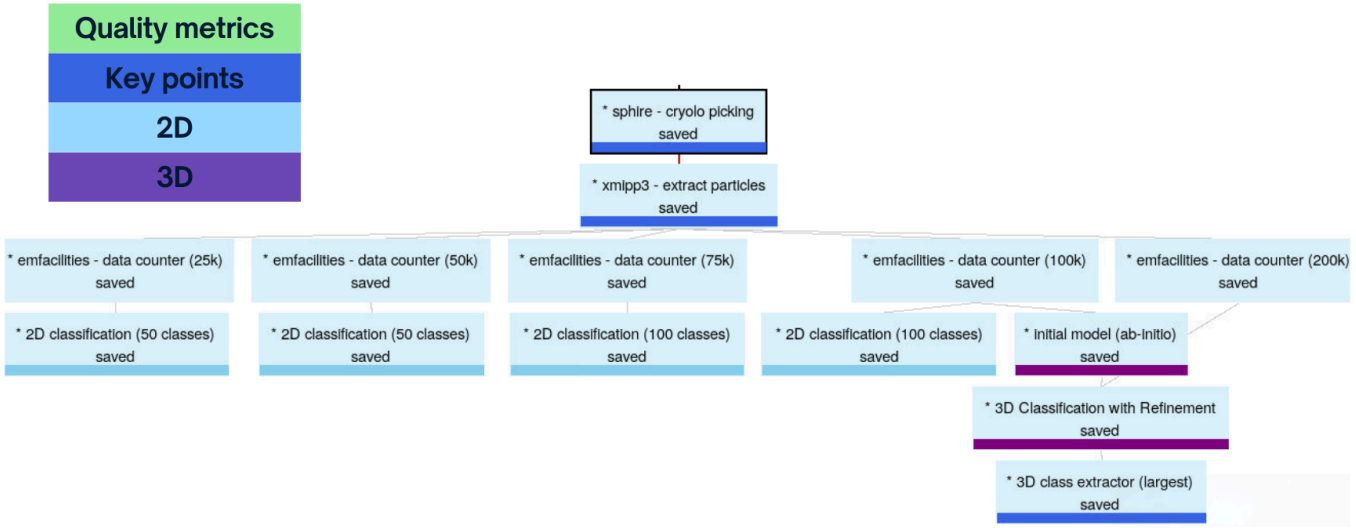


Figure 38. The detailed Scipion workflow diagram for the Initial 2D and 3D Analysis stage. Dark blue labels indicate the final steps of the *Particle Picking strategy* and the key decisions of this stage. Sky blue labels correspond to the image processing steps related to 2D analysis, while purple labels denote the steps pertaining to 3D analysis.

3.4.1 Overview of the importance of 2D and 3D Feedback

The ability to obtain rapid feedback on the structural characteristics of the sample while it is still in the microscope is a transformative advantage. This early analysis can reveal critical information about sample quality that is not apparent from inspecting the raw micrographs alone. At this stage, we can definitively identify issues such as sample heterogeneity, complex dissociation, interactions with the air-water interface, and, most notably, preferred particle orientation.

Preferred orientation, where particles adopt a limited set of views on the grid, is a particularly limiting problem that can prevent the reconstruction of a high-resolution 3D map. It is often not apparent until 2D classification, where the prevalence of overrepresented poses is revealed.

If left unaddressed, this biased orientation distribution translates into stretching artifacts and an anisotropic resolution in the final 3D map. By employing 2D and 3D on-the-fly processing, preferred orientation can be identified early in the data collection process, allowing the microscope operator to implement alternative data collection schemes, such as tilted data collection, to mitigate the issue [83].

Beyond preferred orientation, other factors that negatively impact data quality, such as sample degradation or the presence of multiple conformational states, are also readily identified through 2D and 3D analysis. Importantly, these problems often require further optimization of the sample preparation itself. If this initial analysis can be performed while the grid is still in the microscope, a well-informed decision can be made as to whether to continue with additional data collection, adjust collection parameters, or abort the session and return to the sample preparation stage. The sooner this decision can be made, the more efficiently valuable microscope time and computational resources will be used.

3.4.2 Image processing strategy for 2D and 3D Analysis

A central innovation of this workflow is the branching of two parallel, independent processing streams for 2D and 3D classification. Our principal goal is to generate a comprehensive structural overview of the specimen from two different approaches and dimensions. This strategy is divided into three main parts (Figure 39):

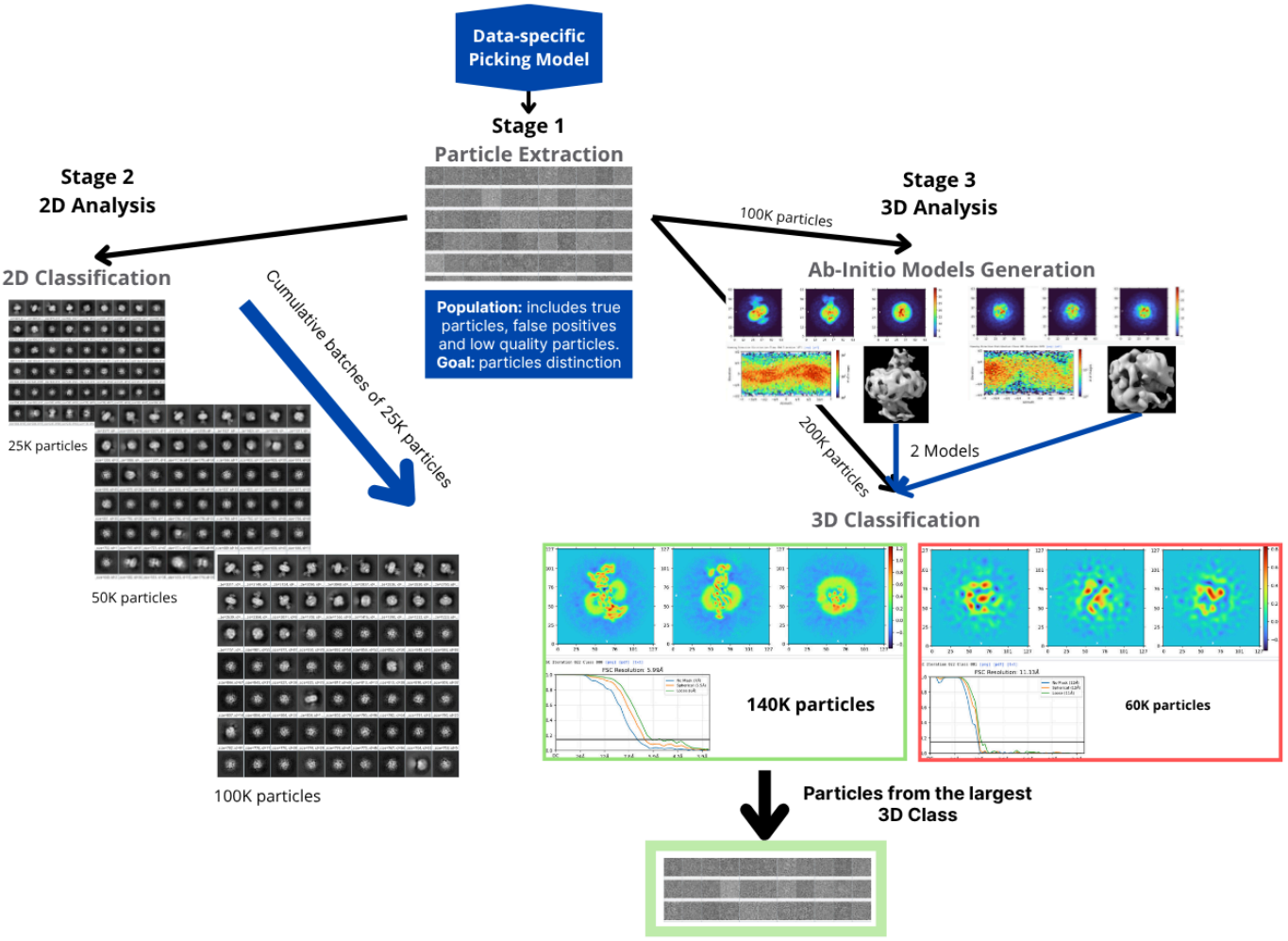


Figure 39. 2D and 3D analysis schematic. Each stage corresponds to the three main processing components involved in this analysis, which are described in detail below. Data is from EMPIAR-11057.

Stage 1: Particle Extraction

Particle extraction involves excising 2D images of the particles from the full micrographs based on the coordinates from the picking stage. As simple as it may look, several key considerations in this step are critical for the performance of subsequent processing:

A. **Extraction Box Size:** The extraction box must be large enough to contain the entire particle and its associated signal but not so large that it includes neighboring particles or introduces excessive computational cost. It needs to be larger than the picking box size because the CTF delocalizes high-resolution information, effectively spreading signal from the particle into the surrounding area. To capture this delocalized signal, which is essential for accurate alignment, a common rule of thumb is to use an extraction box

that is 1.5 to 2 times the particle diameter. Our *pwem - box size related parameters* protocol automates this by setting the extraction box size to twice the picking box size.

B. **Particle Pre-processing:** A series of standard pre-processing steps are applied to the extracted particles. These include inverting the image contrast so that particles appear white on a black background, and normalization, which sets the background noise to have a zero mean and unit standard deviation. Normalization is crucial as it removes background gradients and places all particles on a common contrast scale, preventing high-contrast particles from dominating downstream classification algorithms. Additionally, particles near the micrograph borders are removed, and a dust removal filter is applied to eliminate pixels with unusually high values.

C. **Rescaling and its Importance:** A final, critical step is particle rescaling. Raw micrographs can be very large (e.g., 4k x 4k pixels), resulting in large extracted particle images (e.g., 350x350 pixels). Processing hundreds of thousands of such large particles is computationally prohibitive for initial analysis. While processing algorithms typically load data in batches to manage memory, the total storage and computational overhead remain substantial. For example, a dataset of 100,000 particles with a box size of 350x350 pixels (stored as 32-bit floats) would occupy approximately 45.6 GB of storage. Furthermore, algorithms like 2D classification and initial model generation typically operate at a maximum resolution of 6-12 Å. Processing full-size particles would dramatically increase the computational time, which for many algorithms scales with the number of pixels, without providing any benefit, as the highest-resolution information is not used at this stage. Therefore, to optimize for both speed and memory, our workflow, using the *xmipp - extract particles protocol*, rescales all particles to a standard box size of **128x128 pixels**. This size is large enough to preserve the information needed for initial analysis while drastically reducing the computational burden.

The particle extraction and pre-processing in this workflow are derived from the *xmipp - extract particles protocol* within the *Xmipp* software suite, which provides all of these options.

Stage 2: 2D Analysis

The 2D analysis stream provides the first high-quality look at the particle's structure from different viewpoints. Although not used for particle cleaning in this specific workflow, 2D classification is a powerful tool for separating true particles from false positives. The process is achieved through 2D classification, an unsupervised machine learning method that groups similar particle images together.

The process is iterative: a set of initial random reference images (class averages) is generated, and each raw particle is compared and aligned to every reference to determine the best match. Particles are then assigned to the class of their best-matching reference. In the next iteration, new class averages are calculated from the particles assigned to each class, and these new averages become the references for the next round of alignment and assignment. This process is repeated until the class assignments stabilize. The mean of all particles within a group is calculated to produce a high-SNR "class average," which serves as a detailed representative image of that particular view.

To provide continuous feedback during data acquisition, our pipeline employs a repetitive classification strategy. The process begins once an initial set of 25,000 particles has been extracted. This first batch is classified to provide an early glimpse of the sample quality. Subsequently, new classification jobs are run with each new batch of 25,000 particles, incorporating all previously collected data up to a maximum of 100,000 particles. This approach ensures that the 2D averages are continuously updated and improved as more data becomes available, allowing for near-real-time monitoring of the experiment. The number of classes is a key parameter, as increasing the number of classes also increases the computational time. A common rule of thumb is to use one class per 500 particles, which provides a satisfactory balance between detail and speed. In our workflow, to prioritize rapid feedback, we use 50 classes for the initial 25,000 and 50,000 particle sets and increase this to 100 classes for the 75,000 and 100,000 particle sets.

Our workflow can perform 2D classification using two different, widely used software packages: *RELION* [37] and *CryoSPARC* *[29]*. While both solve the same multi-reference alignment (MRA) problem, they employ different underlying algorithms:

- ***RELION*** relies on a maximum-likelihood approach implemented through an Expectation-Maximization (EM) algorithm. This is a statistically rigorous method known for its robustness and ability to produce high-quality, detailed class averages, though it can be computationally intensive.
- ***CryoSPARC*** utilizes a stochastic gradient descent (SGD) approach with a branch-and-bound algorithm to accelerate the search for the optimal alignment. This method is generally much faster than the EM approach, making it particularly well-suited for the rapid, iterative feedback required in an on-the-fly pipeline.

The choice between these algorithms is often a practical one. Due to licensing models, *CryoSPARC* is freely available for academic use, while industry users require a license. For this reason, our pipeline is designed to be flexible, allowing the use of *CryoSPARC* in academic

settings and *RELION* in industrial contexts, ensuring broad applicability without sacrificing the quality of the on-the-fly feedback.

Stage 3: 3D Analysis

Concurrent to the 2D analysis, the 3D stream generates two initial 3D structures and is set to divide true particles from noise and contaminants. An initial 3D structure determination is necessary when no prior template of the structure exists, and it also serves as a crucial validation step to ensure that the experimental data can independently produce a sensible 3D model without being biased by an external reference. For this reason, no symmetry nor 3D template is imposed during this initial model generation.

The strategy is based on the assumption that the particle set, while high-quality, still contains a mixture of true particles and false positives (e.g., noise, ice artifacts). During the de novo 3D models generation and 3D classification, true particles, which share common structural features, will converge into a class that resolves into a recognizable macromolecular shape. In contrast, false positives, which lack a consistent structure, will be grouped into a separate, unstructured, blob-like class. With the recognizable 3D class, we are going to be able to observe the direction distribution in our sample, starting to obtain hints about problems such as preferred orientations.

Unlike the iterative 2D analysis, the 3D analysis is performed on larger, discrete batches of particles, as a substantial number of particles (typically 50,000-100,000) is required to generate a reliable initial model and to ensure that the subset is representative of the full dataset for subsequent *3D Refinement Steps*. The process is hierarchical and uses a significant subset of particles for the analysis:

1. **Initial Model Generation:** The first subset of up to 100,000 particles is used to generate two independent *de novo* 3D models using an SGD-based approach in either *RELION* or *CryoSPARC*. The algorithm starts with random angular assignments for each particle and iteratively refines both the 3D model and the particle orientations simultaneously. By optimizing a cost function that measures the disagreement between the experimental particle images and projections of the current 3D model, SGD efficiently explores the vast space of possible structures and orientations to converge on a self-consistent model. This approach is well-suited for our automated strategy because it is robust to the low SNR of individual particles and does not require pre-classified 2D averages as input, allowing for a direct path from extracted particles to a 3D model. This is performed without symmetry or an external template to avoid bias.

2. **3D Classification with Refinement for Particle Cleaning:** These two *ab initio* models serve as references for a 3D classification of a larger, second particle subset (up to 200,000 particles). This step is fundamentally a 3D multi-reference alignment, analogous to the 2D classification process. Each particle is compared to 2D projections of the 3D reference models across all possible orientations. The particle is then assigned to the class corresponding to the 3D reference that yields the best alignment score. This process is iterated, and the 3D class volumes are refined based on their assigned particles. This is a powerful and decisive particle-cleaning step because it separates well-behaved particles that align consistently with a structured reference from non-particles, or "junk" that align poorly. Based on the assumption that our picking strategy is effective, the 3D class with the largest particle population is automatically selected, as it typically represents the most prevalent and stable state of the macromolecule, as exhibited in Figure 40 examples. This yields a high-purity particle set for the final refinement stages.

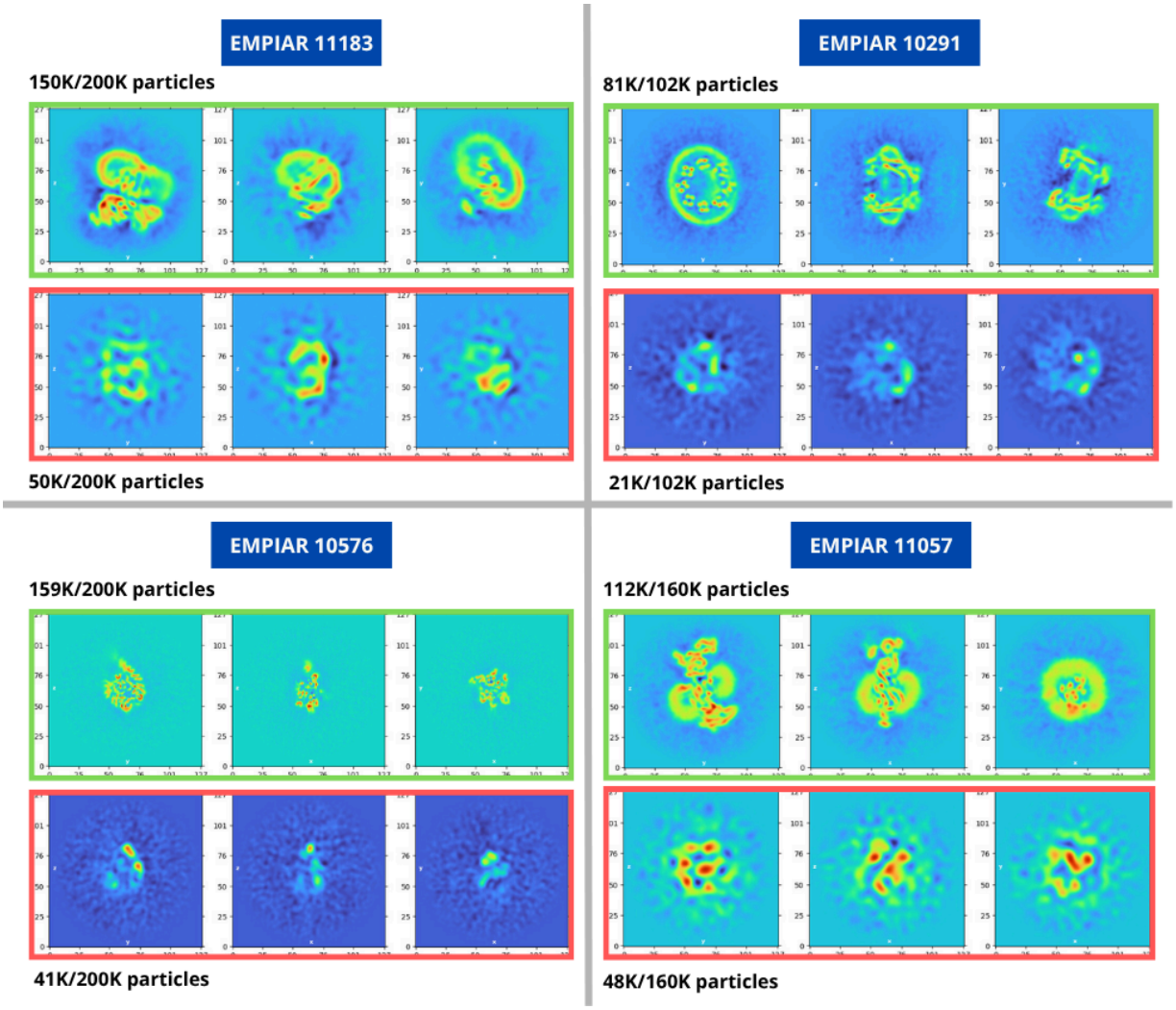


Figure 40. 3D Classification examples. Each quadrant corresponds to an EMPIAR entry used to illustrate particle separation during 3D classification with refinement between two *ab initio* models (using *CryoSPARC* algorithms within Scipion). The class with the larger particle population is highlighted in green, while the smaller class is highlighted in red. Particles belonging to the smaller 3D class are excluded from further processing. Data are from EMPIAR-11183, EMPIAR-10291, EMPIAR-10576, and EMPIAR-11057.

Similar to the 2D analysis, our pipeline offers both *CryoSPARC* and *RELION* for these tasks, allowing for flexibility based on software availability and licensing.

3.5 Refinement and Parallel Validation

Following the initial 2D and 3D analyses, the particle set, while significantly improved, may still contain suboptimal particles or unresolved conformationtal heterogeneity. The presence of these particles can limit the final resolution of the 3D reconstruction. Therefore, a final, more

rigorous stage of particle curation is necessary to produce a particle set of the highest possible quality and yield a refined 3D reconstruction. Figure 41 illustrates the specific Scipion workflow that carries out this final refinement stage.

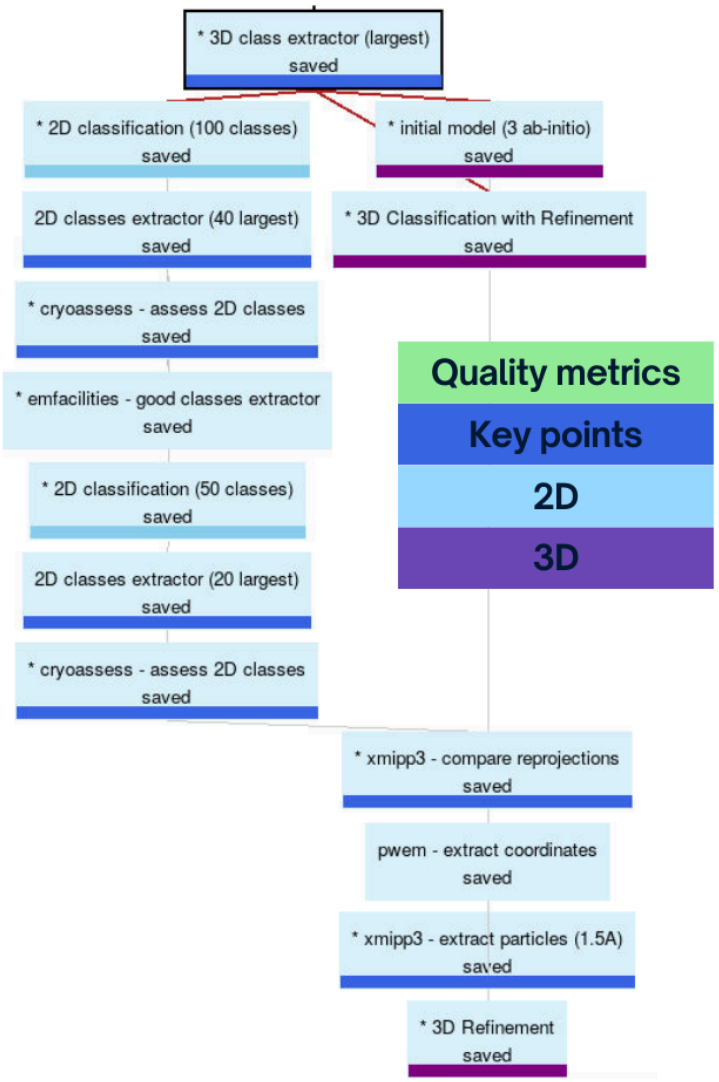


Figure 41. A detailed Scipion workflow diagram for the Refinement and Parallel Validation stage. Dark blue labels highlight the key steps of this stage. Sky blue labels correspond to the image processing steps related to 2D analysis, while purple labels denote those related to 3D analysis.

