uc3m | Universidad Carlos III de Madrid

Master Degree in Biomedical Engineering

2024-2025

Master Thesis

"Design of data analysis workflows for Structural Biology"

Javier Sánchez del Río

CNB-CSIC: Carlos Óscar S. Sorzano

UC3M: María Arrate Muñoz Barrutia

Leganés, 2025

AVOID PLAGIARISM

The University uses the **Turnitin Feedback Studio** for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

AKNOWLEDGMENTS

I would like to take this opportunity to express my gratitude to everyone who supported me throughout this project.

To my tutor, Carlos Óscar Sánchez Sorzano, for giving me the opportunity to work on this project and always finding at least five minutes to answer my endless questions.

To everyone in the Biocomputing Unit at CNB, for creating such a friendly and supportive environment, and for always being there to help whenever I faced a problem.

To my friends, for helping me disconnect and relax when I needed it most.

To my family, whose constant support has always given me the confidence to overcome any obstacle in my path. I am here because of you.

And of course, to her, who has always been by my side, offering strength when mine faltered, calm when I was lost, and wisdom when I needed it most. Thank you for helping me focus on the "bright side of life".

This project would not have been possible without your support.

ABSTRACT

Max Knoll and Ernst Ruska introduced Electron microscopy (EM) in 1931, revolutionizing imaging. Cryo-EM technology solves the limitations presented by EM when working with organic matter.

Despite its advantages, cryo-EM generates large amounts of complex data, which needs to be processed using computational tools to generate good-quality images. Scipion is an open-source framework that integrates various cryo-EM software packages, facilitating streamlined and reproducible image processing workflows. By providing a user-friendly interface, Scipion facilitates the analysis of cryo-EM data, making it more accessible to researchers without extensive computational backgrounds.

The main goal of this bachelor's thesis is to implement new image processing algorithms within Scipion, allowing them to be used by researchers in an automated and easy way. The first implemented algorithm, CryoTEN, enhances the contrast in a cryo-EM density map, achieving better resolution for further molecule processing. The second algorithm helps the user to compare different atomic models of the same molecule and easily visualize the regions of higher confidence between them.

After the implementation of these two algorithms, a new plugin for Scipion is created: scipion-em-cryoten. This plugin installs CryoTEN model and integrates it in Scipion environment to efficiently enhance cryoEM density maps. Moreover, a new protocol (chimeraX - find discrepancies) is added to the available ones in scipion-em-chimeraX plugin. This new protocol compares different atomic models and provides a visual result of the discrepancy regions.

With these results, users can use a pre-designed workflow in Scipion that loads a map from a database, enhances it and improves its quality with cryoTEN, then generate different atomic models from the improved map and compare them using the new ChimeraX protocol: find discrepancies. This workflow provides a fully integrated, reproducible, and accessible pipeline for cryo-EM data analysis, enabling researchers to enhance map quality and validate their results through visual structural comparisons.

Keywords: Cryo-electron microscopy, Structural biology, Scipion, CryoTEN,Image processing workflows, ChimeraX, Deep learning

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

EM	Electron Microscopy
TEM	Transmission Electron Microscopy
SEM	Scanning Electron Microscopy
cryo-EM	Cryo-electron Microscopy
CNB	National Center of Biotechnology
DMM	DeepMainMast
UCSF	University of California San Francisco
RMSD	Root Mean Square Deviation
EMDB	Electron Microscopy Data Bank
PDB	Protein Data Bank
wwPDB	Worldwide Protein Data Bank
DOI	Digital Object Identifier
CNN	Convolutional Neural Network
mmCIF	Macro-molecular Crystallographic Information File
SNR	Signal-to-Noise Ratio
CFT	Contrast Transfer Function

1. INTRODUCTION

The urge for knowledge of the human race has led us to investigate more than what we can see and touch. Romans already tested different glass shapes to magnify little objects barely visible. With the years and the development of science, these lenses were improved, and the first optical microscopes appeared. In 1665, Robert Hooke used one of these initial versions of microscopes to observe and document the first human-seen 'cell' [31].

The optical microscopes kept being improved, but they reached a limit defined as Abbe's diffraction limit. Named after the German physicist Ernst Abbe [30], this principle establishes a theoretical limit to the resolution that an optical microscope can achieve. It is based on the formula:

$$d = \frac{\lambda}{2 \cdot NA} \tag{1.1}$$

According to it, the smallest detail that can be distinguished in a microscope is approximately half of the wavelength of the light used to observe the sample. In the case of optical microscopes, as they are based on the visible light spectrum, which ranges from 380 nm to 740 nm, the smallest detail that can be observed is approximately 200 nm.