3.5.1 Image processing strategy for for Refinement and Parallel Validation

Our pipeline employs a parallel validation strategy where the filtered particle set from the previous stage (the most populous and stable 3D class) is processed through two independent and complementary branches: one focused on consecutive 2D classification jobs and the other

on 3D classification. The results are then cross-validated to select the most promising particle set for final high-resolution refinement. This parallel approach provides a robust internal validation, increasing the confidence in the final particle selection (Figure 42).

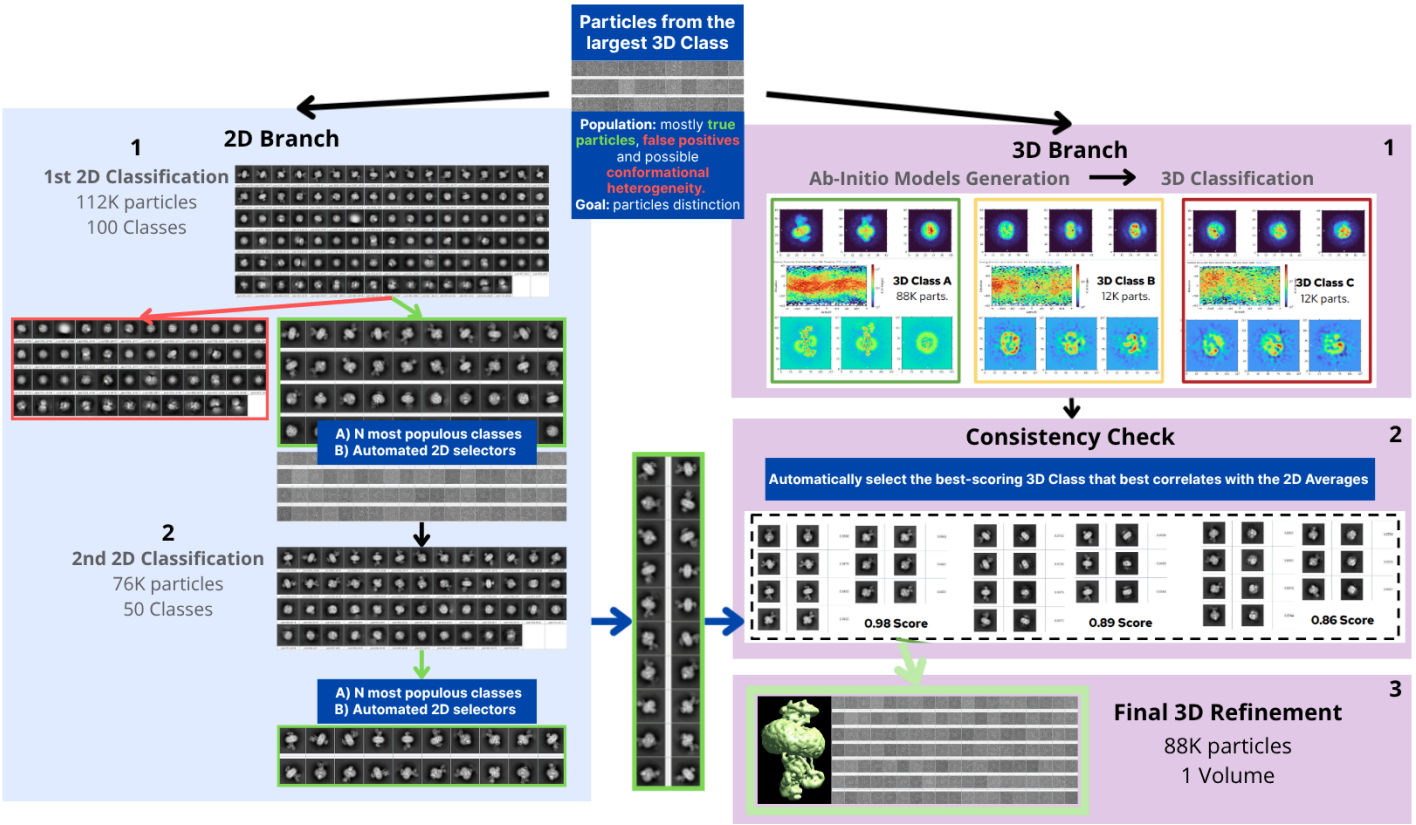


Figure 42. Refinement and parallel validation schematic. Titles in bold black indicate the main processing components involved in this workflow, which are described in detail below. Sky-blue regions correspond to image processing steps related to 2D analysis, while purple regions denote those associated with 3D analysis. Data is from EMPIAR-11057.

2D Branch: Consecutive Classification and Selection

The goal of the 2D Branch is to obtain a high-resolution set of 2D class averages through a two-round strategy combining 2D classification with automated, population-based and deep-learning-based selection:

1. **First Round of Classification and Selection.** The process begins with a 2D classification of the input particle set (typically at least 100,000 particles) into 100 classes. To ensure a robust result, the number of iterations is increased compared to the initial 2D analysis (from 20 to 30 normal iterations, with 3 final full iterations). This more exhaustive search leads to higher quality 2D averages. Once the classification is complete, a two-step automated selection follows:

1.1. **Population-based Filtering:** A custom program, the *numeric classes extractor protocol*, selects the 40 most populous 2D classes. The rationale is that classes with more particles are more likely to represent stable, high-SNR views of the structure, while classes with very few particles are more likely to contain noise or artifacts that could not be properly classified.

1.2. **Deep Learning-based Selection:** These 40 selected classes are then evaluated by *2D Assess*, a deep learning tool for automated 2D class average assessment [36]. This tool uses a pre-trained ResNet50 convolutional neural network to classify each 2D average into one of four categories: *good*, *clip*, *edge*, or *noise*. Only particles belonging to the classes labeled as "*good*" are retained for the next round. While other automated selectors exist, such as Cinderella [89] and Relion's 2D Ranker [37], experimental tests showed that 2D Assess provided the most reliable performance for this task.

2. **Second Round of Classification and Selection.** The curated particle set from the first round undergoes a second, more refined round of 2D classification. The number of classes is reduced to 50 to reflect the smaller number of input particles. A key parameter, the uncertainty factor, is increased from 2 to 4. This parameter controls how quickly the algorithm converges on class assignments; a higher value encourages the algorithm to remain "uncertain" for more iterations, which helps to separate subtly different but still valid particle views, thereby maximizing the diversity of the final class averages. Following this second classification, the same two-step selection process is then applied:

   2.1. **Population-based Filtering:** The 20 most populous classes are selected.

   2.2. **Deep Learning-based Selection:** *2D Assess* is used again to select only the "good" classes from this subset.

This two-round, hybrid selection strategy is designed to be more robust than using a single method alone. While DL-based selectors are powerful, they can sometimes be misled by high-resolution artifacts or well-centered, blob-like "junk" classes. By pre-filtering based on population, we ensure that the DL selector is primarily evaluating classes that are already statistically significant, thereby reducing the chances of it selecting false positives. This yields a high-resolution set of 2D class averages.

3D Branch: 3D Classification with Refinement

In parallel, the 3D branch aims to achieve two goals: first, to generate a set of refined 3D classes that can separate particles based on conformational or compositional differences, and second, to produce a final, highly curated particle set from the best-resolved class.

The process begins with the same input particle set as the 2D branch. This set is used to generate three independent *de novo* 3D models using an ab initio reconstruction algorithm. These volumes then serve as references for a final, exhaustive round of 3D classification with refinement. This step is designed to rigorously separate any remaining noise and heterogeneity, with the expectation that at least one of the three classes will converge to a high-resolution structure representing the main particle state. Other classes will either capture alternative conformations, lower-quality particles, or residual noise and artifacts.

This last 3D classification helps us not only to extract a final set of high-quality particles for refinement but could also give us some hints that heterogeneity is present in the sample being acquired. The class that yields the highest resolution, as determined by the Fourier Shell Correlation (FSC), could be automatically selected. However, relying solely on the FSC value can be misleading, as a "junk" class composed of noise can sometimes produce a high-resolution FSC curve due to overfitting, especially if the mask used for the calculation is too tight. A more robust method is needed to ensure the selected class truly represents the best particle population. Therefore, we apply a more sophisticated approach that takes advantage of the 2D branch results, as described in the next section.

Consistency Check: Cross-Validation of 2D and 3D Results

This parallel structure enables a powerful self-consistency check. This novel step uses the high-resolution 2D class averages from the 2D branch as an independent ground truth to objectively rank and select the best 3D class from the 3D branch. The process is as follows: for each of the three 3D classes, a complete set of 2D projections is generated, covering all possible viewing directions. Then, for each of the "good" 2D class averages selected by the 2D branch, an exhaustive search finds the best-matching projection from each of the three 3D volumes, and a cross-correlation score is calculated for each of these best matches. Finally, for each of the three 3D volumes, the mean of these cross-correlation scores is computed (Figure 43).

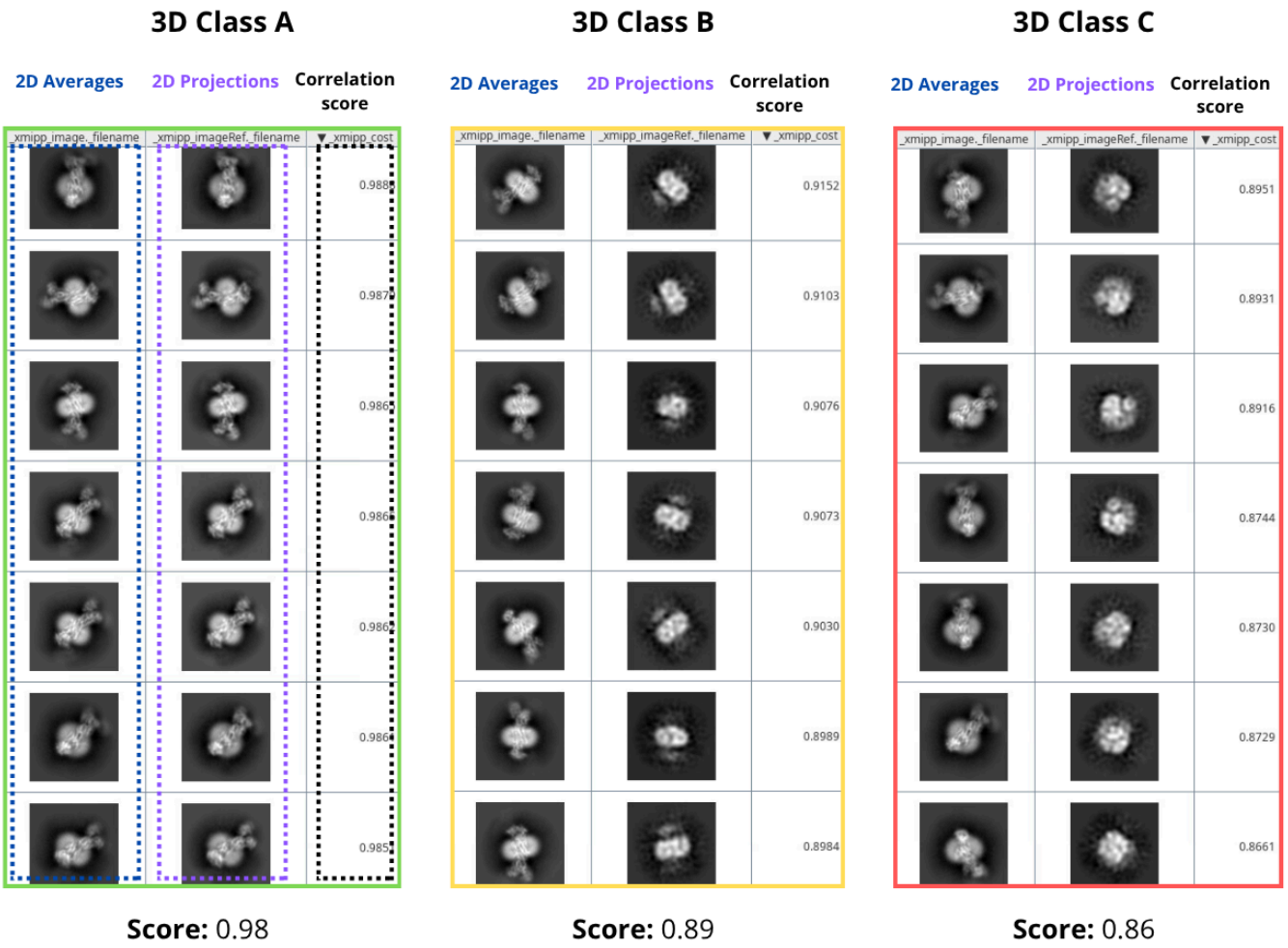


Figure 43. Cross-Validation of 2D and 3D Results. The titles at the top indicate the 3D class being evaluated. The blue column shows the high-quality *2D averages* from the 2D branch, and the adjacent purple column shows the *2D projections* of the corresponding 3D-class volume. These projections represent the best matches to the 2D averages after an exhaustive search. The *Score* below refers to the average correlation between the set of 2D averages and their best-matching 2D projections. The 3D class with the highest score is selected to continue in the image-processing pipeline.

This provides an overall score indicating how well each 3D volume agrees with the 2D data. High-resolution features present in the 2D averages will only correlate well with projections from a 3D volume that also contains those features. A blob-like "junk" volume, on the other hand, will produce featureless projections that will correlate poorly with the detailed 2D averages. The 3D volume with the highest mean correlation score is therefore the one most consistent with the 2D data and is selected as the best representation of the structure.

This ranking is performed automatically by a modified version of the *xmipp compare reprojections protocol* in the *Xmipp* software suite. The program outputs the highest-scoring 3D

class, providing both the 3D volume and its corresponding set of curated particles, which are then passed to the final refinement step.

Final 3D Refinement

With an optimal particle set and a reliable 3D model, the pipeline is ready for the final refinement. The goal of this step is to yield a preliminary high-resolution 3D reconstruction before the data acquisition session has concluded. This is divided into two main parts:

1. **Re-extraction of Particles at Higher Resolution:** For the initial analysis stages, particles were downsampled to a small box size (e.g., 128x128 pixels) to accelerate processing. However, this process discards high-frequency information. To recover this information for the final reconstruction, the curated particles are re-extracted from the original micrographs at a larger box size. While re-extracting at the original pixel size would yield the highest possible resolution, this would also significantly increase the computational cost. To balance speed and quality for on-the-fly feedback, we re-extract the particles at a pixel size of 1.5 Å/px. Given that typical acquisition pixel sizes are around 0.5-1 Å/px, this represents a minimal downsampling that retains most of the high-resolution detail while still offering a significant speed advantage over processing at the original size.
2. **3D Refinement:** A high-resolution 3D refinement job is performed (using *CryoSPARC* or *RELION*) on the re-extracted particles. This is an iterative process that aims to find the optimal alignment for each particle with respect to the 3D reference and then reconstruct a new, improved 3D map from the aligned particles. The refinement continues until the 3D map converges and no further improvement in resolution is observed. The final resolution is estimated using the Fourier Shell Correlation (FSC), which measures the self-consistency between two independently refined half-maps. The FSC curve indicates the spatial frequency at which the two maps are no longer correlated, providing a global measure of the map's resolution (Figure 44). In this pipeline, *CryoSPARC* or *RELION* is used depending on the type of user (academia or industry).

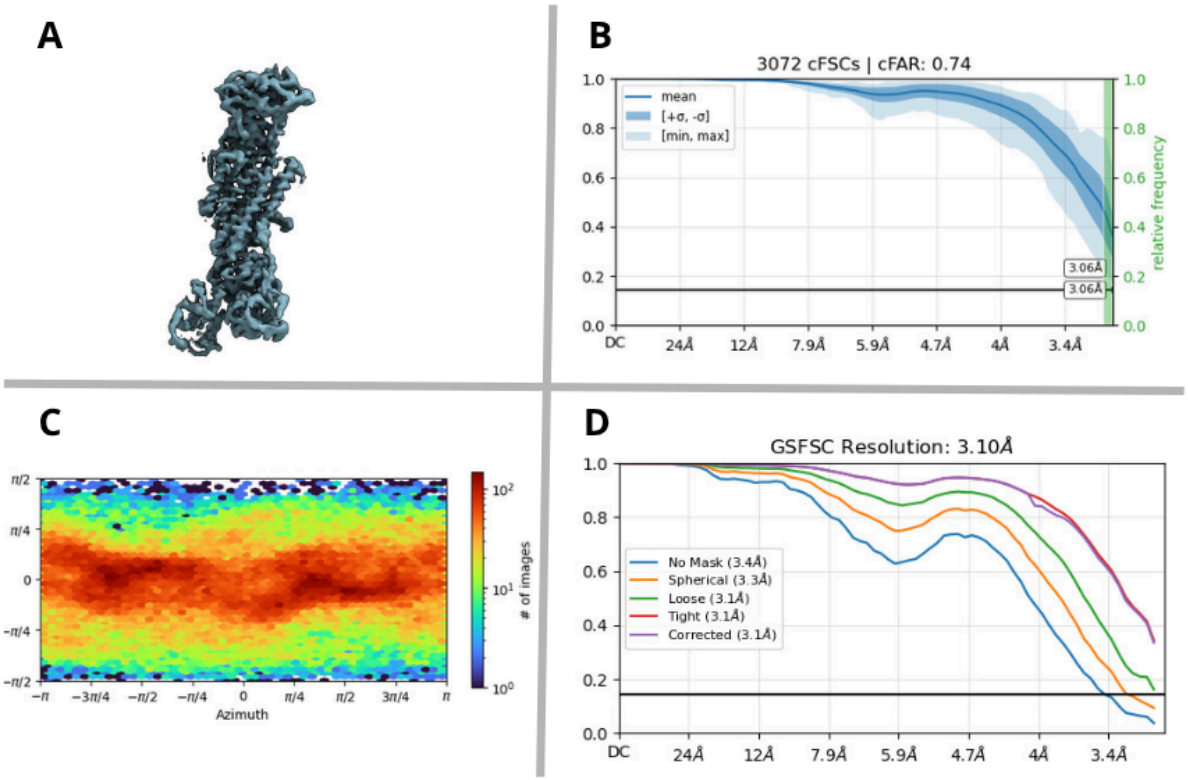


Figure 44. 3D Refinement Overview. Example of cryo-EM data processing results for a protein complex (hydrolase). Data is from EMPIAR-11057. **(A)** Reconstructed density map showing the refined 3D structure. **(B)** Resolution assessment using the **conical Fourier Shell Correlation (cFSCs)**, a method used to evaluate the directional quality of a cryo-EM reconstruction by computing FSCs within localized cones in Fourier space rather than over the entire volume. Multiple cFSCs are calculated along uniformly distributed directions, providing a comprehensive view of anisotropy in the reconstruction. The resulting plot displays the mean, minimum, maximum, and standard deviation of FSC values across directions; ideally, narrow standard deviation bands around the mean indicate isotropic resolution and uniform directional quality. **(C) Particle orientation distribution plot**, which visualizes the angular coverage of particle projections used for reconstruction. Uniform coverage indicates well-sampled orientations, whereas anisotropy or clustering reveals preferred orientations that can limit achievable resolution. **(D)** Global resolution assessment using the **gold-standard Fourier Shell Correlation (GSFSC)**, which measures the correlation between two independently refined half-maps at different spatial frequencies. The FSC curve typically includes unmasked, masked, and corrected versions: the unmasked curve often underestimates resolution due to background noise, the masked curve provides a more accurate estimate of signal correlation within the particle volume, and the **corrected** curve accounts for mask effects, offering the most reliable global resolution estimate. A sharp decay crossing the 0.143 threshold at high frequency indicates a high-quality, well-refined reconstruction with minimal overfitting.

It is important to note that the option to optimize the defocus per particle is deactivated in this step. This choice could considerably improve resolution, but it could potentially introduce systematic errors; therefore, a more thorough, per-particle CTF refinement is recommended as a separate post-processing step for the final structure. The FSC provides a global resolution measurement, which is sufficient for assessing the quality of this preliminary reconstruction.

However, for a more thorough validation, methods that compute the local resolution would be necessary to account for spatial variations in the map's quality.

3.6 3D Workflow implementation

This section details the practical implementation of the pipeline, from the final outputs provided to the user to the underlying software ecosystem and its adaptation for high-performance computing (HPC) environments.

3.6.1 Pipeline Deliverables and Outputs

Upon completion of the data acquisition session, the automated pipeline concludes, providing the user with a comprehensive suite of processed data and quality control metrics. This organized output serves as a robust foundation for subsequent, more detailed post-acquisition analysis. The key deliverables include:

- A complete set of aligned movies and their corresponding CTF estimations.
- A curated set of high-quality micrographs accompanied by detailed quality measurements.
- A data-specific particle picking model trained on the user's sample.
- A complete set of picked particle coordinates from all curated micrographs.
- High-quality 2D class averages representing the particle's different views.
- The best 3D volume from the parallel validation stage, along with its associated high-purity particle set.
- A preliminary high-resolution 3D structure.
- A complete record of all processing steps and their outcomes within the Scipion project, providing critical feedback to guide further, in-depth image processing.

3.6.2 Workflows and documentation

The workflow templates used in this thesis are publicly available in the WorkflowHub collection called “CryoEM Facility Workflows” [97], which also provides detailed documentation. This includes two key documents: (i) a manual for using Scipion in the context of a cryo-EM facility, focusing on three main areas: on-the-fly processing, workflows and templates design, and queue systems integration; and (ii) a more specific guide describing the available templates, image processing details, implementation details, and software requirements (Figure 45). WorkflowHub itself is a community registry for computational workflows [98].