This diffraction limit is the result of the nature of light. As light passes through a small aperture, it spreads out or diffracts, generating a diffraction pattern. This diffraction increases as the object gets smaller, until there is a point where the diffraction is so high that the objective cannot fully resolve its diffractive pattern and the object appears blurred. This is called the diffraction limit.

As this resolution limit cannot be changed, the optical microscopes had an important limitation when going smaller in the samples. A breakthrough occurred in 1931, when Max Knoll and Ernst Ruska introduced Electron microscopy (EM). The main revolution was the replacement of the light source by an electron beam, as electrons have 100,000 times smaller wavelengths than light. This way, the diffraction limit that limits the resolution of microscopes is decreased and the magnification that can be achieved is significantly increased. This type of microscopy has two main variations [26]:

- **Scanning electron Microscope (SEM):** SEM uses a focused beam of electrons to scan the surface of a sample. It generates images of the sample's surface topography and morphology by detecting the secondary electrons emitted from the surface.
- **Transmission electron Microscope (TEM):** TEM focuses on the internal structure of the sample. The electron beam is transmitted through a thin sample. The electrons transmitted through the sample are detected and provide information about its composition and morphology.

However, electron microscopy still presents some limitations in biological imaging. As the EM requires operating in vacuum conditions, the organic samples dehydrate and tend to collapse. Moreover, the electron beam fired at the sample easily rips organic matter. Therefore, the full potential of electron microscopy cannot be reached when working with organic samples.

These problems were solved in 1982, when the three scientists Jacques Dubochet, Joachim Frank, and Richard Henderson presented cryo-electron microscopy (cryo-EM), a technology that would later receive the Nobel Prize in Chemistry in 2017 [52]. They solved the issues with organic molecule imaging with the process of vitrification, which involves quickly freezing the sample to prevent crystals from forming inside it. This way, the biological samples maintain their native hydration and structure even in vacuum conditions.

1.1. Motivation

The cryo-EM introduction is an important milestone in the microscopy field. However, this imaging technology generates large amounts of data that require computational processing to extract meaningful information from the images obtained.

Cryo-EM is widely used in structural biology studies and, for this reason, there exists a wide variety of software tools to process cryo-EM data for many different purposes, such as improving resolution, refining alignment or eliminating distortions.

Scipion is an open-source software created by the CNB - Instruct Image Processing Center [43], in Madrid. This software aims to integrate and make more accessible all these packages and processing tools used for cryo-EM processing. By integrating all these tools and software for EM molecule processing, Scipion creates an environment where this type of analysis can be done more easily, faster and more automatically.

The motivation behind this work is to solve specific limitations in the current processing workflows, such as improving the efficiency of data handling and ensuring better quality control of the results. Despite the availability of several processing tools, challenges remain in integrating them effectively to make them easier to use and reduce the GPU consumption. This work focuses on enhancing the capabilities of Scipion to enable more accurate analysis of cryo-EM data and contributing to the advancement of structural biology research.

1.2. Objectives

The main objective of this bachelor's thesis is to develop, implement, and integrate two image-processing tools into the Scipion framework to enhance the analysis of cryo-electron microscopy (cryo-EM) data.

One of the main challenges when working with cryo-EM maps is the high level of noise present immediately after reconstruction. Effective denoising is essential before further processing, as it directly improves the performance and accuracy of downstream algorithms. While tools like DeepEMhancer [45] are already integrated into Scipion for map enhancement, CryoTEN provides a valuable alternative that is faster and significantly more efficient in terms of GPU memory consumption.

The second development in this thesis involves the integration of a new protocol named "ChimeraX - find discrepancies" into Scipion's ChimeraX plugin. This tool enables the comparison of atomic models—either generated by deep learning-based tools (e.g., ModelAngelo [23], KiharaLab [3]) or retrieved from structural databases—and highlights regions of structural agreement to assess model reliability. This comparison is especially useful for evaluating conformational differences or validating predictive models.

Therefore, the objectives can be established as the following:

1. **Implementation of CryoTEN within Scipion**, ensuring compatibility with existing workflows and enabling faster and more efficient map denoising.