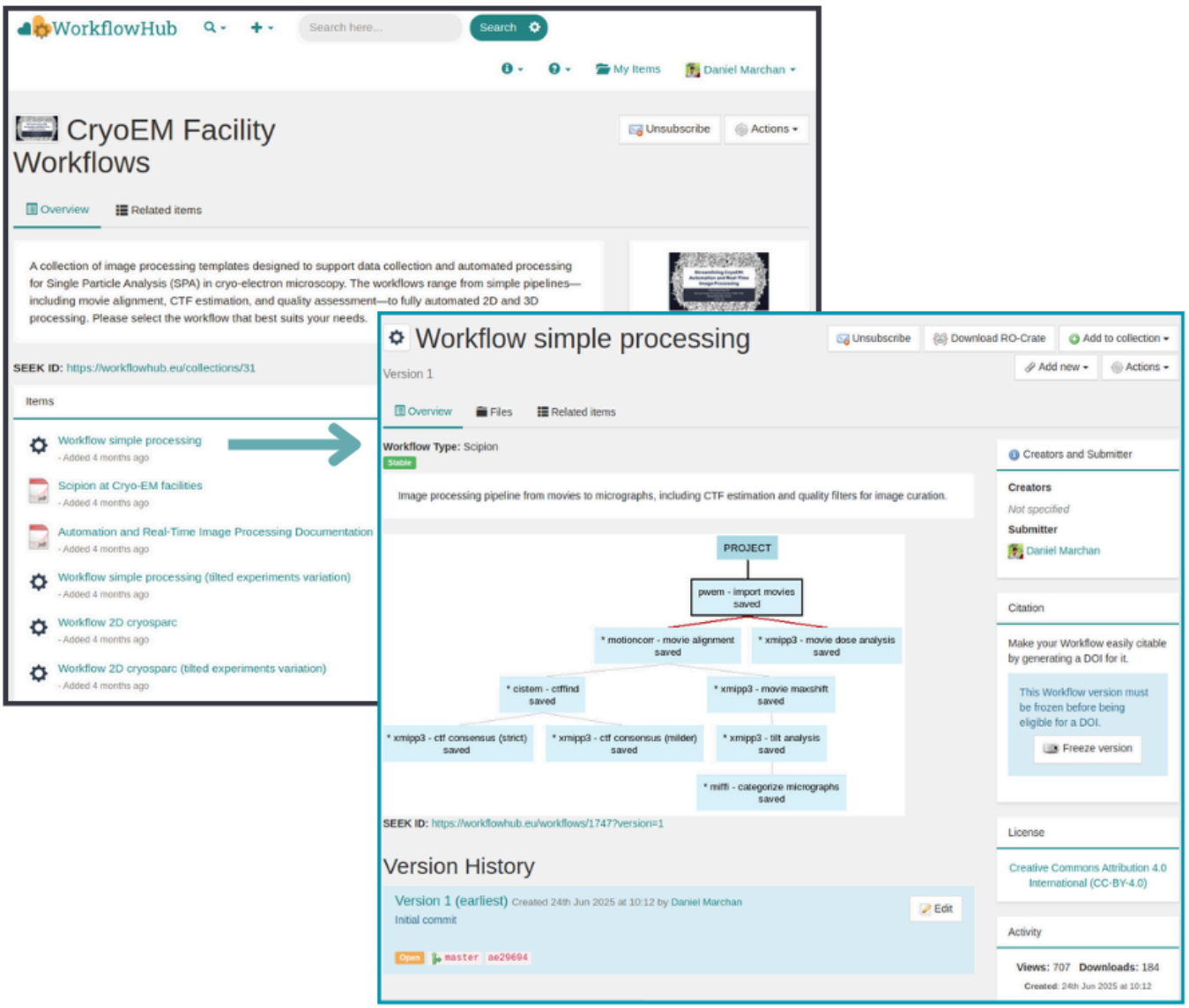


Figure 45. WorkflowHub *Scipion* Webpage. The black-highlighted panel corresponds to the main view of the *CryoEM Facility Workflows* collection uploaded by the *Scipion* Team. This section displays all items related to the collection, along with a brief description of its content. The blue-highlighted panel shows the main view of one of the workflows included in the collection. Here, you can find a short description at the top, all associated files, a representative image of the workflow, version history, license information, and activity metrics (such as views, downloads, and creation date).

Available Templates

The following image processing templates are available:

- **Workflow Simple:** Image processing from movies to micrographs, including CTF estimation and quality filters for image curation.

- **Workflow 2D *CryoSPARC*:** Automated processing from movies to 2D classes. Includes curation, box size estimation, data-specific picking model training, and three 2D classification jobs (25k, 50k, and 100k particles) using *CryoSPARC*.
- **Workflow 2D *Relion*:** Same as the 2D *CryoSPARC* workflow, but using *RELION* for 2D classification.
- **Workflow 3D *CryoSPARC*:** A complete, automated workflow from movies to 3D reconstruction. Includes all steps from the 2D workflow, plus 3D processing of the first 200k particles. It generates multiple *de novo* 3D models without symmetry and performs automatic 2D/3D class selection using *CryoSPARC*.
- **Workflow 3D *Relion*:** Same as the 3D *CryoSPARC* workflow, but using *RELION* for 2D/3D classification and refinement.

Specialized templates are provided for tilted experiments, which are modified to handle the high drift and lower resolution commonly observed in such datasets:

- **Workflow Simple - Tilted Samples**
- **Workflow 2D *CryoSPARC* - Tilted Samples**
- **Workflow 2D *Relion* - Tilted Samples**
- **Workflow 3D *CryoSPARC* – Tilted Samples**
- **Workflow 3D *Relion* – Tilted Samples**

Key adjustments in these "Tilted" templates include:

- **Relaxed Motion Correction Filters**: Max per-frame shift = 30Å; Global drift = 120Å.
- **Relaxed CTF Filters**: Resolution cutoffs adapted to 6.5 Å and 8.5 Å.
- **The *Defocus Sampler Protocol*** uses a larger pool of 500 images to select the 25 training examples, ensuring a more generalizable picking model.

Software and plugins

These complex workflows require 11 external software packages, which are integrated as *Scipion* plugins (*scipion-em-plugin*). All plugins are freely available from the official *Scipion* GitHub organization (https://github.com/scipion-em).

Before importing a template, all required plugins must be installed to prevent the pipeline from breaking in the missing protocol from the missing plugin. The required plugins are the following:

A. **Xmipp** requires its own set of installation and compilation steps:
   - *"xmipp-bundle"*

- *"scipion-em-xmipp"*

B. Plugins with software binaries. These plugins download the necessary external software and binaries**:**

- *"scipion-em-cistem"*
- *"scipion-em-gautomatch"*
- *"scipion-em-miffi"*
- *"scipion-em-motioncorr"*
- *"scipion-em-sphire"*
- *"scipion-em-topaz"*
- *"scipion-em-emfacilities"*
- *"scipion-em-repic"*

C. Plugins without software binaries. External software or models are not downloaded automatically with the plugin:

- *"scipion-em-cryosparc2": CryoSPARC* software is not installed automatically with the plugin. The scipion.conf file must be edited to point to an existing CryoSPARC installation.
- *"scipion-em-cryoassess":* The pre-trained models must be downloaded separately, and the scipion.conf file must be edited to specify their location.

Special instructions for each of the plugins can be found in the *Scipion* GitHub organization.

3.6.3 HPC Queue Systems and Adaptation

In cryo-EM facilities, image processing pipelines can operate under two distinct paradigms depending on the computational environment: **queue-less systems** and **queue-based systems**. In a queue-less setup, GPU resources must be assigned manually to each protocol, which is both cumbersome and error-prone. This approach provides only limited control over resource usage and user access, often leading to inefficient GPU utilization and potential conflicts when multiple users share the same hardware. In contrast, queue-based systems automate the allocation of available computational resources, such as GPUs and CPUs, thereby reducing manual intervention. This automation minimizes assignment errors, enhances control over resource usage, and improves access management across users (Figure 46).

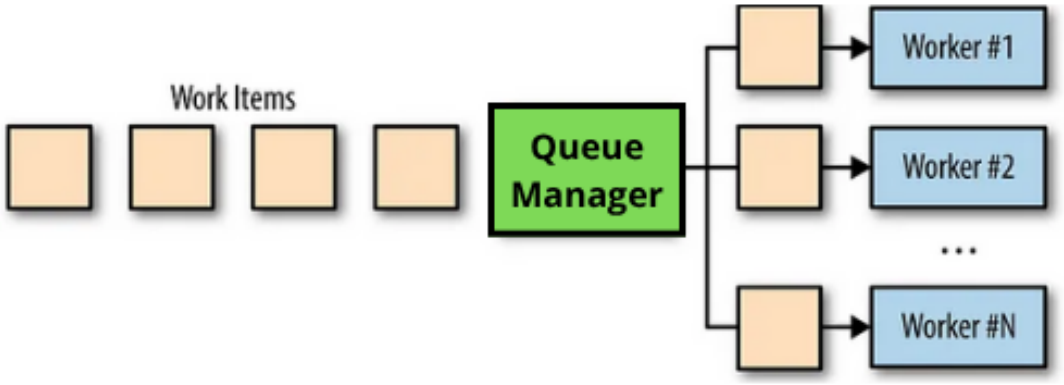


Figure 46. Representation of a Queue Management System. This diagram illustrates how work items are distributed across different worker nodes (cluster nodes) by the Queue Manager (e.g., SLURM).

The use of a queue system becomes essential in shared computing environments for several reasons. First, computational resources in HPC clusters and data centers are inherently limited and must be distributed efficiently among multiple concurrent users. A structured queue mechanism ensures that these finite resources are shared fairly, preventing overload or conflicts. Moreover, adopting a queue-based workflow significantly improves overall efficiency by maximizing resource utilization and minimizing idle time. It also provides fine-grained control over resource assignment and enables prioritization of critical jobs, ensuring that high-priority or time-sensitive tasks can be executed preferentially. In summary, while queue-less systems may suffice for small-scale or single-user setups, queue-based systems are indispensable in multi-user or resource-constrained environments. Their integration within frameworks such as *Scipion* ensures efficient task scheduling, stable operation, and optimal use of GPU and CPU resources across the entire cryo-EM data-processing workflow.

A key feature of the *Scipion* workflow is its ability to initiate image processing immediately after data collection and to manage computational resources automatically via job scheduling systems such as **SLURM**. This design ensures scalability from small laboratory workstations to large high-performance computing (HPC) clusters, supporting complex multi-GPU workflows that operate robustly and autonomously.

In Scipion, this integration is particularly critical for automating a heterogeneous pipeline. Instead of executing a single, monolithic job, the workflow is divided into multiple discrete steps, each with different computational requirements. Acting as a meta-scheduler, *Scipion* submits each protocol as a separate job to SLURM, which enables:

- **Efficient Resource Allocation:** CPU-intensive tasks (e.g., file I/O, preprocessing scripts) run on CPU-only nodes, while computationally demanding steps (e.g., motion correction, picking model training, and 2D/3D classification) are executed on GPU-equipped nodes.
- **Job Dependency Management:** Scipion automatically creates dependencies between jobs. For example, a 2D classification job will not start until all its prerequisite particle extraction tasks have successfully finished.
- **Resilience and Error Handling:** If a job fails due to a transient hardware or resource issue, the entire workflow does not collapse. The error is logged, allowing users to fix and restart the workflow from the point of failure without reprocessing earlier steps.

This architecture significantly enhances control over resource usage, minimizes errors from manual intervention, and is essential for deploying the pipeline in a multi-user facility environment. Further configuration details are available in the *Scipion for Facilities Course* (Section 3; [97]).

New implementation

During this project, a key limitation related to GPU resource usage was identified, particularly in environments where data is processed on a **fat node** (a single, high-capacity workstation) rather than a full HPC cluster. In such systems, traditional queue mechanisms are often suboptimal. The default behavior of Scipion when operating with SLURM is to submit an entire protocol as a single queued job. If the protocol requires GPU computation, SLURM reserves one GPU for the entire duration of that protocol, even if GPU usage occurs only during specific steps. This approach restricts the number of concurrent GPU protocols to the number of available GPUs and can lead to significant underutilization of resources (Figure 47).

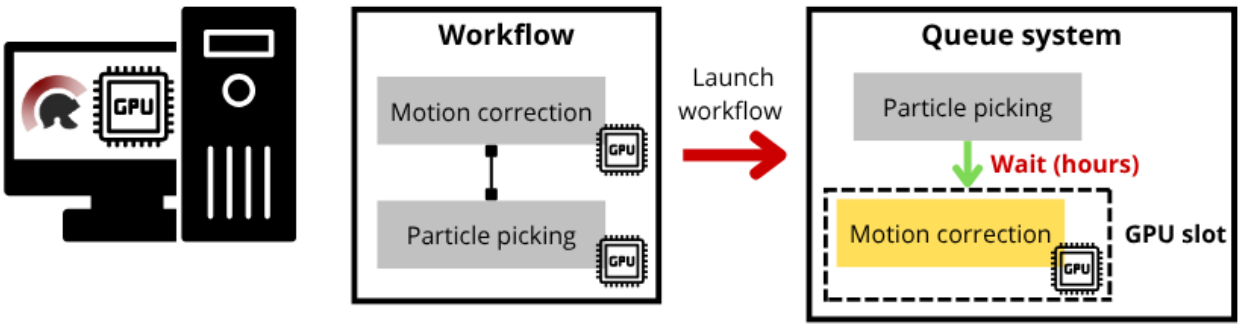


Figure 47. GPU number limiting factor. This diagram illustrates a use case in which a workstation has only one GPU and an on-the-fly image processing workflow consisting of motion correction and particle picking. Since both steps require GPU resources, the particle picking job (protocol) must remain in the queue until the entire motion correction process finishes and releases the GPU, creating a processing bottleneck.

For instance, this behavior limits us to running only one GPU-dependent protocol simultaneously in a one GPU machine, as shown in Figure 47. To work around this constraint, no queue system is to be applied, GPUs need to be manually assigned per protocol and shared. A process that required detailed knowledge of the pipeline and was both inefficient and error-prone. Incorrect GPU assignments could result in multiple protocols attempting to share the same GPU, leading to memory overload, execution failures (*core dumped* errors), and overall workflow instability.

To overcome this, we revisited and improved an experimental development within *Scipion* that allows submitting **individual actions within a protocol** as independent queue jobs rather than submitting the entire protocol at once. This approach enables protected and unattended execution in a more granular manner: GPU resources are allocated dynamically only for the duration of each action and then released. As a result, GPUs remain in active use rather than being reserved by idle protocols.

This granular execution unit corresponds to what we define as an **image processing step**. A Scipion protocol represents an image processing operation — for example, Motion Correction — but this operation must be applied to as many images as there are movies in the dataset. By splitting the execution into image processing steps per image or per image batch, we achieve finer task granularity, submitting smaller jobs to the queue rather than sending the entire protocol at once as it is represented in Figure 48.

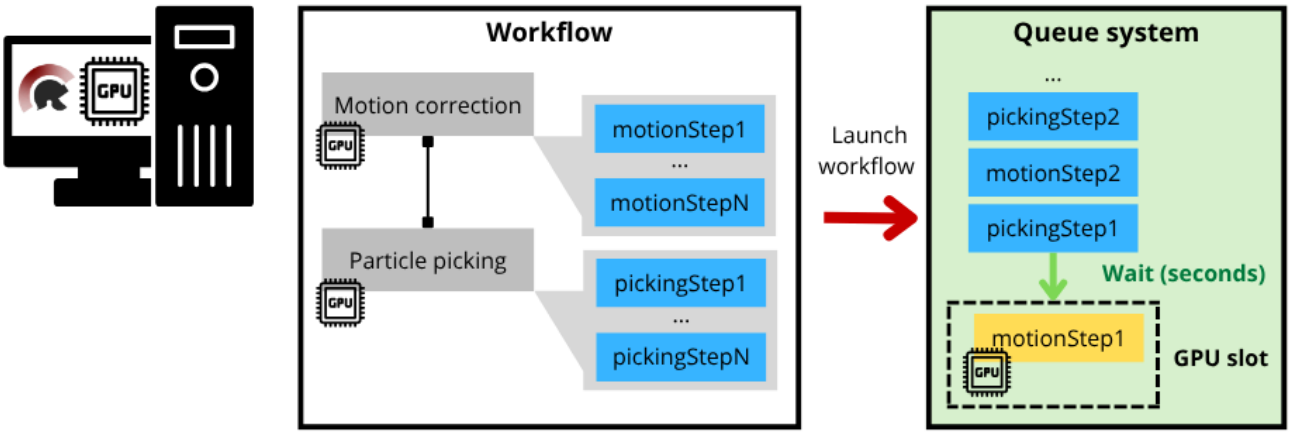


Figure 48. Resolving the GPU number limiting factor. This diagram illustrates the same use case of a workstation with a single GPU and an on-the-fly image processing workflow consisting of motion correction and particle picking. With the new implementation, protocols are divided into smaller image processing steps, allowing the submission of shorter jobs and reducing waiting times from hours to seconds. This enables multiple protocols to share and use the GPU slots in an orderly and efficient manner.

This new implementation offers several advantages:

- **Optimized GPU utilization:** GPUs are used only when needed and are available for other tasks when idle.
- **Improved stability:** Only one action per GPU is executed at a time, preventing memory saturation and reducing failures.
- **Increased parallelism:** Multiple protocols can now submit actions concurrently to the same GPU, improving throughput.

This development was crucial for enabling more complex multi-GPU workflows (Figure 49) to be executed during on-the-fly processing helping us reach further in the image processing pipeline.

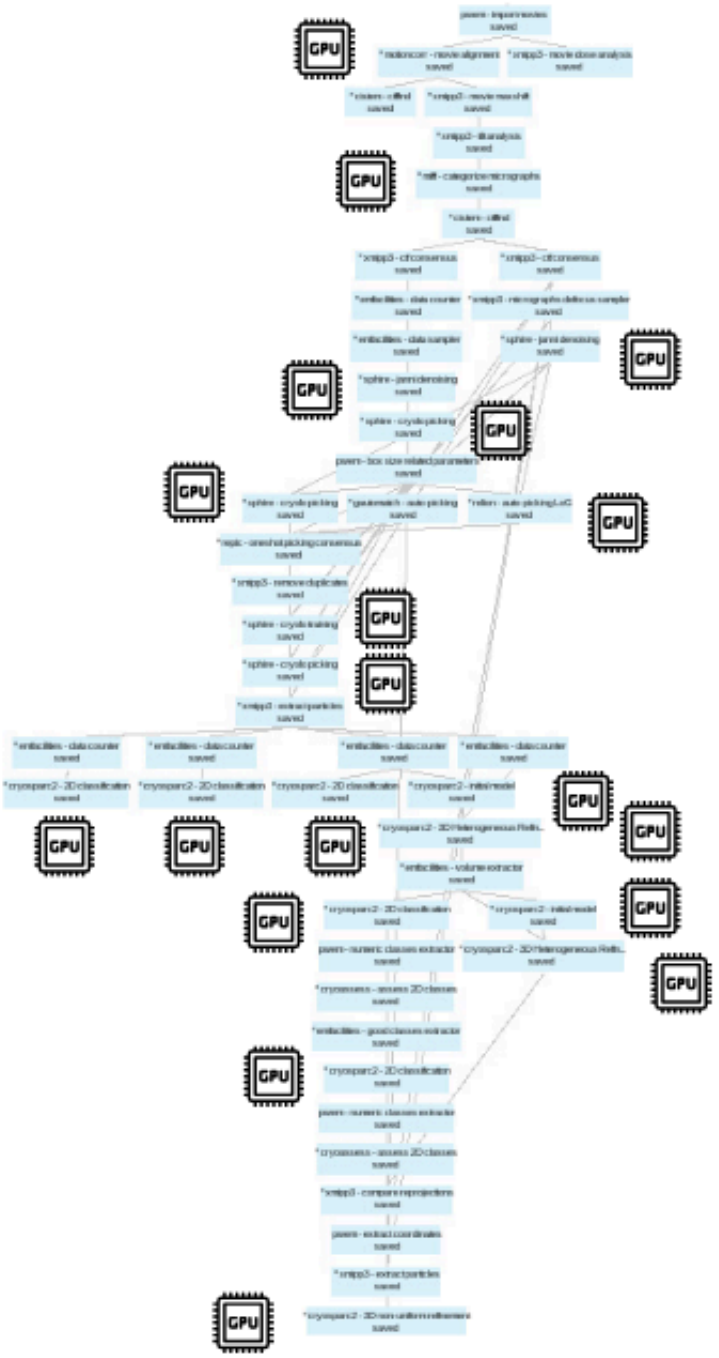


Figure 49. Example workflow illustrating GPU usage. This diagram shows which protocols in the automated image processing pipeline require GPU execution, indicated by the GPU icon. As illustrated, 19 out of 37 protocols depend on GPU computation, highlighting the importance of efficiently managing computational resources. Although not all protocols run simultaneously, manual resource allocation would be tedious and error-prone. Without a proper queue system with a granular job implementation, this workflow would require at least a 6-GPU machine to operate efficiently. With the new implementation, it could theoretically run on a single-GPU workstation—though multiple GPUs are still recommended for proper on-the-fly processing.

To implement this functionality, several modifications were made to the *Scipion* source code in collaboration with the *Scipion* development team. The system was designed to track all submitted jobs, ensuring that they can be monitored or canceled if a protocol is interrupted. What began as an experimental prototype has now become a stable feature within Scipion. This innovation has been requested by several cryo-EM facilities and represents a major step toward more flexible, general, and reliable management of GPU resources in automated image-processing workflows.

3.6.4 Data availability

This thesis analyzes both publicly available cryo-EM data from EMPIAR and private user data collected at the ESRF, which cannot yet be shared or disclosed. The accession numbers for the public datasets correspond to those included in the CryoPPP collection [3], all of which are shared within the thesis.

3.6.5 Courses and dissemination

As part of this thesis, the event *"Scipion for Facilities: Practical Course and Workshop"* (Madrid, November 5th–7th, 2024) was organized (Figure 50). Traditionally, only the workshop component had been conducted, delivered entirely online. In this edition, a significant step forward was taken by designing a hybrid event combining in-person and online sessions. The main objective was to present the technological developments achieved during this thesis and to strengthen the collaboration between the *Scipion* development team and the broader cryo-EM facility community.

Figure 50. *Scipion* for Facilities: Practical Course and Workshop (Madrid, November 5th–7th, 2024)

The event was structured in two main components:

Practical On-site Course

The first two days were dedicated to an intensive hands-on course focused on the use of Scipion within the operational context of a cryo-EM facility. For this purpose, preconfigured virtual machines were deployed on Amazon Web Services (AWS), each containing Scipion and all required software dependencies, along with representative cryo-EM datasets. The course was organized into several thematic modules (*Scipion for Facilities Course* [97]):

- Launching and analyzing on-the-fly image processing workflows.
- Adapting workflows to specific facility requirements using templates, enabling easy export and import of predefined pipelines.
- Configuring Scipion to operate with queue systems such as SLURM, emphasizing the importance of controlled computing environments for robust processing.
- One-to-one personalized sessions, during which participants received tailored guidance to design processing workflows adapted to their own institutional setups.

Many attendees were staff members from established cryo-EM facilities who were particularly interested in implementing Scipion for real-time data processing and facility-level automation.

Hybrid Workshop

The second component of the event consisted of a two-day workshop featuring technical presentations and invited talks.

- **Day 1:** Focused on *Laboratory Information Management Systems* (LIMS) and the concept of *Smart Data Collection*, highlighting the integration between data collection and computational processing.
- **Day 2:** Dedicated to automation in image processing for both single-particle analysis (SPA) and tomography. As part of this session, I participated as a speaker, presenting the main results and developments of this thesis related to automated on-the-fly processing. Additionally, several international cryo-EM facilities shared real-world use cases demonstrating their adoption of *Scipion* in production environments.

The event achieved remarkable success, with 15 in-person participants representing 12 different cryo-EM facilities, and more than 30 additional participants attending online. The hybrid format fostered an enriching exchange of knowledge, diverse perspectives, and constructive discussions regarding the future needs of cryo-EM data processing facilities.

From a personal and professional perspective, this experience was especially valuable, as it allowed me to participate both as an instructor in the practical course and as a speaker in the workshop, receiving direct feedback and validation on the research and developments carried out in this thesis. Furthermore, this event served as the foundation for formalizing the collaboration with the ESRF, where a subsequent three-month research stay was conducted to implement the automated processing workflow, results are discussed in the following section.

CHAPTER 4 – RESULTS

The following sections present the results obtained from applying this pipeline to a wide range of scenarios, from extensive benchmark datasets [3] to its real-world deployment as the standard processing solution at the high-throughput cryo-EM facility of the European Synchrotron (ESRF). The workflow templates used in this study are publicly available in the WorkflowHub collection called "*CryoEM Facility Workflows*" [97], which also provides detailed documentation.

4.1 Extensive benchmark

To validate the pipeline's robustness and general applicability, we performed an extensive benchmark on the CryoPPP dataset [3]. This dataset, derived from the Electron Microscopy Public Image Archive (EMPIAR), comprises 32 non-redundant, diverse protein targets that vary significantly in terms of size, molecular weight, symmetry, and sample characteristics. While the original CryoPPP dataset limits entries to approximately 300 micrographs, this number is often insufficient to yield enough particles for a meaningful 3D reconstruction. To create a test scenario that more closely mirrors a real-world, on-the-fly processing session, we expanded our test set by selecting up to 1,000 micrographs for each entry, ordered by filename in ascending order. This standardized subset size ensures sufficient data for a robust 3D reconstruction attempt while reinforcing the pipeline's diagnostic goal: to determine sample viability from a representative fraction of the data during the on-the-fly acquisition, rather than requiring a complete, multi-terabyte dataset.

A crucial aspect of this study is defining a success metric suitable for an automated diagnostic pipeline. While the gold-standard Fourier Shell Correlation (FSC) is indispensable for reporting the final global resolution of a manually processed structure, it can be a misleading metric for success in a fully automated context. An automated pipeline may converge on non-particle features (*e.g.,* ice contaminants, carbon edges) that still produce a high-resolution FSC curve, leading to a false-positive result. Conversely, a pipeline may correctly identify a challenging biological sample. Still, inherent issues, such as conformational heterogeneity or flexibility, which require expert manual intervention to resolve, may limit the achievable resolution, leading to a false-negative assessment of the pipeline's performance if judged solely by a resolution threshold.

Therefore, we defined the primary success criterion as the visual inspection of the final 3D reconstruction. A test case was considered successful when the resulting 3D map exhibited recognizable, protein-like features consistent with the expected size and shape of the target

macromolecule. This qualitative metric was chosen because it directly addresses the central question of whether the workflow can robustly handle the complexity of diverse specimens and acquisition conditions, providing a more meaningful measure of its value as a diagnostic tool for assessing sample quality and the feasibility of high-resolution structure determination.

These tests were performed on a CentOS 7 Linux server with 40 cores (2 × Intel Xeon Gold 6230, 2.20 GHz) and 384 GB of RAM. The workstation also featured four GPUs (Tesla T4 Driver Version 460.27.04, CUDA Version: 11.2) with 16 GB each. In terms of storage, it has four 8 TB SATA HDDs in a RAID 5 configuration for mass storage (where data was stored), two 1 TB SATA SSDs in a RAID 0 configuration for scratch, and two 240 GB SATA SSDs. This machine is housed within the Biocomputing Unit data center at the Spanish National Centre for Biotechnology (CNB-CSIC). It is important to note that *CryoSPARC* was used through the *Scipion* framework, via the *scipion-em-cryosparc* plugin, as the processing backend for the 2D and 3D steps.

As shown in Table 1, the pipeline successfully processed the majority of the datasets (94%), adapting to significant variations in particle distribution, noise, defocus, and sample conditions without manual tuning. In 78% of the samples, the workflow achieved a robust 3D reconstruction, revealing the symmetry and proper shape of the protein, and even reached the Nyquist limit of 3Å in some reconstructions (all particles were extracted at 1.5 Å/px). In the other 16% of cases tagged as successful, even when the study particles were present, the final 3D reconstruction was not optimal, lacking high-resolution details or specific particle orientations. For the entries that did not yield a structure (6%), a more in-depth analysis was performed to determine if the problem lay with the processing or was inherent to the dataset's complexity.

Table 1. General overview of CryoPPP benchmarking results. Green indicates successful and robust 3D reconstructions; yellow denotes suboptimal reconstructions; and red corresponds to non-conclusive 3D reconstructions.

3D Final Reconstruction Assessment	**Total Percentage of the 32 entries (%)**	**FSC range resolution (Å)**
High-quality	78 (25/32)	3.1 - 6.7
Suboptimal	16 (5/32)	3.7 - 11.9
Failed	6 (2/32)	N/A

To demonstrate the robustness of this diagnostic automated pipeline, we gathered six cases in which the pipeline performed **High-quality**, all the **Suboptimal** instances, and all the **Failed**

cases to facilitate a more in-depth analysis of the feedback provided by our tool. The image processing summary is presented in Table 2.

Table 2. CryoPPP Image processing summary. Green indicates successful and robust 3D reconstructions; yellow denotes suboptimal reconstructions; and red corresponds to non-conclusive 3D reconstructions. **Blue highlights mark key aspects of the image processing** that may have influenced the final 3D structure. *3D Map Result* indicates whether the resulting map displayed recognizable, protein-like features consistent with the target macromolecule. *Used mics* refers to the percentage of total micrographs deposited in EMPIAR that were included in the processing, with a maximum of 1,000 micrographs per entry. *Accepted data curation* indicates the percentage of micrographs that passed the quality filters from the initial processing set. *Final/Initial Parts.* refers to the number of particles retained in the final refinement compared to those initially selected from the curated set of micrographs, capped at 200,000 particles.

EMPIAR	Protein Type	Mol. Weight (kDa)	Used Mics (%)	Deposit/ FSC resolution (A)	3D Map Result	Accepted data curation (%)	Particle Diam. /Est. (A)	Final/ Initial Parts.	Features of Micrographs
10061	Beta-galactosidase	467.06	65	2.2 /3.1	High-quality	89.9	150/ 147	77K/ 92K	aggregated particles + sub optimal particle concentration
10406	Ribosome (70S)	632.89	36.8	2.7/3.1	High-quality	58.8	240/ 232	23K/ 74K	mono-dispersed particles + moderate protein edge texture
10576	Nuclear Protein (DNA)	290.21	50.4	2.9/3.2	High-quality	97.4	180/ 208	103K/ 200K	low contrast micrographs + difficult to recognize and pick particles
10737	Membrane Protein (E-coli)	155.83	29	2.2/4.1	High-quality	75.2	179/ 172	56K/ 177K	mono-dispersed particles + sufficient contrast
10184	Aldolase	NA	62	2.4/3.2	High-quality	47.5	100/ 90	111K/ 200K	mono-dispersed compact particles
11057	Hydrolase	149.43	11.9	2.8/3.1	High-quality	80.4	140/ 118	88K / 161K	difficult to identify particles + sub optimal particle concentration + ice issues
10389	Metal Binding Protein	1,042.17	23.1	2/4	Suboptimal	65.4	200/ 165	6K/ 30K	abundance of ice patches + dispersed protein particles + low number of particles per micrograph
10671	Signaling Protein	77.14	17	3.5/6.1	Suboptimal	96.6	110/ 86	90K/ 200K	extremely small protein particles + high density protein particles
10669	Proteasome (Plant Protein)	1,681.81	2.2	3.2/5.9	Suboptimal	95.3	500/ 229	27K/ 101K	carbon edges presence + disperse and distinct top, side and inclined views of particles
10387	Viral Protein (DNA)	185.87	49.5	2.8/3.7	Suboptimal	84.3	168/ 200	21K / 37K	highly aggregated protein particles + difficult to pick particles
10590	TRPV1 with DkTx and RTX	N/A	20.6	7.8/11.9	Suboptimal	94	236/ 250	29K / 181K	high contrast + mono dispersed particles + ice contaminations

10526	Ribosome (50S)	1085.81	90.7	2.8/(NA)	Failed	1.9	400	N/A	extremely high ice contamination + variation in ice thickness
10760	Membrane Protein	321.69	26	4.5/8.3	Failed	99.6	130/122	66K/200K	abundance of ice patches + mono-disperse distribution + particles with sufficient contrast

4.1.1 High-quality examples

For the high-quality examples, 6 out of the 25 different proteins were carefully selected to illustrate the workflow's ability to adapt to different experimental conditions, shapes, molecular sizes, biochemical compositions, and dataset-specific challenges of these EMPIAR entries. As an introductory case, we use the **β-galactosidase dataset** (EMPIAR-10061, first row of Table 2) to provide a detailed walkthrough of the entire image-processing pipeline. This example serves as a complete, step-by-step demonstration of how the automated workflow behaves under automated conditions, before presenting a more concise, result-oriented overview of the remaining five high-quality examples (second to sixth rows of Table 2).

The processing of EMPIAR-10061 is summarized in two complementary figures. Figure 51 presents the **pre-processing results**, including quality metrics and the particle-picking strategy, while Figure 52 shows the **processing results**, focusing on structural outcomes. Together, these figures illustrate how the workflow performs from the earliest quality control stages to the final 3D reconstruction. For the remaining high-quality datasets, a more general comparison between our reconstructions and the corresponding EMPIAR-deposited maps is shown in Figure 53.

The **β-galactosidase dataset** (EMPIAR-10061) is an excellent example for demonstrating the full capabilities of the automated workflow, as it combines several desirable properties that make it particularly informative for method evaluation. First, β-galactosidase is a well-established cryo-EM benchmark specimen: a moderately large, symmetric (tetrameric) enzyme with strong structural features that provide clear signal for motion correction, CTF estimation, particle picking, and refinement. Second, the dataset exhibits realistic imperfections, most notably particle aggregation and suboptimal, uneven particle concentration across micrographs. This means that some images contain densely packed particles while others contain very few, offering a valuable test of how the automated particle picker adapts to both extremes. Finally, the dataset offers a broad and favorable distribution of particle orientations, representative of good experimental conditions and conducive to reliable 2D and 3D convergence. Together, these characteristics make EMPIAR-10061 a compelling example for illustrating how the workflow's quality filters, scoring mechanisms, and pruning strategies operate in practice.

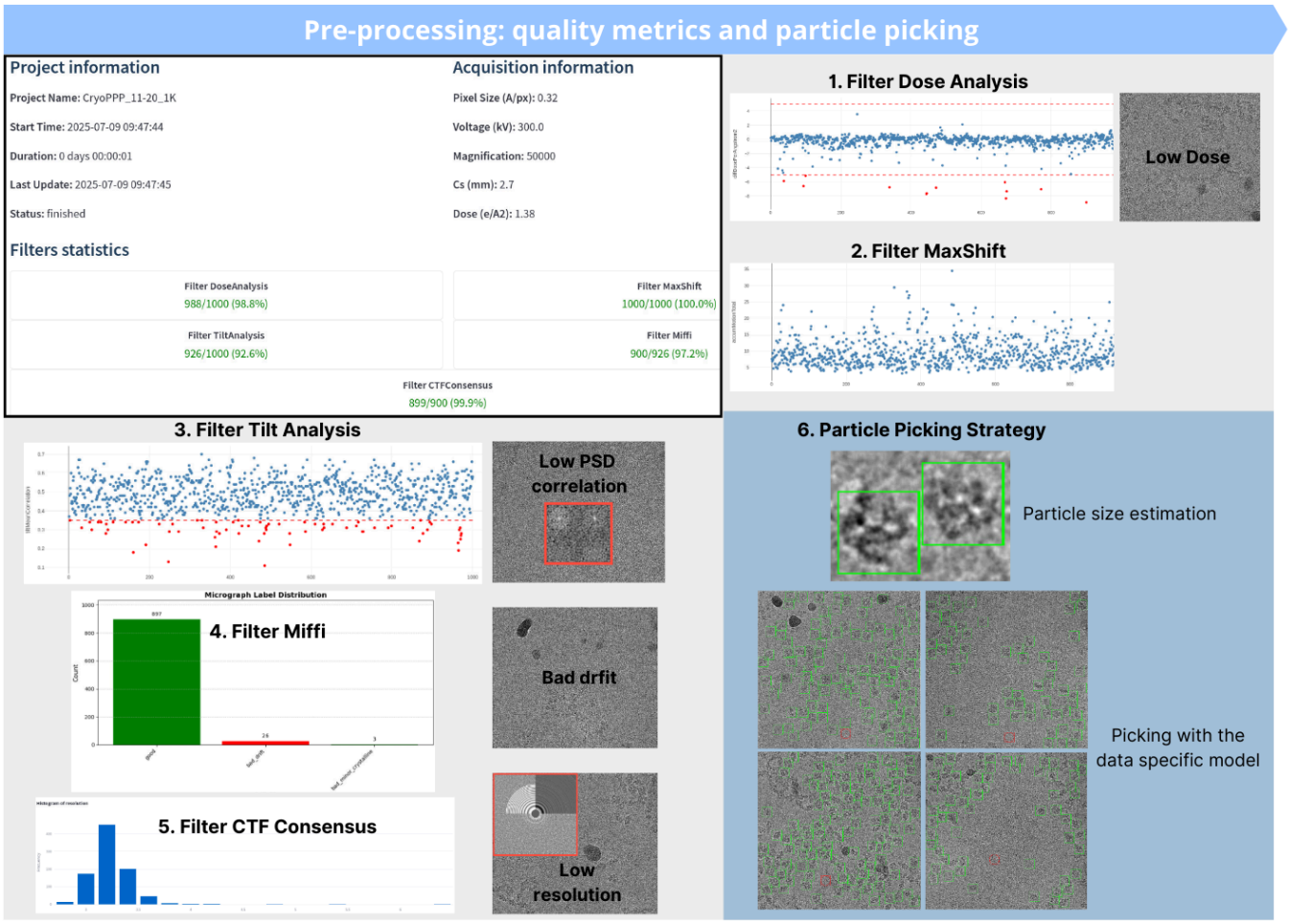


Figure 51. Pre-processing summary. The corner panel outlined in bold black provides a complete summary of the project, including acquisition parameters and the data-curation filters statistics. This panel is directly extracted from the Main View of the Dashboard. To the right and below it, the numbered items correspond to the cascade of quality filters, with each filter's diagnostic plot shown on the left and representative discarded micrographs on the right. Finally, the section highlighted in blue shows the Particle Picking Strategy, the last step of the pre-processing phase. The first image displays the estimated particle size, followed by four examples of particle picking across micrographs with very different particle concentrations. Data correspond to EMPIAR-10061.

In the preprocessing phase (Figure 51), and using the ***Quality Monitor Dashboard*** described in the Methods section, the pipeline identified several quality issues present in this dataset; however, none of them were considered critical. The **Dose** plot revealed a stable and uniform dose rate across the dataset. Although a small number of movies fell outside the ±5% deviation from the mean dose, these deviations were minor and did not indicate problematic dose variability or significant changes in ice thickness. Consistently, the **Maxshift filter** also indicated good data quality: the movie-alignment algorithm detected no strong drift, and consequently no micrographs were rejected based on this criterion.

The filter with the lowest acceptance rate was the **Tilt Analysis**, which retained 92.6% of the micrographs. Using the Dashboard's plotting tools, we examined the PSD Analysis mean-correlation values over time and identified several micrographs with unusually low correlation. These images exhibited reduced contrast, which explains their rejection by this criterion. For the **Miffi Filter**, the model flagged 26 out of 926 micrographs as affected by severe drift, indicating that these movies were not properly corrected by the movie-alignment algorithm, likely due to a systematic alignment failure. These were therefore discarded.

The **CTF filter** showed a very high acceptance rate. The interactive histogram in the Dashboard revealed that the vast majority of micrographs had high-quality CTF fits, with estimated resolution values in the 3–4 Å range. Only a single micrograph failed to meet the resolution threshold of <6 Å. Overall, only a small fraction of the dataset was rejected, with 101 out of 1000 micrographs discarded (approximately 10%).

Regarding the **Particle Picking Strategy**, the estimated particle diameter proved to be accurate, as confirmed both by Figure 51 and by expert annotations summarized in Table 2. The micrograph examples picked with the on-the-fly–trained picking model illustrate its high precision: in densely populated micrographs, individual particles are correctly identified without duplication, whereas in sparsely populated regions the picker effectively avoids false positives arising from background noise or ice features.

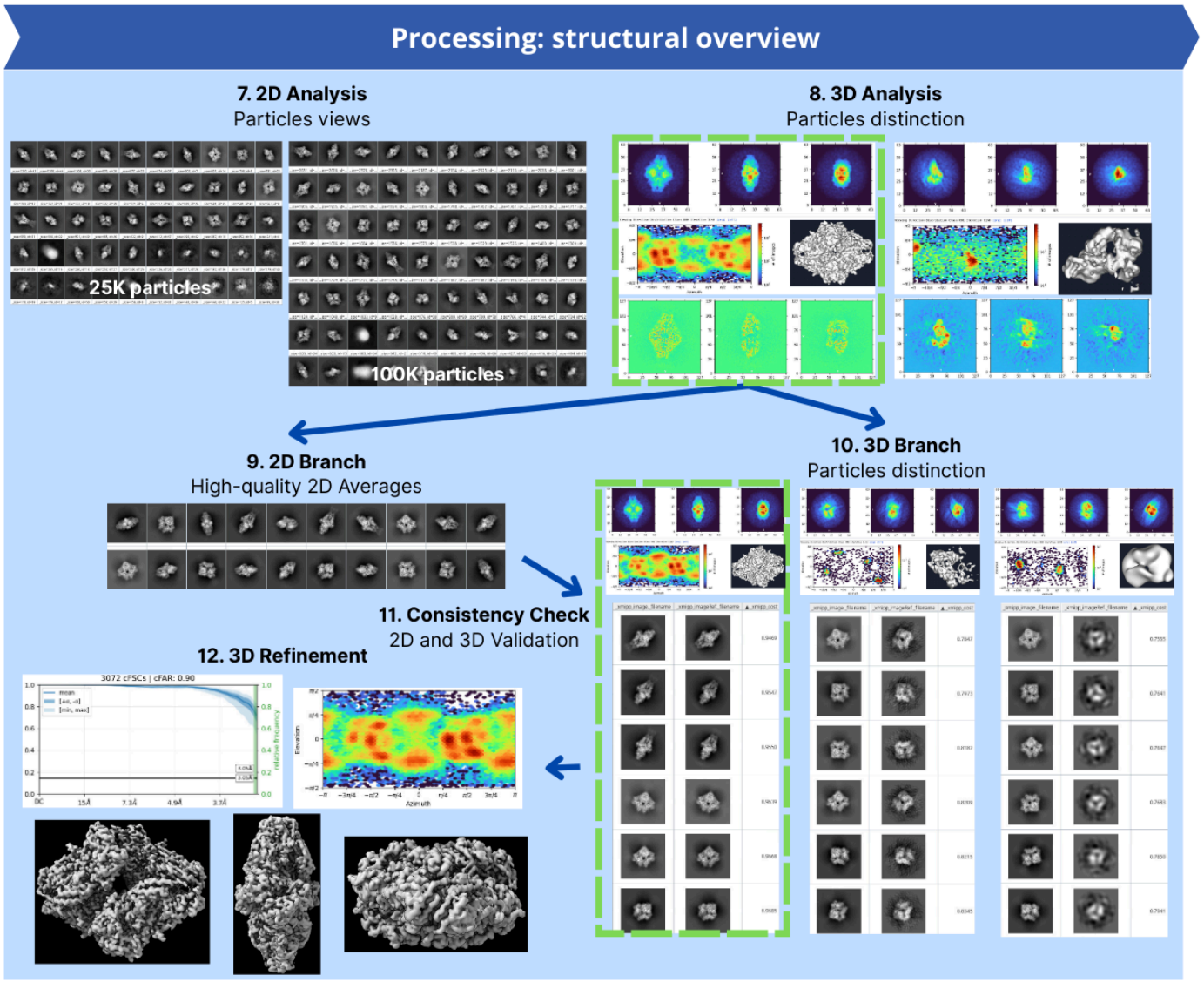


Figure 52. Processing summary: structural overview. The numbered titles correspond to each step of the image-processing pipeline, summarizing the key outputs and decisions made at every stage. Data correspond to EMPIAR-10061.

After achieving a low rejection rate during the quality-control phase, and obtaining a reliable particle-picking model, the initial 2D analysis revealed a broad distribution of particle views (Figure 52, Step 7). This is typically a positive indicator, suggesting the absence of strong preferred orientation. This interpretation was later confirmed by the 3D analysis performed in parallel (Figure 52, Step 8). In the 3D analysis, two ab initio models were generated. One model clearly captured the overall shape of the target protein (highlighted in green), while the other failed to converge to a meaningful structure. Selecting the volume with the largest particle population was therefore the appropriate decision. The angular distribution plot further

confirmed a well-sampled orientation space, supporting the conclusion that no preferred-orientation issue was present.

Using the selected particle set, high-quality 2D averages were obtained (Figure 52, Step 9). Visual inspection confirmed that no poor-quality class averages had been mistakenly selected in the 2D branch. In parallel, the 3D branch produced three ab initio volumes (Figure 52, Step 10). Two of these displayed protein-like features. Through the Consistency Check (Figure 52, Step 11), the workflow automatically identified the best 3D class, the one whose projections showed the highest correlation with the high-quality 2D averages.

Using this 3D volume and its associated particles, 3D refinement proceeded to yield an almost Nyquist-limited reconstruction (3.1 Å), given that particles were extracted at 1.5 Å per pixel. Resolution assessment using conical Fourier Shell Correlation (cFSC) displayed narrow standard-deviation bands around the mean curve, indicating isotropic resolution and uniform directional quality of the final reconstruction. Together, these results demonstrate that the automated pipeline consistently makes correct decisions at each step of the image-processing pipeline. It performs rigorous data-quality assessment and filtering, achieves robust particle picking, provides an immediate structural overview of the sample on-the-fly, and reliably converges to a high-resolution 3D structure without user intervention.

The next five examples offer a more general, results-oriented perspective, illustrating how the automated workflow successfully reconstructed 3D structures that closely match those deposited by the original authors of each EMPIAR dataset (Figure 53).

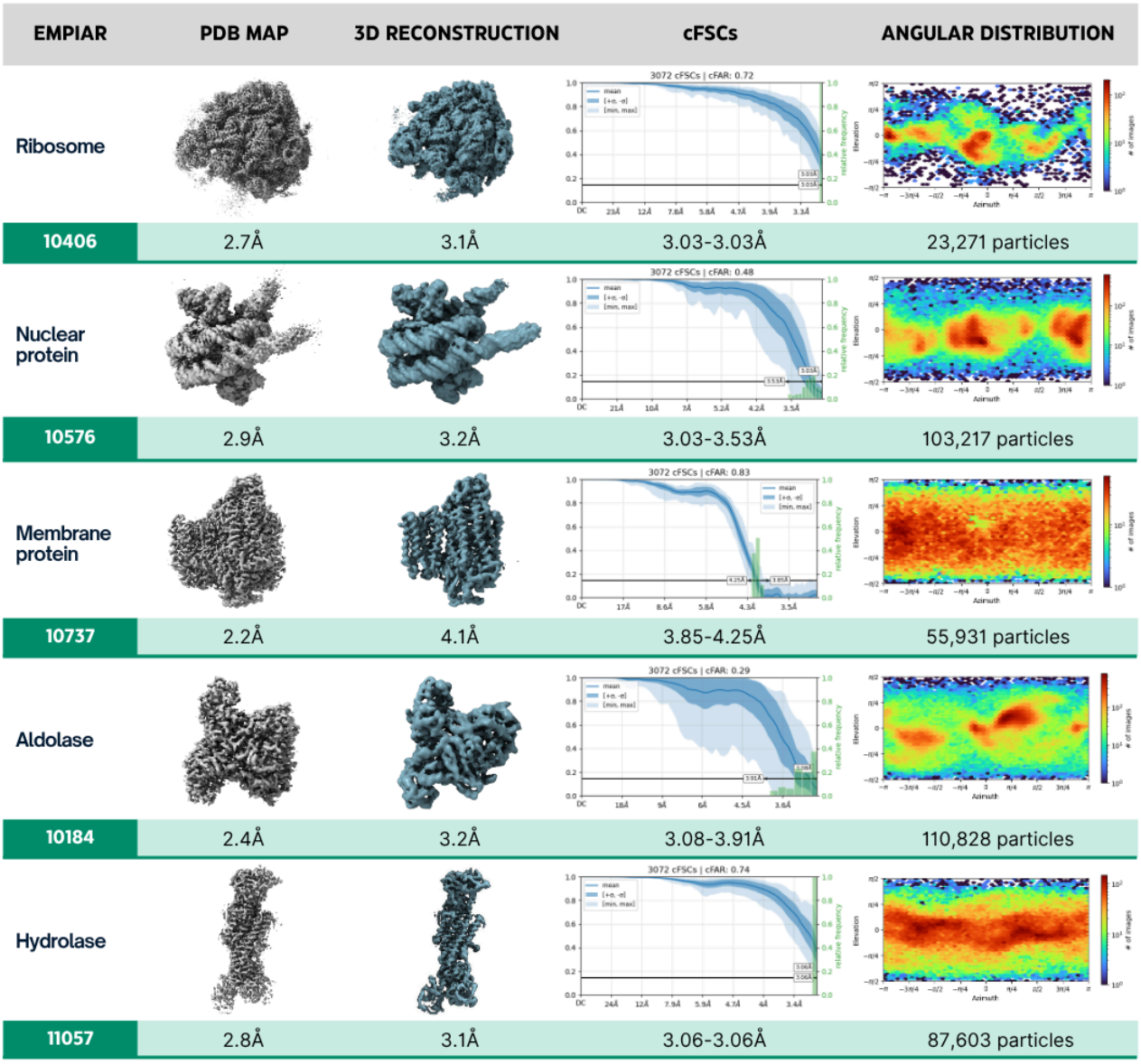


Figure 53. High-quality examples. Each entry corresponds to one of five distinct proteins with unique shapes and molecular weights. For each case, the deposited PDB volume, the reconstructed 3D map, and its associated quality metrics (cFSCs and angular distribution) are shown.

For the **70S ribosome dataset (EMPIAR-10406)**, our workflow was challenged with a very large ribonucleoprotein particle composed of ribosomal RNA and numerous protein subunits. This system is not only a widely used benchmark in cryo-EM but also a representative of highly heterogeneous and asymmetric macromolecular assemblies. Despite processing only 36.8% of the deposited data, the workflow successfully reconstructed the conformation corresponding to EMD-10809. It even detected the alternative state reported as EMD-10892 during the first round of 3D classification, which was then excluded from subsequent steps. This demonstrates that the pipeline can both handle large particles and resolve conformational heterogeneity

without requiring manual intervention. Importantly, automated curation retained only 58.8% of the micrographs, with most rejections flagged by Miffi as bad films, in agreement with the dataset's known issues of moderate protein edge texture. These results highlight the pipeline's ability to robustly process large, complex, and heterogeneous assemblies while also informing the user of data quality issues.

The **nucleosome dataset (EMPIAR-10576)** provided a distinct test case, consisting of a DNA–protein complex of moderate size with a characteristic elongated helical structure of DNA wrapped around histone proteins. Unlike compact globular proteins, nucleosomes often generate micrographs with uneven contrast, where the DNA end segments contribute weaker signals compared to the protein core, resulting in the ends of the DNA appearing blurred in 2D averages. Despite these particularities, and using only 50% of the deposited micrographs, the workflow achieved reliable particle picking, with 2D classes that clearly captured the DNA-wrapped contours that became more defined after refined 3D alignment. Data curation was highly effective, with 97.4% of the first 1,000 micrographs retained, confirming the overall quality of the dataset. This case demonstrates the pipeline's adaptability in handling nucleic acid–protein complexes, producing high-quality 2D classes and 3D reconstructions without requiring manual intervention.

The **cytochrome bo3 dataset in MSP nanodiscs (EMPIAR-10737)** represented a third distinct scenario, testing the pipeline on a membrane protein embedded in a lipid bilayer mimic. The lipid nanodisc introduces additional background that complicates particle picking, as empty nanodiscs or lipid belts can dominate early classifications. Even with only 29% of the deposited micrographs, the workflow effectively identified the protein. After the initial 3D classification, the nanodisc boundaries were evident in the classes. However, subsequent classification steps allowed the lipid background to be minimized, ensuring the reconstruction focused on the protein core. This resulted in the recovery of high-resolution features in the membrane protein itself. Automated data curation accepted 75.2% of the initial 1,000 micrographs, ensuring that only the best-quality data continued into processing. This example highlights how the pipeline addresses the unique challenges of membrane proteins, where empty lipid signals must be distinguished from protein density.

The **aldolase dataset (EMPIAR-10184)** served as an example of a small, symmetric, and well-behaved soluble enzyme. Its high particle count and intrinsic symmetry make it an ideal benchmark for testing throughput, efficiency, and resolution. In this case, the workflow processed 62% of the deposited data using C1 symmetry (non-symmetry), yet still managed to recover symmetric features and achieve an almost Nyquist-limited reconstruction, given that particles were extracted at 1.5 Å in all cases. Data curation retained only 47.5% of the first 1,000

micrographs, effectively discarding those with low resolution or empty fields, illustrating the ability of automated filtering to clean the dataset. This case demonstrates the efficiency and reliability of the pipeline under suboptimal conditions, showing that it can achieve near-maximum resolution automatically without manual supervision.

Finally, the **hydrolase dataset (EMPIAR-11057)** provided another contrasting test case, consisting of a gastric proton pump complexed with revaprazan, a relatively small membrane protein (149 kDa) solubilized in detergent with a non-symmetric, cylindrical shape. This sample was notoriously difficult to process due to their elongated geometry, small size, relatively low contrast and the presence of a micelle coating its transmembrane region, which complicates particle picking and alignment. Moreover, this particular dataset was labeled as suboptimal due to difficulties in particle identification, low particle concentration, and issues with ice contamination. Despite these challenges, and using only 12% of the deposited micrographs, the workflow was able to perform reliable particle picking. However, some false positives in empty micrographs were also detected in early 2D classes. Automated data curation was effective, with 80.4% of the first 1,000 micrographs accepted, showing that the system could still distinguish usable data under challenging conditions. This case highlights the robustness of the pipeline in handling weak-signal particles and irregularly elongated shapes.

4.1.2 Suboptimal cases

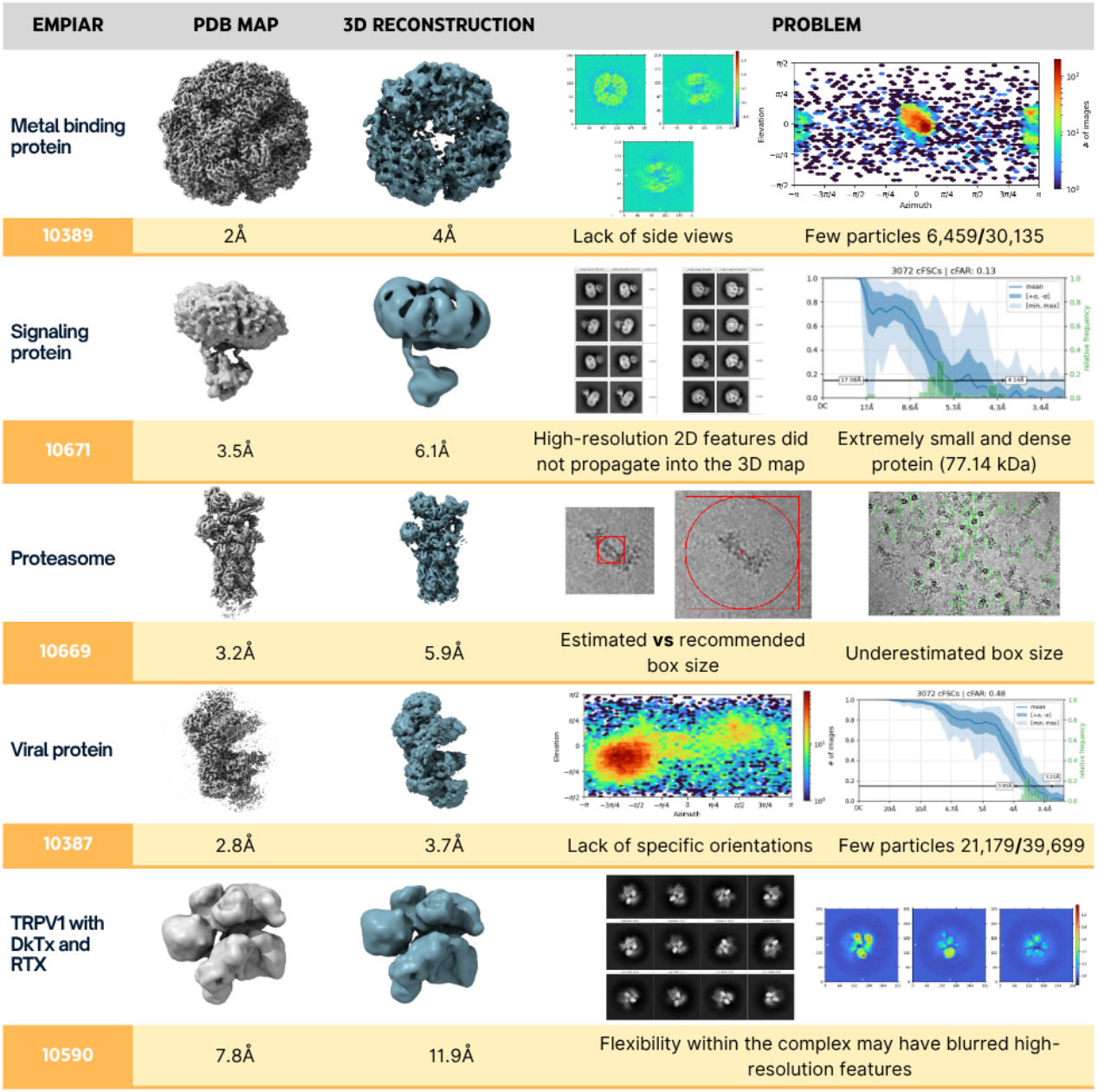


Figure 54. Suboptimal cases. Each entry corresponds to one of five datasets in which the final 3D reconstruction was suboptimal. Each entry displays the deposited PDB volume, the reconstructed 3D map, and the issues identified by the pipeline.

For the suboptimal cases, reviewing the image processing results helps us identify why the 3D reconstruction was not optimal (Figure 54). Nevertheless, it is essential to note that in all five cases the target protein was still visible in the reconstructions:

For the **metal-binding protein dataset (EMPIAR-10389)**, corresponding to urease from *Yersinia enterocolitica*, our workflow successfully identified the target protein; however, the final 3D reconstruction lacked side views. This limitation is likely related to the relatively small number of initial particles (30,135), the fact that we used only 23.1% of the deposited micrographs, and a strong pruning effect during automated data curation, which retained just 65.4% of the first 1,000 images. The dataset itself is annotated as being affected by abundant ice patches, dispersed particles, and low particle density per micrograph, and our results corroborated these issues. A potential improvement for future processing would be to increase the number of initial micrographs, thereby boosting particle counts and sampling missing orientations.

The **signaling protein dataset (EMPIAR-10671)**, corresponding to the CGRP receptor bound to peptide in detergent micelles, yielded a low-resolution 3D reconstruction. Although the protein was clearly visible and recognizable in 2D classes, high-resolution features failed to propagate into the 3D map. The nature of the dataset partly explains this: the target is a very small and compact 77 kDa protein embedded in detergent micelles, which obscures structural detail. While we respected our 200,000-particle threshold, typically sufficient to reach high resolution, the presence of the micelle has limited resolution. In this case, additional manual intervention, such as masking the micelle or carefully filtering particle sets during 2D classification, could improve outcomes.

The **proteasome dataset (EMPIAR-10669)**, corresponding to the substrate-engaged 26S proteasome, highlighted the challenges of working with elongated complexes. The reconstruction problem stemmed from an underestimation of the box size. Due to its very long cylindrical shape, side views extended beyond the extraction box and were truncated in the final 3D map. While the protein was clearly recognizable, the whole complex was not reconstructed. This issue emphasizes the need for box-size adaptation when processing very elongated particles; increasing the box dimensions would likely recover the missing density and enable a more complete reconstruction.

For the **viral protein dataset (EMPIAR-10387)**, which investigates HIV intrasomes in complex with strand-transfer inhibitors, the reconstructed map appeared noisy with unresolved regions. Only 39,699 particles were initially picked, and processing relied on just 50% of the deposited data. Moreover, the dataset itself is annotated as containing highly aggregated particles that are difficult to pick out, a challenge that our results confirmed. Although the target complex was detectable, the particle scarcity limited reconstruction quality. A straightforward solution would be to include a larger portion of the dataset to increase the number of particles.

Finally, the **TRPV1 dataset (EMPIAR-10590)**, which contains the ion channel in complex with DkTx and RTX, yielded a low-resolution 3D reconstruction consistent with the deposited resolution of 7.8 Å. The overall shape matched the reference, but many finer details remained unresolved. The relatively large particle set (181,267 initial particles) suggests that insufficient sampling was not the main issue. Instead, flexibility within the complex may have blurred high-resolution features: in the 2D classes, some regions appeared well defined while others were noticeably diffuse. In such cases, advanced flexibility analysis or targeted manual refinement would be necessary to overcome these limitations.

4.1.3 Failed cases

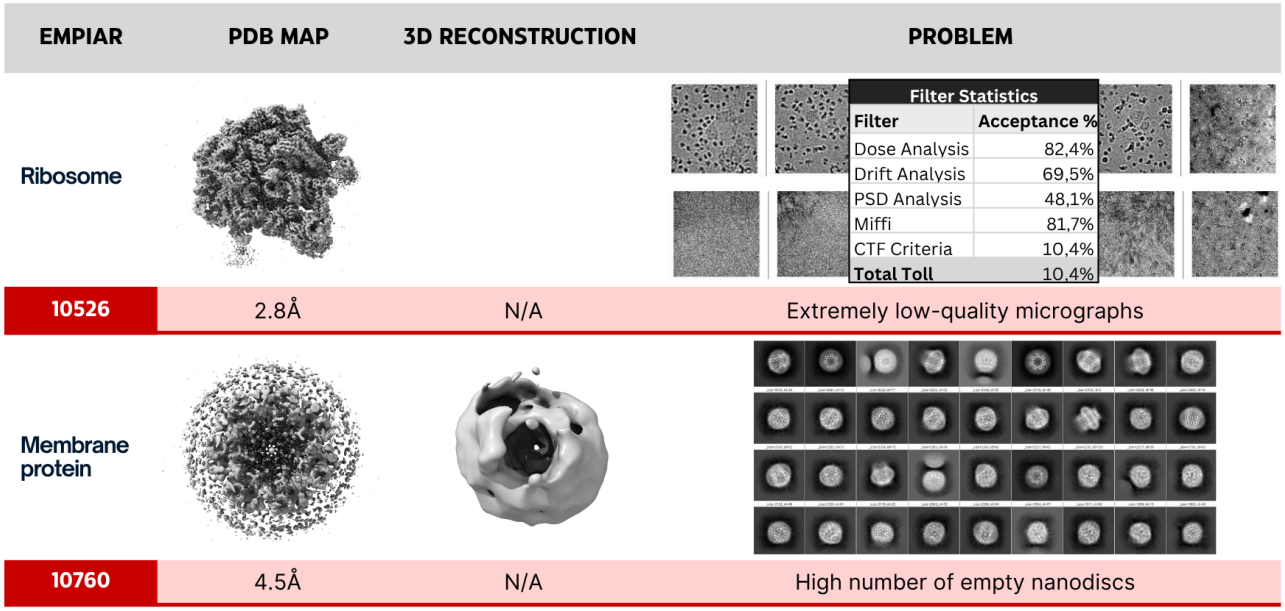


Figure 55. Failed cases. Each entry corresponds to one of two datasets that did not yield a valid 3D structure. Each entry displays the deposited PDB volume, the attempted reconstruction, and the problems identified by the pipeline.

For the entries that did not yield a structure (2/32), it is necessary to determine whether the limitation was due to our image processing pipeline or inherent to the dataset's complexity (Figure 55):

For the **50S ribosome dataset (EMPIAR-10526)**, the issue was not that the data were unreconstructable, but that our workflow's quality filters correctly discarded the majority of micrographs due to extremely high ice contamination and variations in thickness, as annotated in Table 2. Only 1.9% of the images (19/1,000) passed the automated image curation, which would have prompted, in a real experiment, the termination of data collection and a return to sample optimization. Nevertheless, due to their size, ribosomes are often used as benchmark

targets in cryo-EM, such that with limited manual intervention, it is possible to reconstruct a 3D structure even from such suboptimal data. Indeed, this explains why the dataset was deposited in EMPIAR. This case thus highlights how automated rejection can serve as an early-warning mechanism during acquisition, helping to save microscope time by flagging when conditions are too poor for efficient processing.

The **pannexin-1 dataset in nanodiscs (EMPIAR-10760)** posed a different challenge. Although particle picking was excellent, reaching our maximum of 200,000 initial particles, the dataset was heavily dominated by empty nanodiscs. As a result, most 2D and 3D classes converged to lipid-only shapes, with only a minority of 2D classes showing traces of high-resolution protein features. The inability of the protein signal to dominate the 3D classification prevented reconstruction of the intended structure. Without manual selection to isolate these minority classes, any processing would not yield a viable reconstruction. This case highlights the biochemical and sample-preparation challenges of membrane proteins, where empty nanodiscs can overwhelm image processing.

4.2 Real-life deployment

The workflow has been successfully implemented and validated as the standard on-the-fly processing solution at the European Synchrotron (ESRF) Cryo-EM beamline, CM01 [99], and the newly commissioned French Collaborative Research Group cryo-EM beamline, CM02. This work was carried out during a three-month **international research stay** as part of the PhD program (Figure 56). The CM01 user program accommodates a wide range of samples with varying sizes and shapes. The cryo-EM microscopes operate at high throughput, averaging 600-700 images per hour and supporting approximately 100 experiments per year. Each 48-hour experiment may include up to four different samples, thereby broadening the diversity of protein specimens that can be assessed using our protocol.

Figure 56. CM01 CryoEM Facility at the ESRF, France.

4.2.1 Network setup and IT infrastructure for on-the-fly processing

Scipion was installed in a shared folder on the ESRF CernVM file system, together with all external software required by the workflow, provided as Scipion plugins. Data is transferred in real time from the K3 camera computer to the ESRF data center over a 25 Gb fibre-optic network connection and accessed in real time, such that a Scipion project was automatically created as soon as acquisition began. Both raw and processed data are stored on GPFS shared filesystem. This setup ensured efficient data transfer between storage systems with minimum latency and enabled processing to run in parallel with data collection (Figure 57).

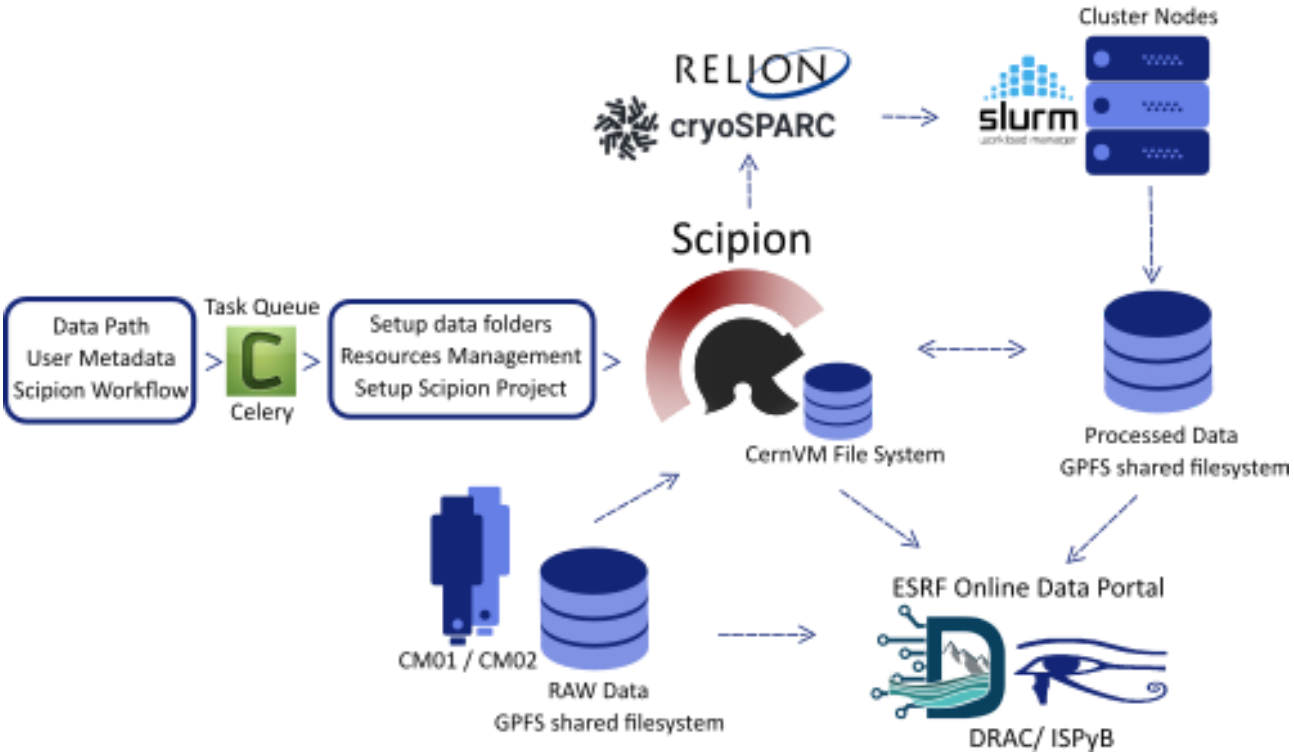


Figure 57. Overview of the Scipion on-the-fly workflow deployment at the ESRF. Scipion and all associated plugins are installed on the ESRF CernVM File System (CVMFS), while raw and processed data are stored on the GPFS shared filesystem. Users initiate Scipion workflows by specifying the data path, relevant metadata, and the desired processing template. A Celery task is then launched to handle the preprocessing steps, which include setting up the data directories, allocating computational resources, and injecting metadata into the corresponding Scipion protocols. Scipion subsequently orchestrates downstream protocols (e.g., CryoSPARC, Relion), executed on dedicated SLURM cluster partitions. A dedicated Scipion protocol automatically transfers acquisition and processing metadata to the ISPyB/Drac online platform, enabling real-time monitoring of experiment progress and facilitating user access and data retrieval.

4.2.2 Launching the on-the-fly processing

To enable unattended processing, the ESRF data management system was extended to load and launch workflows as soon as acquisitions began automatically. These developments, maintained in a version-controlled GitLab repository [100], introduced a **template-based strategy** that replaced the earlier protocol-by-protocol workflow construction. This new approach significantly improved maintainability and flexibility by:

A. **Reducing code size and complexity:** Workflow generation was reduced from ~1,800 lines of code to fewer than 250 lines.
B. **Improving modularity:** pipelines for different use cases (*e.g.,* academic vs. industrial, tilted vs. non-tilted samples) are dynamically loaded from templates, eliminating code duplication.
C. **Minimizing code dependency:** templates can be edited, validated, and approved without requiring modifications to the central workflow launching system.

A variety of workflow templates are available, ranging from **minimal preprocessing** (*i.e.,* motion correction and CTF estimation) to **comprehensive pipelines** that include 2D classification and preliminary 3D reconstructions. Templates also accommodate tilted datasets by adjusting motion correction filters, CTF criteria, and defocus sampling strategies. Both *CryoSPARC* (default for academic projects) and *Relion* (used for industrial clients) are supported as processing backends. Templates are stored as *JSON* files within the *scipion-em-esrf* plugin and version-controlled through GitLab [100], ensuring traceability and approval before deployment.

For cryo-EM facility services, the workflows are fully integrated with ISPyB for SPA and ICAT for cryo-ET through dedicated Scipion protocols embedded within the processing pipelines, ensuring that results are automatically reported and made accessible to users. All jobs are submitted via the ESRF SLURM scheduling system, guaranteeing efficient and fair use of shared computational resources. Up to three workflows can run concurrently, with CPU and GPU requirements dynamically estimated prior to launch to optimize hardware utilization across the shared cluster. Users can monitor both data-collection progress and processing quality in real time through the ISPyB or ICAT web interfaces. Raw and processed datasets remain available through the ESRF Data Portal for three months following acquisition and are subsequently archived for long-term preservation for a period of five years.

4.2.3 Workflow results

The deployment of the workflow in user experiments at ESRF's CM01 beamline [99] yielded informative results and robust statistics. Although each dataset was limited to 200,000 particles due to computational constraints, the outcomes fully met the intended performance objectives. The analyses revealed three main categories of reconstruction outcomes: **high-quality**, in which the data converged to high-resolution structures of the target protein; **suboptimal**, where convergence was achieved but the reconstructions exhibited heterogeneity or preferred orientations that limited the attainable resolution; and **failed**, where no meaningful structure could be recovered and only low-resolution, non-specific density was obtained.

Of the 34 experiments analyzed (Table 3), more than 70% of the datasets converged to interpretable 3D structures, with approximately half of these achieving high-quality reconstructions at resolutions in the 3–4 Å range (with particles extracted at 1.5 Å/px). A smaller fraction of datasets fell into the failed category (26.5%).

This slightly lower success rate compared with the benchmark CryoPPP tests is expected. Benchmark datasets are typically curated EMPIAR entries accompanied by deposited 3D structures in the PDB (Protein Data Bank), meaning that a dataset is only deposited when the underlying experiment produced a successful reconstruction. In other words, benchmark data reflect ideal or near-ideal conditions: high-quality purified samples, good grids and vitrification, stable imaging conditions, and favorable particle-orientation distributions—all factors that strongly influence the ability to recover a high-resolution 3D structure.

By contrast, even though CM01 enforces strict standards for high-resolution data collection, these conditions cannot guarantee a successful reconstruction, not even when processing is performed manually by experts. The variability observed in our real-life experiments therefore reflects the inherent challenges of routine cryo-EM data-collection samples, where outcomes depend strongly on sample quality, preparation, vitrification, and imaging conditions. Importantly, the suboptimal datasets proved particularly informative, as they revealed issues such as sample heterogeneity and strong preferred orientations. Likewise, the failed cases were valuable for anticipating and diagnosing problematic samples early in the experiment, providing users with immediate feedback about potential processing difficulties.

Table 3. General overview of ESRF real-life results. Green indicates successful and robust 3D reconstructions; yellow denotes suboptimal reconstructions; and red corresponds to non-conclusive 3D reconstructions.

3D Final Reconstruction Assessment	Total Percentage of the 34 experiments (%)	FSC range resolution (Å)
High-quality	35.3	3 - 4
Suboptimal	38.2	4 - 8
Failed	26.5	6 - 10

While Table 3 summarizes the overall distribution of reconstruction outcomes for all experiments, the practical value of the workflow becomes particularly evident when examining representative real-life cases that achieved high-quality results. To illustrate this, four representative datasets from the high-quality category (4 out of 12) were selected (Figure 58). In all four cases, users granted permission to publish the corresponding 3D reconstructions and confirmed that the target protein was successfully identified and resolved at high resolution in a fully automated, on-the-fly manner during data collection. For confidentiality reasons, the protein identities and associated scientific context cannot be disclosed. All four reconstructions reached the maximum achievable resolution (Nyquist ≈ 3Å), consistent with the image-processing constraint of re-extracting particles at a maximum pixel size of 1.5 Å/px. These representative examples are shown in Figure 58.

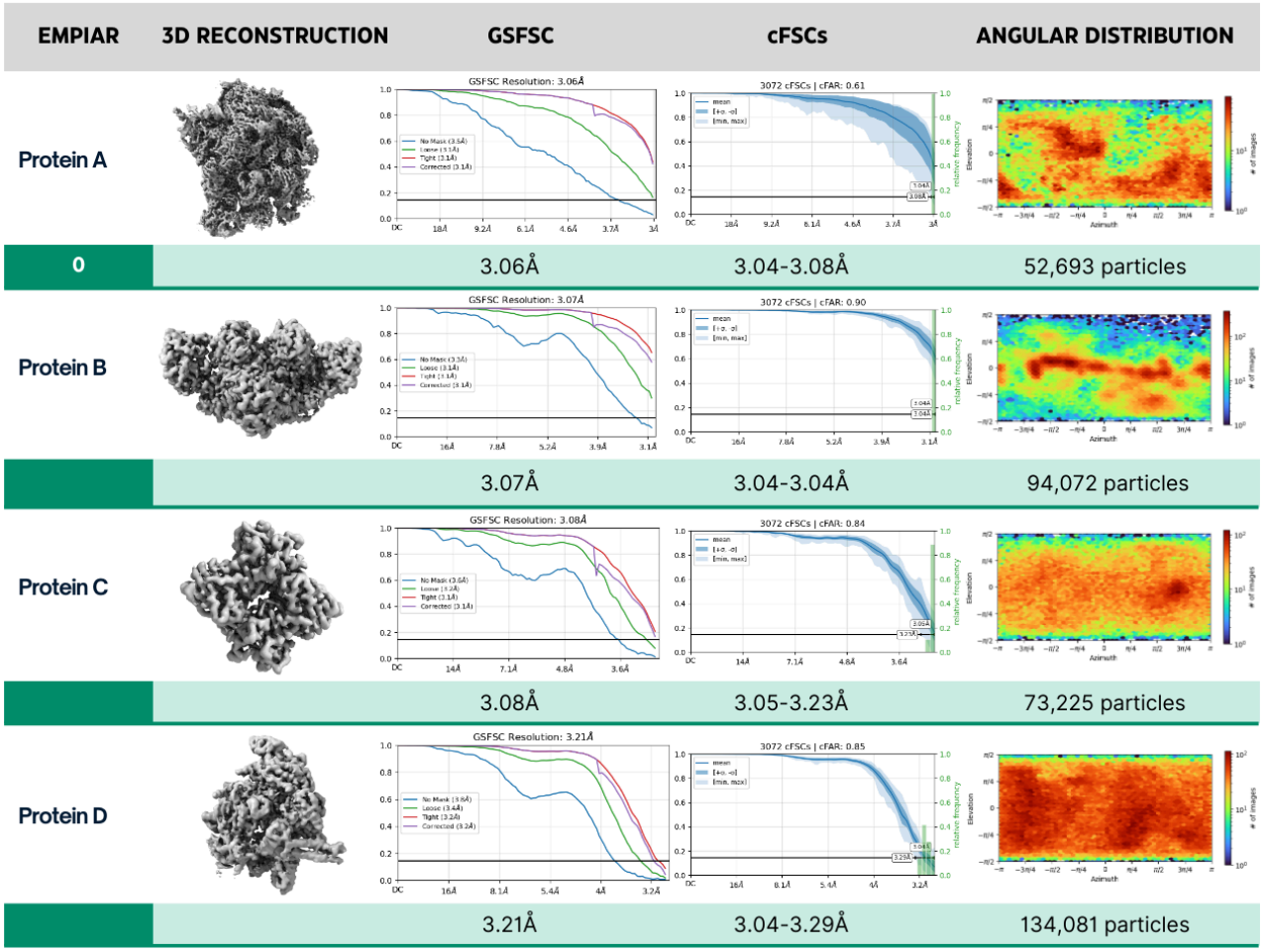


Figure 58. Overview of real-case examples from the CM01 CryoEM Facility.

Protein A displayed regions with very high-resolution features, in some cases potentially suitable for direct model building from the 3D reconstruction. The **GSFSC** tight and corrected curves were almost identical, confirming an accurate and reliable global resolution estimate. Although both curves exhibited a sharp decay approaching Nyquist, neither crossed the 0.143 threshold, indicating that Nyquist was reached. Re-extraction at a smaller pixel size could likely push the resolution further, supporting the interpretation that this is a high-quality, well-refined reconstruction with minimal overfitting. The **cFSCs** allowed us to evaluate the directional quality of the reconstruction by computing FSCs within localized cones in Fourier space. While the standard deviation band was not very wide, the minimum curve was noticeably separated from the mean, indicating that certain regions did not capture the same high-resolution detail. This observation is consistent with the **particle orientation distribution**, which shows a non-uniform angular sampling with certain orientations more populated than others. Some localized loss of resolution is therefore expected. Importantly, the distribution does not indicate a preferred-orientation problem.

Protein B showed a robust 3D reconstruction, with GSFSC tight and corrected curves again presenting a sharp decay at Nyquist without crossing the resolution threshold. This suggests that re-extraction at full-size pixels could further improve the final resolution. The cFSCs displayed an ideal behavior, with narrow standard-deviation bands and min/max curves remaining close to the mean, indicating strong directional isotropy. The particle orientation distribution revealed broadly sampled views, with some regions slightly overrepresented, but without negatively impacting the final reconstruction.

Protein C exhibited a well-resolved 3D reconstruction; however, in this case the GSFSC tight and corrected curves showed a sharp decay before reaching Nyquist, crossing the resolution threshold just prior to 3 Å. The cFSC analysis displayed a correct behavior, with narrow standard-deviation bands and minimum and maximum curves closely tracking the mean, indicative of strong directional isotropy. This observation was further supported by the particle orientation distribution plot, which revealed an uniform angular sampling across orientations.

Protein D displayed high-resolution features consistently across the entire 3D map. The GSFSC tight and corrected curves showed a sharp decay just before Nyquist, crossing the resolution threshold at 3.21 Å. The cFSC analysis revealed minimal directional variability, with narrow standard-deviation bands and minimum and maximum curves remaining very close to the mean. This result was further supported by the particle orientation distribution plot, which showed a uniform angular sampling across orientations. A notable feature of this dataset is the large number of particles retained for the final refinement: approximately 134,000 particles out of the initial 200,000. This high retention rate highlights the accuracy of the automated particle-picking strategy, where the majority of selected particles were of sufficient quality to pass subsequent filtering steps and contribute meaningfully to the final high-resolution reconstruction.

Overall, these results demonstrate precise and robust automated image processing, confirming the workflow's ability to deliver high-resolution structures on-the-fly without manual intervention. The data used for these examples were generously shared by ESRF users worldwide. The contributing groups are:

- **Protein A:** Institute for Biochemistry and Molecular Biology, Signalling Research Centres BIOSS (Biological Signalling Studies) and CIBSS (Centre for Integrative Biological Signalling Studies) at the University of Freiburg, Germany. The study is conducted by Prof. Dr. Carola Hunte, Dr. rer. nat. Wei-Chun Kao, and Staff Scientist Dr. Christophe Wirth.

- **Protein B:** Department of Microbiology, Infection and Immunity at the Institut de Biologie Structurale (IBS), France. The study is conducted by Henri Gröger (PhD candidate) and Prof. Wim Burmeister.
- **Protein C:** Instituto de Química Física Blas Cabrera (CSIC, Spain). The study is conducted by Postdoctoral Senior Scientist Dr. Martín Alcorlo Pagés.
- **Protein D**: Structure of Macromolecular Assemblies group (PI: Dr. Carlos Fernández-Tornero) at the Centro de Investigaciones Biológicas Margarita Salas (CSIC, Spain). The study is conducted by Alicia Santos (PhD candidate).

4.2.4 Operational times

Figure 59 illustrates the progression of image processing over time during a live acquisition. Based on experimental observations, the Titan Krios microscope at the ESRF operates at a pace of approximately 6 seconds per acquired movie. Our preprocessing steps (alignment, CTF estimation, and particle picking) have been shown to keep up with — and even surpass — this pace, achieving processing times as fast as 2–3 seconds per movie. Thus, the limiting factor lies on the microscope side.

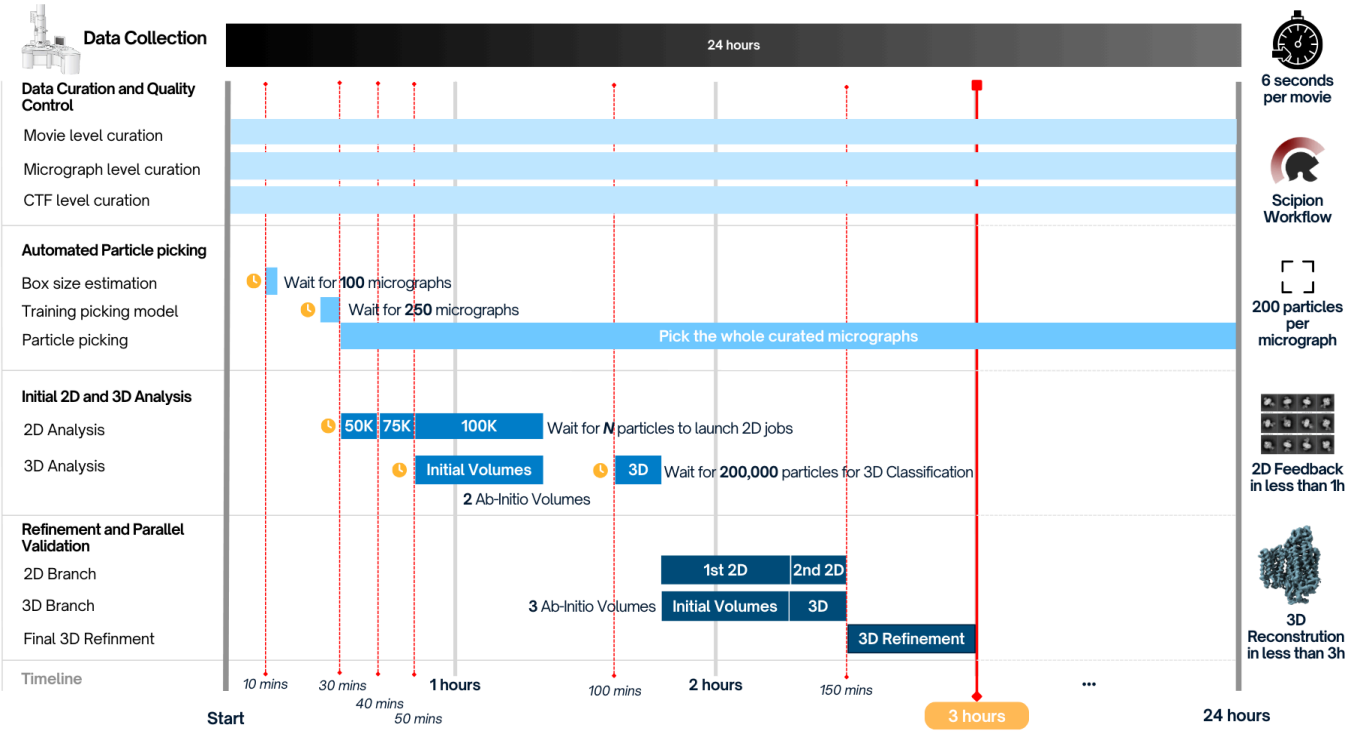


Figure 59. Gantt diagram of the Unattended Image Processing Pipeline at the ESRF. The left panel shows the four main stages of the image processing pipeline, while the right panel illustrates the timeline once acquisition has started. For this example, we assume, based on experimental cases, a microscope acquisition rate of up to 6 seconds per movie and an average of 200 particles per micrograph.

Under the following operational assumptions: (i) a limiting acquisition rate of 6 seconds per movie, (ii) an average of 200 particles per micrograph, which is a reasonable intermediate value, and (iii) a high acceptance rate during data curation, indicating good acquisition quality. We were able to estimate predictive times for obtaining structural feedback during on-the-fly processing.

Within the first 30–40 minutes, initial particle features and orientations can be assessed through 2D class averages, which are continuously updated in cumulative batches of 25,000 particles. The first ab-initio volumes can typically be obtained within the first 1.5 hours of acquisition, already providing insight into symmetry and angular distribution. At this stage, we must wait until approximately 200,000 particles have been accumulated to launch 3D classification with refinement of the two ab-initio volumes.

Given the assumptions above (200 particles per micrograph and a 6-second acquisition pace), the pipeline reaches the 200K particles subset before the 2 hour mark, allowing us to filter down with 3D classification protein-like particles for subsequent processing. Entering the final processing phase, before 2.5 hours of acquisition, we are able to perform both 2D and 3D validation, comparing three refined ab-initio 3D classes against high-quality 2D averages. After selecting the best-correlated volume and particle subset, the final 3D refinement is expected to complete before 3 hours of acquisition under these operational conditions.

Throughout this entire period, a continuous assessment of acquisition quality is performed from the beginning via the data-curation stage, providing both diagnostic and structural feedback as the session progresses.

The time benchmarking was performed on a Ubuntu 24.04 Linux workstation equipped with 64 CPU cores (2×AMD EPYC 9354, 3.25 GHz), 1.48 TB RAM, and 8 NVIDIA A40 GPUs (Driver Version 550.163.01, CUDA Version 12.4).

CHAPTER 5 – DISCUSSION

In this work, we have presented a fully unattended pipeline for on-the-fly cryo-EM data processing, designed as a robust diagnostic tool for high-throughput facility environments. Its main contribution lies not merely in automating sequential steps, but in creating an intelligent workflow that actively assesses data quality, adapts to the specimen, and provides a comprehensive, curated, analysis-ready output that significantly accelerates the path to a final structure. This pipeline and the associated diagnostic tools were developed, validated, and deployed during a three-month International PhD Stay at the ESRF Cryo-EM Facility, where close interaction with real user projects informed much of its design.

A key advantage of the pipeline is the value of its comprehensive output. Unlike workflows that conclude after pre-processing, ours delivers a complete suite of actionable results, including a data-specific picking model, curated micrographs and particle sets, high-quality 2D classes, and a preliminary 3D map. This provides researchers with a substantial head start, eliminating many of the repetitive and time-consuming early stages and enabling immediate focus on high-resolution refinement and heterogeneity analysis. Moreover, the workflows can be applied not only to data acquired in real time but also to pre-existing or transferring datasets. Leveraging streaming execution, the pipeline processes each new input as soon as it becomes available rather than waiting for the entire dataset to finish before advancing to the next step, thereby saving valuable time. In this way, it provides an effective and reproducible entry point for any reconstruction project from which users can seamlessly continue toward detailed structural refinement.

The extensive benchmarking on the diverse CryoPPP dataset demonstrates the pipeline's robustness and adaptability. With a high overall success rate (94% successful processing, with 78% yielding featured 3D reconstructions), the workflow proved effective across a wide range of challenging targets, including large ribonucleoprotein assemblies, small enzymes, flexible or elongated particles, and membrane proteins. This performance validates the core design principles of our approach. Key to this success are the multi-stage quality filters, the consensus-based training of a dedicated particle-picking model, and the integrated parallel 2D/3D validation steps. Furthermore, the analysis of "suboptimal" and "failed" cases highlights the pipeline's diagnostic utility: these outcomes did not reflect failures of the image processing logic, but rather successes in identifying inherent sample or data limitations, such as preferred orientation, extreme contamination, or a dominant population of empty nanodiscs. In a real-world scenario, this rapid feedback is invaluable, allowing users to recognize problematic

acquisitions early, adjust data collection strategies, or terminate unproductive experiments, ultimately saving days of valuable microscope time.

To complement the automated processing, we developed a centralized, user-friendly quality-monitoring tool that consolidates processing metadata across high-throughput acquisitions. Instead of requiring users to navigate the numerous specialized viewers inside *Scipion*, the monitor provides a single, coherent interface where key metrics (from movies, micrographs, and CTF estimates) are aggregated and presented through intuitive plots. This centralization dramatically reduces cognitive load during acquisition, makes quality issues immediately visible, and shortens the feedback loop between data collection and decision-making. For example, if a micrograph is rejected by the Max-Shift filter, users can instantly inspect the corresponding diagnostic plots to understand the underlying drift behaviour. These detailed visualizations ensure that data rejections are not opaque events but become meaningful, actionable insights.

Moreover, the dashboard is read-only, ensuring that processing results remain unaltered, and it is fully responsive across devices. Its design allows deployment either locally or via a secure facility endpoint for remote monitoring. The visual tools emphasize transparency in the curation logic and help operators quickly diagnose issues such as excessive drift, poor CTF fits, or anomalous dose behaviour. Beyond its immediate operational utility, the monitor also provides a structured, model-oriented data representation that facilitates downstream analysis and future automation. The same consolidated metadata can be reused for threshold tuning, retrospective performance studies, or even training supervised models for automated decision support. Finally, its modular architecture enables straightforward integration of new quality metrics or filter views without disrupting existing workflows, making the system well suited to evolving facility requirements and to the pipeline improvements.

Deployment at the ESRF's CM01 beamline further confirmed the pipeline's robustness under operational conditions. Approximately 70% of the 34 user datasets converged to interpretable structures, with nearly half achieving high-resolution reconstructions (3–4 Å). This slightly lower success rate compared with benchmark CryoPPP tests is expected, as facility experiments often involve new or unstable samples whose biochemical or biophysical properties remain uncertain. Unlike EMPIAR datasets, which represent well-characterized and already successful projects. In this context, the pipeline's ability to rapidly identify promising samples from unproductive ones and produce preliminary structures during acquisition is particularly valuable, enabling researchers to confirm that their target protein has been successfully captured while also detecting potential issues such as heterogeneity, dissociation, or preferred orientation within hours.

An equally important aspect of the workflow is its proven real-time performance. The fact that the pipeline's pre-processing speed surpasses the data acquisition rate of a Titan Krios ensures that analysis never falls behind the microscope and provides truly real-time feedback. The ability to generate a preliminary 3D reconstruction in under three hours, often before the data collection session has even finished, fundamentally changes the experimental paradigm. It allows operators to move from passive data collection to active, informed decision-making, such as implementing tilted data collection [39] when preferred orientation is detected. Furthermore, working directly with ESRF beamline scientists accelerated refinement of the workflow templates, queue configurations, and streaming logic, ensuring that the system could operate reliably under demanding conditions of high-throughput user sessions. The stay also provided a unique environment for transferring knowledge to facility staff, coordinating development priorities with the Scipion team, and establishing the foundations for long-term collaboration.

A central requirement for deploying a fully automated 3D workflow is robust integration with high-performance computing (HPC) queue systems. Without these adaptations, throughput would be fundamentally constrained by available GPUs. Queue systems such as SLURM or LSF provide protected, fair access to shared resources, enabling the workflow to scale from a single laboratory workstation to multi-GPU clusters. Within Scipion, this integration is not merely a convenience but a prerequisite for reliable automation: different stages of the workflow exhibit heterogeneous computational demands, and only an explicit queue-aware execution model can orchestrate them without user intervention. A major technical contribution of this thesis is the development of a granular per-action job submission layer that extends Scipion's queue capabilities to both large HPC infrastructures and constrained fat-node environments. By submitting each image-processing action as an independent job, GPU usage becomes dynamic rather than statically allocated, eliminating idle GPU time, increasing throughput, and enabling concurrent multi-protocol execution in the same GPU. This represents a substantial advance making the workflow adaptable to a wide range of facility setups.

To ensure transparency, reproducibility, and long-term maintainability, all workflow templates developed in this thesis have been deposited in WorkflowHub (public workflow registry) under the entry *CryoEM Facility Workflows* [97]. This public registration guarantees versioning, metadata standardization, documentation and compliance with FAIR principles (Findable, Accessible, Interoperable, Reusable). It also ensures that the workflows can be cited, reused, extended, and audited by the community, forming a stable foundation for future developments in automated cryo-EM image processing.

Beyond the technical implementation, this work also contributed to community training and dissemination. The methodological advances and software developments achieved during this period were integrated into courses, hands-on tutorials, and training materials presented in the *Scipion for Facilities* events. These activities reinforced adoption across multiple international facilities and ensured that the innovations in workflow design and HPC adaptation were effectively communicated to the broader cryo-EM community.

While the pipeline represents a major advance, there is still room for future enhancements. Benchmark analyses revealed challenges with highly elongated particles, where automated box size estimation was suboptimal, and with datasets dominated by contaminants, where classification struggled to isolate the minority protein-containing class. Future improvements could include developing more sophisticated algorithms to address these specific challenges, such as implementing dynamic box size adjustments for non-globular particles or integrating specialized processing steps to sort particles from common contaminants, such as empty nanodiscs or micelles. Furthermore, the current workflow relies on a static analysis of a fixed particle set. A significant improvement would be to transition to a dynamic, streaming-based model where 2D and 3D analyses are continuously updated as new particles are acquired, offering an even greater level of real-time feedback.

The implementation of this pipeline within an open-source framework such as *Scipion* provides a strong foundation for the suggested improvements and future expansions. *Scipion's* modular design allows seamless integration of new algorithms and tools as they emerge, ensuring the workflow can adapt to evolving challenges in cryo-EM. Its open-source nature guarantees transparency and reproducibility, as every processing step can be inspected, validated, and shared. Moreover, the framework enforces traceability, with parameter choices and processing decisions automatically recorded, a crucial requirement for both scientific rigor and compliance with data-sharing standards. These characteristics make *Scipion* not only an ideal environment for the current deployment, but also a sustainable platform for the continued evolution of automated cryo-EM processing pipelines.

CHAPTER 6 – CONCLUSION

We have developed and validated a fully automated, on-the-fly processing pipeline that successfully addresses key bottlenecks in the cryo-EM workflow. Its strength lies not only in its function as an intelligent diagnostic tool, delivering rapid and reliable data quality assessment rather than merely pursuing ultimate resolution, but also in the new, centralized quality-monitoring dashboard that complements and extends these capabilities. By consolidating processing metadata into an intuitive interface, the dashboard transforms the pipeline's internal diagnostics into a transparent, user-friendly environment where quality issues become immediately interpretable. This ensures that automated decisions, such as micrograph rejection, are no longer opaque but instead supported by clear, actionable visual evidence.

Through extensive benchmarking, facility-level deployment, and real-world user feedback, we have demonstrated the robustness of the pipeline across diverse and challenging samples. Its speed meets and exceeds the pace of modern high-throughput data acquisition, ensuring that a comprehensive assessment of data quality is available before a collection session is completed. This empowers researchers to make informed, on-the-fly decisions that maximize the efficiency and scientific value of precious microscope time.

A key advantage of our approach is its accessibility: the pipeline requires no user scripting, and workflows can be launched directly from Scipion using publicly available *JSON* templates hosted in the WorkflowHub collection "CryoEM Facility Workflows" [97]. The dashboard, equally lightweight and modular, can be deployed locally or behind secure institutional endpoints and readily extended with new metrics or views as facility needs evolve.

Together, the automated pipeline and the quality-monitoring dashboard represent a significant advance toward a more efficient, transparent, and intelligent cryo-EM ecosystem. They lay the groundwork for future developments in adaptive acquisition, threshold optimization, and machine-learning-driven decision support, ultimately contributing to a more accessible and data-driven cryo-EM workflow for the community.

6.1 Future Work

While the pipeline presented in this Thesis represents a significant advance, there remains room for improvement across several complementary directions. Based on the experience gained during development, deployment, and facility-scale testing, we identify four main areas for future work:

- **Pipeline enhancements:** Future versions of the workflow would benefit from integrating more sophisticated algorithms to address specific challenges observed in diverse samples. Examples include implementing dynamic box-size adaptation for non-globular or highly elongated particles, or adding dedicated modules to detect and filter common contaminants such as empty nanodiscs, micelles, or overlapping particles. Another important improvement would be transitioning from the current static analysis of a fixed particle subset to a fully dynamic streaming model. In such a system, 2D and 3D analyses would update continuously as new particles are acquired, enabling a new level of real-time structural feedback during data collection.
- **Dashboard expansion:** At present, the Quality Monitor focuses mainly on the Data Curation stage. Thanks to its modular and extensible design, the dashboard could be expanded to incorporate additional layers of feedback, including particle-picking statistics, 2D-class evolution, 3D *ab initio* and refinement convergence, or directional-resolution metrics. This would provide users with a unified, facility-ready platform where the entire on-the-fly processing workflow becomes transparent, traceable, and easy to interpret at a glance.
- **Advances in adaptive acquisition:** In single-particle analysis (SPA), there is growing interest in adaptive acquisition strategies that close the loop between microscope decisions and real-time processing feedback. Future work could explore classifying micrographs by ice thickness (hole type) and quality parameters associated with each hole type, correlating particle orientations with specific grid regions, or identifying underrepresented views during on-the-fly 3D analysis. These approaches would allow the microscope to actively target the most informative regions of the grid, improving angular coverage and overall resolution.
- **Extending automation to CryoET:** Cryo-electron tomography (CryoET) has recently gained significant momentum and is rapidly becoming a central technique for *in situ* structural biology. However, its processing pipelines remain far less mature and considerably more complex than those of SPA. A natural continuation of this work would be the development of an automated, on-the-fly processing pipeline for CryoET, inspired by the methods and principles established in this Thesis, particularly as the *Scipion* framework is rapidly evolving in this direction. This could include automated tilt-series alignment, real-time tomogram reconstruction, and early tomogram-quality assessment, ultimately contributing to a fully automated CryoET data-processing ecosystem.

CONCLUSIÓN

Hemos desarrollado y validado un pipeline de procesamiento completamente automatizado y *on-the-fly* que aborda con éxito los principales cuellos de botella del flujo de trabajo en *cryo-EM*. Su fortaleza no reside únicamente en su capacidad como herramienta diagnóstica inteligente, ofreciendo una evaluación rápida y fiable de la calidad de los datos, más allá de perseguir exclusivamente la resolución final, sino también en el nuevo panel centralizado de monitorización de calidad que amplía y complementa dichas capacidades. Al reunir todos los metadatos de procesamiento en una interfaz intuitiva, el panel convierte los diagnósticos internos del pipeline en un entorno transparente y fácil de interpretar, donde los problemas de calidad se identifican de inmediato. Esto garantiza que decisiones automatizadas como la limpieza de micrografías dejen de ser procesos opacos y pasen a estar respaldadas por evidencia visual clara y accionable.

Gracias a una evaluación comparativa exhaustiva, su implementación en centros de adquisición de datos y la retroalimentación de usuarios en condiciones operativas reales, hemos demostrado la robustez del flujo de procesamiento frente a muestras diversas y desafiantes. Su velocidad iguala, y en muchos casos supera, el ritmo de la adquisición de microscopios modernos de alto rendimiento, proporcionando una evaluación completa de la calidad de los datos antes de que la sesión de adquisición llegue a su fin. De este modo, los investigadores pueden tomar decisiones informadas en tiempo real, maximizando la eficiencia y el valor científico del tiempo de microscopio, un recurso extremadamente valioso.

Una ventaja clave de nuestro enfoque es su accesibilidad: el *pipeline* no requiere ningún tipo de *scripting* por parte del usuario, y los flujos de trabajo pueden lanzarse directamente desde Scipion utilizando las plantillas *JSON* públicas disponibles en la colección de WorkflowHub "CryoEM Facility Workflows" [97]. El panel, igualmente ligero y modular, puede desplegarse localmente o detrás de puntos de acceso institucionales seguros, y ampliarse fácilmente con nuevas métricas o vistas según evolucionen las necesidades del campo y de cada centro.

En conjunto, el flujo de procesamiento automatizado y el panel de monitorización de calidad representan un avance importante hacia un ecosistema de *cryo-EM* más eficiente, transparente e inteligente. Ambos establecen los cimientos para futuros desarrollos en adquisición adaptativa, optimización automática de umbrales y sistemas de apoyo a la decisión basados en aprendizaje automático, contribuyendo en última instancia a un flujo de trabajo de *cryo-EM* más accesible y guiado por datos para toda la comunidad.

Trabajo futuro

El flujo de procesamiento presentado en esta Tesis supone un avance importante, pero aún existen múltiples líneas de desarrollo que pueden ampliar sus capacidades. A partir de la experiencia obtenida durante su implementación, despliegue y validación en entornos reales de instalación, destacamos cuatro direcciones principales para el trabajo futuro:

- **Mejoras del *pipeline*:** Las próximas versiones del flujo de trabajo podrían incorporar algoritmos más avanzados para afrontar retos observados en muestras especialmente complejas. Por ejemplo, sería útil desarrollar una adaptación dinámica del tamaño de caja para partículas no globulares o muy elongadas, o añadir programas específicos que detecten y filtren contaminantes habituales como nanodiscos vacíos, micelas o partículas solapadas. Otra mejora relevante consiste en pasar del análisis estático actual, basado en un conjunto fijo de partículas, a un modelo de *streaming* completamente dinámico. En este enfoque, los análisis 2D y 3D se actualizarían continuamente a medida que se adquieren nuevas partículas, permitiendo una retroalimentación estructural más completa en tiempo real durante la adquisición.
- **Ampliación del panel de calidad:** Actualmente, el panel de monitorización se centra en la etapa de *Data Curation*. Gracias a su diseño modular, puede ampliarse con facilidad para incluir nuevas capas de información: estadísticas del picado de partículas, evolución temporal de las clases 2D, estado de convergencia de reconstrucciones 3D o métricas de resolución direccional, entre otras. Esto permitiría disponer de una plataforma unificada, clara y pensada para centros de adquisición de datos, donde todo el procesamiento *on-the-fly* sea completamente transparente y fácil de interpretar.
- **Avances en adquisición adaptativa:** En el campo de SPA está creciendo el interés por estrategias que conectan directamente las decisiones del microscopio con la información que llega del procesamiento en tiempo real. Algunas posibilidades incluyen clasificar las micrografías según el grosor del hielo y el tipo de agujero con métricas de calidad, correlacionar orientaciones de partículas con regiones concretas de la rejilla o detectar vistas poco representadas durante el análisis 3D en tiempo real. Estas aproximaciones permitirían dirigir la adquisición hacia las zonas más informativas de la rejilla, mejorando la cobertura angular y, en consecuencia, la resolución final.
- **Extensión de la automatización a *CryoET*:** Tomografía en *CryoEM* (*CryoET*) está ganando protagonismo como herramienta clave para estudiar estructuras *in situ*, pero sus flujos de trabajo son más inmaduros, más complejos y menos automatizados que los de *SPA*. Un paso natural tras esta Tesis sería desarrollar un pipeline automático y *on-the-fly* para *CryoET*, aprovechando los principios y metodologías aquí descritos y

beneficiándose de la rápida evolución de *Scipion* en esta dirección. Esto incluiría desde la alineación automática de *tilt series* y la reconstrucción de tomogramas en tiempo real hasta la evaluación temprana de la calidad de tomogramas, acercándonos a un ecosistema de procesamiento para *CryoET* completamente automatizado.

BIBLIOGRAPHY

[1] H. Wang, "Cryo-electron microscopy for structural biology: current status and future perspectives," *Sci. China Life Sci.*, vol. 58, no. 8, pp. 750–756, Aug. 2015, doi: 10.1007/s11427-015-4851-2.

[2] V. Raimondi and A. Grinzato, "A basic introduction to single particles cryo-electron microscopy," *AIMS Biophys.*, vol. 9, no. 1, pp. 5–20, 2021, doi: 10.3934/biophy.2022002.

[3] A. Dhakal, R. Gyawali, L. Wang, and J. Cheng, "CryoPPP: A large expert-labelled cryo-EM image dataset for machine learning protein particle picking," *bioRxivorg*, Feb. 22, 2023. doi: 10.1101/2023.02.21.529443.

[4] A. J. G. Hey, *Feynman and computation*. Philadelphia, PA: CRC Press, 2018.

[5] Y. Cheng, "Single-particle cryo-EM-How did it get here and where will it go," *Science*, vol. 361, no. 6405, pp. 876–880, Aug. 2018, doi: 10.1126/science.aat4346.

[6] "What is Structural Biology - Instruct-ERIC." Accessed: Nov. 06, 2025. [Online]. Available: https://instruct-eric.org/what-is-structural-biology

[7] J. Dubochet *et al.*, "Cryo-electron microscopy of vitrified specimens," *Q. Rev. Biophys.*, vol. 21, no. 2, pp. 129–228, May 1988, doi: 10.1017/s0033583500004297.

[8] J. Frank, "Averaging of low exposure electron micrographs of non-periodic objects," *Ultramicroscopy*, vol. 1, no. 2, pp. 159–162, Dec. 1975, doi: 10.1016/s0304-3991(75)80020-9.

[9] J.-H. Chung and H. M. Kim, "The Nobel prize in chemistry 2017: High-resolution cryo-electron microscopy," *Appl. Microsc.*, vol. 47, no. 4, pp. 218–222, Dec. 2017, doi: 10.9729/am.2017.47.4.218.

[10] A. van Leewenhoeck, "Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd: concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused," *Philos. Trans. R. Soc. Lond.*, vol. 12, no. 133, pp. 821–831, Mar. 1677, doi: 10.1098/rstl.1677.0003.

[11] S. Bhakta and M. van Heel, "Single-particle cryo-EM," in *Cryo-Electron Microscopy in Structural Biology*, Boca Raton: CRC Press, 2024, pp. 87–109. doi: 10.1201/9781003326106-8.

[12] L. de Broglie, "XXXV. A tentative theory of light quanta," *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, vol. 47, no. 278, pp. 446–458, Feb. 1924, doi: 10.1080/14786442408634378.

[13] E. Ruska, "The development of the electron microscope and of electron microscopy," *Rev. Mod. Phys.*, vol. 59, no. 3, pp. 627–638, Jul. 1987, doi: 10.1103/revmodphys.59.627.

[14] W. Kühlbrandt, "Biochemistry. The resolution revolution," *Science*, vol. 343, no. 6178, pp. 1443–1444, Mar. 2014, doi: 10.1126/science.1251652.

[15] S. Sunil, "Making the Electron Microscope," Asimov Press. Accessed: Oct. 27, 2025. [Online]. Available: https://www.asimov.press/p/electron-microscope

[16] "Krios 5 Cryo TEM." Accessed: Oct. 27, 2025. [Online]. Available: https://www.thermofisher.com/es/es/home/electron-microscopy/products/transmission-electron-microscopes/krios-cryo-tem/accessories.html

[17] S. Brenner and R. W. Horne, "A negative staining method for high resolution electron microscopy of viruses," *Biochim. Biophys. Acta*, vol. 34, pp. 103–110, Jul. 1959, doi: 10.1016/0006-3002(59)90237-9.

[18] E. F. J. van Bruggen, E. H. Wiebenga, and M. Gruber, "Negative-staining electron microscopy of proteins at pH values below their isoelectric points. Its application to hemocyanin," *Biochim. Biophys. Acta*, vol. 42, pp. 171–172, Jan. 1960, doi: 10.1016/0006-3002(60)90771-x.

[19] D. J. DeRosier and P. B. Moore, "Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry," *J. Mol. Biol.*, vol. 52, no. 2, pp. 355–369, Sep.

1970, doi: 10.1016/0022-2836(70)90036-7.
[20] R. A. Crowther, "Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 261, no. 837, pp. 221–230, May 1971, doi: 10.1098/rstb.1971.0054.
[21] M. Radermacher, T. Wagenknecht, A. Verschoor, and J. Frank, "Three-dimensional structure of the large ribosomal subunit from Escherichia coli," *EMBO J.*, vol. 6, no. 4, pp. 1107–1114, Apr. 1987, doi: 10.1002/j.1460-2075.1987.tb04865.x.
[22] K. A. Taylor and R. M. Glaeser, "Electron diffraction of frozen, hydrated protein crystals," *Science*, vol. 186, no. 4168, pp. 1036–1037, Dec. 1974, doi: 10.1126/science.186.4168.1036.
[23] K. A. Taylor and R. M. Glaeser, "Retrospective on the early development of cryoelectron microscopy of macromolecules and a prospective on opportunities for the future," *J. Struct. Biol.*, vol. 163, no. 3, pp. 214–223, Sep. 2008, doi: 10.1016/j.jsb.2008.06.004.
[24] M. Adrian, J. Dubochet, J. Lepault, and A. W. McDowall, "Cryo-electron microscopy of viruses," *Nature*, vol. 308, no. 5954, pp. 32–36, 1984, doi: 10.1038/308032a0.
[25] M. van Heel and J. Frank, "Use of multivariates statistics in analysing the images of biological macromolecules," *Ultramicroscopy*, vol. 6, no. 1, pp. 187–194, Jan. 1981, doi: 10.1016/s0304-3991(81)80197-0.
[26] M. van Heel, "Multivariate statistical classification of noisy images (randomly oriented biological macromolecules)," *Ultramicroscopy*, vol. 13, no. 1–2, pp. 165–183, 1984, doi: 10.1016/0304-3991(84)90066-4.
[27] R. Henderson and P. N. Unwin, "Three-dimensional model of purple membrane obtained by electron microscopy," *Nature*, vol. 257, no. 5521, pp. 28–32, Sep. 1975, doi: 10.1038/257028a0.
[28] S. H. W. Scheres, "RELION: implementation of a Bayesian approach to cryo-EM structure determination," *J. Struct. Biol.*, vol. 180, no. 3, pp. 519–530, Dec. 2012, doi: 10.1016/j.jsb.2012.09.006.
[29] A. Punjani, "Algorithmic advances in single particle cryo-EM data processing using CryoSPARC," *Microsc. Microanal.*, vol. 26, no. S2, pp. 2322–2323, Aug. 2020, doi: 10.1017/s1431927620021194.
[30] D. Tegunov and P. Cramer, "Real-time cryo-electron microscopy data preprocessing with Warp," *Nat. Methods*, vol. 16, no. 11, pp. 1146–1152, Nov. 2019, doi: 10.1038/s41592-019-0580-y.
[31] J. M. de la Rosa-Trevín *et al.*, "Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy," *J. Struct. Biol.*, vol. 195, no. 1, pp. 93–99, Jul. 2016, doi: 10.1016/j.jsb.2016.04.010.
[32] T. Bhamre, T. Zhang, and A. Singer, "Denoising and covariance estimation of single particle cryo-EM images," *J. Struct. Biol.*, vol. 195, no. 1, pp. 72–81, Jul. 2016, doi: 10.1016/j.jsb.2016.04.013.
[33] T. Wagner *et al.*, "SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM," *Commun. Biol.*, vol. 2, no. 1, p. 218, Jun. 2019, doi: 10.1038/s42003-019-0437-z.
[34] T. Bepler *et al.*, "TOPAZ: A positive-unlabeled convolutional neural network CryoEM particle picker that can pick any size and shape particle," *Microsc. Microanal.*, vol. 25, no. S2, pp. 986–987, Aug. 2019, doi: 10.1017/s143192761900566x.
[35] D. Xu and N. Ando, "Miffi: Improving the accuracy of CNN-based cryo-EM micrograph filtering with fine-tuning and Fourier space information," *J. Struct. Biol.*, vol. 216, no. 2, p. 108072, Jun. 2024, doi: 10.1016/j.jsb.2024.108072.
[36] Y. Li, J. N. Cash, J. J. G. Tesmer, and M. A. Cianfrocco, "High-throughput cryo-EM enabled by user-free preprocessing routines," *Structure*, vol. 28, no. 7, pp. 858–869.e3, Jul. 2020, doi:

10.1016/j.str.2020.03.008.
[37] 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," *bioRxiv*, Sep. 30, 2021. doi: 10.1101/2021.09.30.462538.
[38] D. N. Mastronarde, "Automated electron microscope tomography using robust prediction of specimen movements," *J. Struct. Biol.*, vol. 152, no. 1, pp. 36–51, Oct. 2005, doi: 10.1016/j.jsb.2005.07.007.
[39] J. Bhandari, D. Kompaniiets, A. K. Singh, C. Bator, J. Porta, and B. Liu, "Efficient strategies and troubleshooting for single particle cryoEM data collection using EPU," *BMC Methods*, vol. 2, no. 1, Feb. 2025, doi: 10.1186/s44330-025-00025-8.
[40] J. Bouvette, Q. Huang, A. A. Riccio, W. C. Copeland, A. Bartesaghi, and M. J. Borgnia, "Automated systematic evaluation of cryo-EM specimens with SmartScope," *Elife*, vol. 11, no. e80047, Aug. 2022, doi: 10.7554/eLife.80047.
[41] P. R. Baldwin *et al.*, "Big data in cryoEM: automated collection, processing and accessibility of EM data," *Curr. Opin. Microbiol.*, vol. 43, pp. 1–8, Jun. 2018, doi: 10.1016/j.mib.2017.10.005.
[42] M. Liao, E. Cao, D. Julius, and Y. Cheng, "Structure of the TRPV1 ion channel determined by electron cryo-microscopy," *Nature*, vol. 504, no. 7478, pp. 107–112, Dec. 2013, doi: 10.1038/nature12822.
[43] E. Cao, M. Liao, Y. Cheng, and D. Julius, "TRPV1 structures in distinct conformations reveal activation mechanisms," *Nature*, vol. 504, no. 7478, pp. 113–118, Dec. 2013, doi: 10.1038/nature12823.
[44] S. M. Fica and K. Nagai, "Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine," *Nat. Struct. Mol. Biol.*, vol. 24, no. 10, pp. 791–799, Oct. 2017, doi: 10.1038/nsmb.3463.
[45] Y. Shi, "Mechanistic insights into precursor messenger RNA splicing by the spliceosome," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 11, pp. 655–670, Nov. 2017, doi: 10.1038/nrm.2017.86.
[46] A. Iudin *et al.*, "EMPIAR: The electron microscopy public image archive," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1503–D1511, Jan. 2023, doi: 10.1093/nar/gkac1062.
[47] A. Chari and H. Stark, "Prospects and limitations of high-resolution single-particle cryo-electron microscopy," *Annu. Rev. Biophys.*, vol. 52, pp. 391–411, May 2023, doi: 10.1146/annurev-biophys-111622-091300.
[48] F. J. Sigworth, "Principles of cryo-EM single-particle image processing," *Microscopy (Oxf.)*, vol. 65, no. 1, pp. 57–67, Feb. 2016, doi: 10.1093/jmicro/dfv370.
[49] T. R. D. Costa, A. Ignatiou, and E. V. Orlova, "Structural analysis of protein complexes by cryo electron microscopy," *Methods Mol. Biol.*, vol. 1615, pp. 377–413, 2017, doi: 10.1007/978-1-4939-7033-9_28.
[50] D. Lee, H. Lee, J. Lee, S.-H. Roh, and N.-C. Ha, "Copper oxide spike grids for enhanced solution transfer in cryogenic electron microscopy," *Mol. Cells*, vol. 46, no. 9, pp. 538–544, Sep. 2023, doi: 10.14348/molcells.2023.0058.
[51] D. Lyumkis, "Challenges and opportunities in cryo-EM single-particle analysis," *J. Biol. Chem.*, vol. 294, no. 13, pp. 5181–5197, Mar. 2019, doi: 10.1074/jbc.REV118.005602.
[52] Q. Chen *et al.*, "A large-scale curated and filterable dataset for cryo-EM foundation model pre-training," *Sci. Data*, vol. 12, no. 1, p. 960, Jun. 2025, doi: 10.1038/s41597-025-05179-2.
[53] 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," *Nat. Methods*, vol. 14, no. 4, pp. 331–332, Apr. 2017, doi: 10.1038/nmeth.4193.

[54] 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. Pt 1, pp. 5–17, Jan. 2019, doi: 10.1107/S205225251801463X.
[55] X. Li *et al.*, "Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM," *Nat. Methods*, vol. 10, no. 6, pp. 584–590, Jun. 2013, doi: 10.1038/nmeth.2472.
[56] D. Střelák, D. Marchán, J. M. Carazo, and C. O. S Sorzano, "Performance and quality comparison of movie alignment software for cryogenic electron microscopy," *Micromachines (Basel)*, vol. 14, no. 10, p. 1835, Sep. 2023, doi: 10.3390/mi14101835.
[57] J. H. Mendez, E. Y. D. Chua, M. Paraan, C. S. Potter, and B. Carragher, "Automated pipelines for rapid evaluation during cryoEM data acquisition," *Curr. Opin. Struct. Biol.*, vol. 83, no. 102729, p. 102729, Dec. 2023, doi: 10.1016/j.sbi.2023.102729.
[58] Z. A. Ripstein and J. L. Rubinstein, "Processing of cryo-EM movie data," *Methods Enzymol.*, vol. 579, pp. 103–124, Jun. 2016, doi: 10.1016/bs.mie.2016.04.009.
[59] Y. Zhang *et al.*, "Single-particle cryo-EM: alternative schemes to improve dose efficiency," *J. Synchrotron Radiat.*, vol. 28, no. Pt 5, pp. 1343–1356, Sep. 2021, doi: 10.1107/S1600577521007931.
[60] A. Rohou and N. Grigorieff, "CTFFIND4: Fast and accurate defocus estimation from electron micrographs," *bioRxiv*, bioRxiv, Jun. 16, 2015. doi: 10.1101/020917.
[61] 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.
[62] J. M. de la Rosa-Trevín *et al.*, "Xmipp 3.0: an improved software suite for image processing in electron microscopy," *J. Struct. Biol.*, vol. 184, no. 2, pp. 321–328, Nov. 2013, doi: 10.1016/j.jsb.2013.09.015.
[63] Zhang, K., Li, M., Sun, F., "Gautomatch: an efficient and convenient gpu-based automatic particle selection program," *Unpublished manuscript*, 2011.
[64] F. Wang *et al.*, "DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM," *J. Struct. Biol.*, vol. 195, no. 3, pp. 325–336, Sep. 2016, doi: 10.1016/j.jsb.2016.07.006.
[65] U. Zarzecka *et al.*, "Functional analysis and cryo-electron microscopy of Campylobacter jejuni serine protease HtrA," *Gut Microbes*, vol. 12, no. 1, pp. 1–16, Nov. 2020, doi: 10.1080/19490976.2020.1810532.
[66] J. Frank, *Three-dimensional electron microscopy of macromolecular assemblies: Visualization of biological molecules in their native state*, 2nd ed. Cary, NC: Oxford University Press, 2006.
[67] E. V. Orlova and H. R. Saibil, "Structural analysis of macromolecular assemblies by electron microscopy," *Chem. Rev.*, vol. 111, no. 12, pp. 7710–7748, Dec. 2011, doi: 10.1021/cr100353t.
[68] S. J. Ludtke, T. Durmaz, M. Chen, and J. M. Bell, "New strategies for improving CryoEM single particle analysis in EMAN2.2," *Microsc. Microanal.*, vol. 23, no. S1, pp. 810–811, Jul. 2017, doi: 10.1017/s1431927617004718.
[69] B. Zhu and H. Liu, "A method to fast ab initio cryo-electron microscopy initial volume reconstruction," *Preprints*, Jun. 14, 2023. doi: 10.20944/preprints202306.1023.v1.
[70] S. H. W. Scheres, "Classification of structural heterogeneity by maximum-likelihood methods," *Methods Enzymol.*, vol. 482, pp. 295–320, 2010, doi: 10.1016/S0076-6879(10)82012-9.
[71] F. J. Sigworth, P. C. Doerschuk, J.-M. Carazo, and S. H. W. Scheres, "An introduction to maximum-likelihood methods in cryo-EM," *Methods Enzymol.*, vol. 482, pp. 263–294, 2010, doi: 10.1016/S0076-6879(10)82011-7.
[72] D. Haselbach *et al.*, "Structure and conformational dynamics of the human spliceosomal bact

complex," *Cell*, vol. 172, no. 3, pp. 454–464.e11, Jan. 2018, doi: 10.1016/j.cell.2018.01.010.
[73] C. O. Sanchez Sorzano, A. L. Alvarez-Cabrera, M. Kazemi, J. M. Carazo, and S. Jonić, "StructMap: Elastic distance analysis of electron microscopy maps for studying conformational changes," *Biophys. J.*, vol. 110, no. 8, pp. 1753–1765, Apr. 2016, doi: 10.1016/j.bpj.2016.03.019.
[74] E. J. Verbeke, Y. Zhou, A. P. Horton, A. L. Mallam, D. W. Taylor, and E. M. Marcotte, "Separating distinct structures of multiple macromolecular assemblies from cryo-EM projections," *J. Struct. Biol.*, vol. 209, no. 1, p. 107416, Jan. 2020, doi: 10.1016/j.jsb.2019.107416.
[75] L. Anton, D. W. Cobb, and C.-M. Ho, "Structural parasitology of the malaria parasite Plasmodium falciparum," *Trends Biochem. Sci.*, vol. 47, no. 2, pp. 149–159, Feb. 2022, doi: 10.1016/j.tibs.2021.10.006.
[76] J. Zivanov *et al.*, "RELION-3: new tools for automated high-resolution cryo-EM structure determination," *bioRxiv*, bioRxiv, Sep. 19, 2018. doi: 10.1101/421123.
[77] A. D. Parvate *et al.*, "Cryo-EM structure of the diapause chaperone artemin," *Front. Mol. Biosci.*, vol. 9, no. 998562, p. 998562, Nov. 2022, doi: 10.3389/fmolb.2022.998562.
[78] P. B. Rosenthal and R. Henderson, "Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy," *J. Mol. Biol.*, vol. 333, no. 4, pp. 721–745, Oct. 2003, doi: 10.1016/j.jmb.2003.07.013.
[79] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of coot," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 66, no. Pt 4, pp. 486–501, Apr. 2010, doi: 10.1107/S0907444910007493.
[80] 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. Pt 10, pp. 861–877, Oct. 2019, doi: 10.1107/S2059798319011471.
[81] T. I. Croll, "ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps," *Acta Crystallogr. D Struct. Biol.*, vol. 74, no. Pt 6, pp. 519–530, Jun. 2018, doi: 10.1107/S2059798318002425.
[82] C. Suloway *et al.*, "Automated molecular microscopy: the new Leginon system," *J. Struct. Biol.*, vol. 151, no. 1, pp. 41–60, Jul. 2005, doi: 10.1016/j.jsb.2005.03.010.
[83] Y. Z. Tan *et al.*, "Addressing preferred specimen orientation in single-particle cryo-EM through tilting," *Nat. Methods*, vol. 14, no. 8, pp. 793–796, Aug. 2017, doi: 10.1038/nmeth.4347.
[84] P. Conesa *et al.*, "Scipion3: A workflow engine for cryo-electron microscopy image processing and structural biology," *Biol. Imaging*, vol. 3, p. e13, Jun. 2023, doi: 10.1017/S2633903X23000132.
[85] C. L. Lawson and W. Chiu, "Comparing cryo-EM structures," *J. Struct. Biol.*, vol. 204, no. 3, pp. 523–526, Dec. 2018, doi: 10.1016/j.jsb.2018.10.004.
[86] J. M. de la Rosa-Trevín *et al.*, "Using Scipion for stream image processing at cryo-EM facilities," *Acta Crystallogr. A Found. Adv.*, vol. 74, no. a1, pp. a161–a161, Jul. 2018, doi: 10.1107/s0108767318098380.
[87] D. Maluenda *et al.*, "Flexible workflows for on-the-fly electron-microscopy single-particle image processing using Scipion," *Acta Crystallogr. D Struct. Biol.*, vol. 75, no. Pt 10, pp. 882–894, Oct. 2019, doi: 10.1107/S2059798319011860.
[88] H.-F. Liu *et al.*, "nextPYP: a comprehensive and scalable platform for characterizing protein variability in situ using single-particle cryo-electron tomography," *Nat. Methods*, vol. 20, no. 12, pp. 1909–1919, Dec. 2023, doi: 10.1038/s41592-023-02045-0.
[89] M. Stabrin, F. Schoenfeld, T. Wagner, S. Pospich, C. Gatsogiannis, and S. Raunser, "TranSPHIRE: Automated and feedback-optimized on-the-fly processing for cryo-EM," *bioRxiv*, bioRxiv, Jun. 17, 2020. doi: 10.1101/2020.06.16.155275.

[90] D. Strelak *et al.*, "Advances in Xmipp for cryo-electron microscopy: From Xmipp to scipion," *Molecules*, vol. 26, no. 20, p. 6224, Oct. 2021, doi: 10.3390/molecules26206224.
[91] K. Maruthi, M. Kopylov, and B. Carragher, "Automating decision making in the cryo-EM pre-processing pipeline," *Structure*, vol. 28, no. 7, pp. 727–729, Jul. 2020, doi: 10.1016/j.str.2020.06.004.
[92] J. Elferich, L. Kong, X. Zottig, and N. Grigorieff, "CTFFIND5 provides improved insight into quality, tilt, and thickness of TEM samples," *Elife*, vol. 13, no. RP97227, p. RP97227, Dec. 2024, doi: 10.7554/eLife.97227.
[93] A. Heimowitz, J. Andén, and A. Singer, "APPLE picker: Automatic Particle Picking, a low-Effort cryo-EM framework," *arXiv [cs.CV]*, Feb. 01, 2018. [Online]. Available: http://arxiv.org/abs/1802.00469
[94] R. Langlois, J. Pallesen, J. T. Ash, D. Nam Ho, J. L. Rubinstein, and J. Frank, "Automated particle picking for low-contrast macromolecules in cryo-electron microscopy," *J. Struct. Biol.*, vol. 186, no. 1, pp. 1–7, Apr. 2014, doi: 10.1016/j.jsb.2014.03.001.
[95] T. Wagner and S. Raunser, "The evolution of SPHIRE-crYOLO particle picking and its application in automated cryo-EM processing workflows," *Commun. Biol.*, vol. 3, no. 1, p. 61, Feb. 2020, doi: 10.1038/s42003-020-0790-y.
[96] C. J. F. Cameron, S. J. H. Seager, F. J. Sigworth, H. D. Tagare, and M. B. Gerstein, "REliable PIcking by Consensus (REPIC): a consensus methodology for harnessing multiple cryo-EM particle pickers," *Commun. Biol.*, vol. 7, no. 1, p. 1421, Oct. 2024, doi: 10.1038/s42003-024-07045-0.
[97] Scipion team, "CryoEM Facility Workflows," WorkflowHub, 2025. [Online]. Available: https://workflowhub.eu/collections/31
[98] O. J. R. Gustafsson *et al.*, "WorkflowHub: a registry for computational workflows," *Sci. Data*, vol. 12, no. 1, p. 837, May 2025, doi: 10.1038/s41597-025-04786-3.
[99] E. Kandiah *et al.*, "CM01: a facility for cryo-electron microscopy at the European Synchrotron," *Acta Crystallogr. D Struct. Biol.*, vol. 75, no. Pt 6, pp. 528–535, Jun. 2019, doi: 10.1107/S2059798319006880.
[100] ESRF, *scipion-em-esrf*. 2025. [GitLab]. Available: https://gitlab.esrf.fr/sb/scipion-em-esrf

APPENDICES

Appendix A: Image processing table for the complete CryoPPP dataset

This table summarizes the image processing results for the complete CryoPPP dataset. Unlike (Table 2), which presents only a few representative high-quality cases (5 out of 25), this table includes the full set of 32 CryoPPP EMPIAR entries. "3D Map Result Assessment" indicates whether the resulting map displayed recognizable, protein-like features consistent with the target macromolecule. "Used mics" refers to the percentage of total micrographs deposited in EMPIARthat were included in processing, with a maximum of 1,000 per entry. "Accepted data curation" indicates the percentage of micrographs that passed the quality filters from the initial processing set. "Final Particles" shows the number of particles retained in the final refinement relative to those initially selected ("Initial Particles") from the curated micrographs, capped at 200,000 particles.

Appendix A Table. CryoPPP Full Dataset Image Processing Summary. Green indicates successful and robust 3D reconstructions; yellow denotes suboptimal reconstructions; and red corresponds to non-conclusive 3D reconstructions. For the **High-quality cases**, **grey highlights** mark the **entries that were not analyzed in-depth in the Results Chapter**. Blue highlights mark key aspects of the image processing that may have influenced the final 3D structure. *3D Map Result* indicates whether the resulting map displayed recognizable, protein-like features consistent with the target macromolecule. *Used mics* refers to the percentage of total micrographs deposited in EMPIAR that were included in the processing, with a maximum of 1,000 micrographs per entry. *Accepted data curation* indicates the percentage of micrographs that passed the quality filters from the initial processing set. *Final/Initial Parts.* refers to the number of particles retained in the final refinement compared to those initially selected from the curated set of micrographs, capped at 200,000 particles.

EMPIAR	Protein Type	Mol. Weight (kDa)	Used Mics (%)	Deposit/ FSC resolution (A)	3D Map Result	Accepted data curation (%)	Particle Diam. /Est. (A)	Final/ Initial Parts.	Features of Micrographs
10061	Beta-galactosidase	467.06	65	2.2 /3.1	High-quality	89.9	150/ 147	77K/ 92K	aggregated particles + sub optimal particle concentration
10406	Ribosome (70S)	632.89	36.8	2.7/3.1	High-quality	58.8	240/ 232	23K/ 74K	mono-dispersed particles + moderate protein edge texture
10576	Nuclear Protein (DNA)	290.21	50.4	2.9/3.2	High-quality	97.4	180/ 208	103K/ 200K	low contrast micrographs + difficult to recognize and pick particles
10737	Membrane Protein (E-coli)	155.83	29	2.2/4.1	High-quality	75.2	179/ 172	56K/ 177K	mono-dispersed particles + sufficient contrast
10184	Aldolase	NA	62	2.4/3.2	High-quality	47.5	100/ 90	111K/ 200K	mono-dispersed compact particles

EMPIAR	Protein Type	Mol. Weight (kDa)	Used Mics (%)	Deposit/ FSC resolution (A)	3D Map Result	Accepted data curation (%)	Particle Diam. /Est. (A)	Final/ Initial Parts.	Features of Micrographs
11057	Hydrolase	149.43	11.9	2.8/3.1	High-quality	80.4	140/ 118	88K/ 161K	difficult to identify particles + sub optimal particle concentration + ice issues
10081	Transport Protein	298.57	100	3.5/4.2	High-quality	98.6	200/ 239	34K/ 93K	distinct protein particles visible though Naked eyes + ideal micrographs with less ice patches + moderate density of particles
10289	Transport Protein	361.39	100	3.8/4	High-quality	91.1	200/ 204	75K/ 185K	mono-disperse distribution + moderate contrast + different particle orientation
10444	Membran e Protein	295.89	32.5	3.6/3.3	High-quality	73.8	180/ 170	70K/ 170K	monodisperse distribution + particles with sufficient contrast easily identified by eye
10816	Transport Protein	166.62	25.8	4/6.7	High-quality	55.7	180/ 177	22K/ 73K	variation in ice thickness
11051	Transcript ion/DNA/ RNA	357.31	22.3	3.4/3.4	High-quality	79.7	180/ 151	33K/ 134K	non differentiable textured particles + wide variations in protein particle conformations
11183	Signaling Protein	139.36	26.3	2.9/4.2	High-quality	71	140/ 110	68K/ 200K	sub optimal particle concentration + varying ice thickness
10291	Transport Protein	361.39	100	3.6/3.9	High-quality	73.3	160/ 138	67K/ 102K	distinct protein particles distinguishable through naked eyes + ideal micrographs with low ice patches + moderate density
10077	Ribosome (70S)	2198.7 8	16.3	4.4/4.5	High-quality	26.2	250/ 230	22K/ 32K	moderate contrast and moderate edge textured particles + monodisperse distribution
10028	Ribosome (80S)	2135.8 9	100	3.2/4.3	High-quality	100	300/ 251	13K/ 16K	particles with sufficient contrast + mono-dispersed particles
10096	Viral Protein	NA	100	4.2/3.7	High-quality	82.3	110/ 120	70K/ 200K	differentiable textured particles + high number of ice patches + distributed top, side and inclined views of particles
10532	Viral Protein	191.76	64.3	4.1/3.7	High-quality	55.6	179/ 156	20K/ 90K	extreme ice contamination + moderate contrast particles
10240	Lipid Transport Protein	171.72	46.6	3.6/4.2	High-quality	69.4	170/ 168	59K/ 131K	particles with moderate contrast
10005	TRPV1 Tansport Protein	272.97	100	3.3/4.4	High-quality	33.8	172/ 237	12K/ 32K	high carbon edge regions + diverse orientation of particles
10017	β-galacto sidase	NA	100	4.2/4.4	High-quality	47.6	190/ 184	15K/ 22K	mono-dispersed particles + sufficient contrast identifiable by naked eye

EMPIAR	Protein Type	Mol. Weight (kDa)	Used Mics (%)	Deposit/ FSC resolution (A)	3D Map Result	Accepted data curation (%)	Particle Diam. /Est. (A)	Final/ Initial Parts.	Features of Micrographs
10075	Bacteriophage MS2	NA	100	7.8/6.5	High-quality	74.7	270/ 285	16K/ 27K	ice contaminations + sufficient contrast identifiable by naked eye
10059	Transport Protein (TRPV1)	317.88	83.3	2/3.3	High-quality	99.4	160/ 161	83K/ 200K	mono-dispersed particles + sufficient contrast
10093	Membrane Protein	779.4	53.4	3.6/4.2	High-quality	97.9	208/ 193	51K/ 189K	particles with heterogenous confirmation
10345	Signaling Protein	244.68	60.9	3.5/6.3	High-quality	29.6	200/ 240	5K/ 12K	sub-optimal concentration + ice patches + non-differentiable textured particle edges
10097	Viral Protein	NA	100	4.2/6.7	High-quality	45.7	140/ 131	89K/173K	tilted sample
10389	Metal Binding Protein	1,042.17	23.1	2/4	Suboptimal	65.4	200/ 165	6K/ 30K	abundance of ice patches + dispersed protein particles + low number of particles per micrograph
10671	Signaling Protein	77.14	17	3.5/6.1	Suboptimal	96.6	110/ 86	90K/ 200K	extremely small protein particles + high density protein particles
10669	Proteasome (Plant Protein)	1,681.81	2.2	3.2/5.9	Suboptimal	95.3	500/ 229	27K/ 101K	carbon edges presence + disperse and distinct top, side and inclined views of particles
10387	Viral Protein (DNA)	185.87	49.5	2.8/3.7	Suboptimal	84.3	168/ 200	21K / 37K	highly aggregated protein particles + difficult to pick particles
10590	TRPV1 with DkTx and RTX	N/A	20.6	7.8/11.9	Suboptimal	94	236/ 250	29K / 181K	high contrast + mono dispersed particles + ice contaminations
10526	Ribosome (50S)	1085.81	90.7	2.8/(NA)	Failed	1.9	400	N/A	extremely high ice contamination + variation in ice thickness
10760	Membrane Protein	321.69	26	4.5/8.3	Failed	99.6	130/ 122	66K/200K	abundance of ice patches + mono-disperse distribution + particles with sufficient contrast