2. **Development of a model comparison tool using ChimeraX**, based on root mean square deviation (RMSD) alignment, to visualize discrepancy regions between atomic models.

3. **Testing and validation** of both tools using cryo-EM datasets and standard processing pipelines within Scipion.

With the implementation of a new plugin for Scipion software called "scipion-em-cryoTEN", the users could easily automate the denoising process of cryo-EM samples making use of CryoTEN software and its advantages in this field. Additionally, with the development of the new ChimeraX protocol "find discrepancies" inside the existing ChimeraX plugin in Scipion, the users can benefit from easier, automated, and reproducible analysis when studying the atomic models generated by different software to test their confidence for later applications such as molecule interactions for drug development.

2. BACKGROUND

This section is intended to provide the basic knowledge assumed for the complete understanding of the thesis.

2.1. EM Microscopy

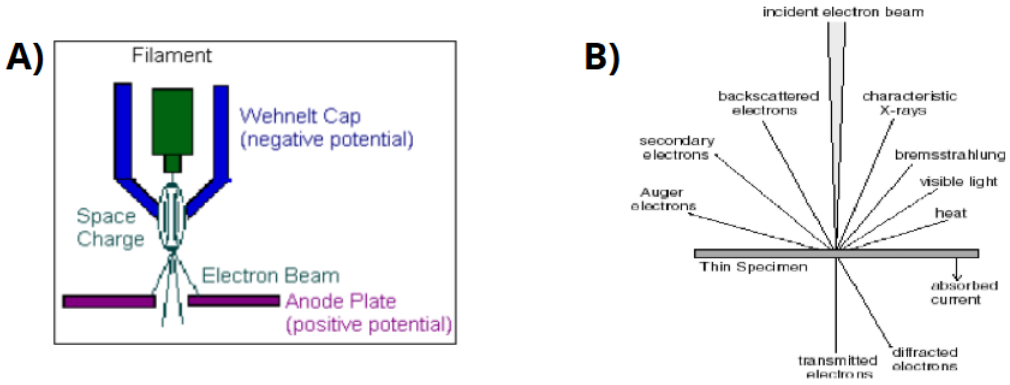
By using electron beams instead of light, electron microscopy allows unprecedented magnification. These electrons interact with the sample and are afterwards detected and processed to create an image of the sample [22]. For it to work, an electron microscope requires the following components:

- Electron source, typically consists of a V-shaped filament made of tungsten (W), surrounded by a Wehnelt electrode (also known as a Wehnelt cap). The procedure to generate the beam of electrons is the following: a positive electrical potential is applied to the anode, while the filament (cathode) is heated until it emits a stream of electrons. A schematic of this process can be seen in Figure 2.1.
- Electromagnetic lenses, which control the direction followed by the electron beam once it is sent. They are composed of parallel electric coils that generate a magnetic field that pulls the electrons to direct them.
- Electron detectors, which nowadays are usually digital cameras with scintillators or electron detectors. The type of electrons detected depends on the type of electron microscopy used

When this electron beam interacts with matter, e.g. the sample introduced in the EM, scattering of the electrons is produced. This scattering has two modalities: elastic and inelastic. In elastic scattering, electrons change their trajectory after interacting with matter, but kinetic energy and velocity are unchanged. Meanwhile, when inelastic scattering occurs, electrons collide with the ones on the sample and they move them from their orbits. These two types of scattering produce certain types of scattered electrons that, depending on the electron microscopy modality chosen, will be detected and processed to generate the final image.

Figure 2.1

A) Schematic representation of an electron source in electron microscopy. Electrons are emitted from a heated filament (cathode), focused by the negatively charged Wehnelt cap, and accelerated toward the anode plate under positive potential, generating a focused electron beam; B) Types of scattered electrons generated during electron-sample interaction. When an incident electron beam strikes a thin specimen, various signals are produced, each providing different information depending on the microscopy technique used



Source: [73]

There are two main types of electron microscopy (EM): SEM and TEM.

- **Scanning Electron Microscopy (SEM)**: this type of EM takes advantage of the secondary electrons and the back-scattered electrons. The first ones are mainly emitted from the surface of the sample and provide information about surface topography and material properties. In the case of back-scattered electrons, they are the beam electrons reflected from the sample and their intensity changes depending on the atomic number.

  In SEM, a probe scans the surface of the sample in vacuum conditions and detects the previously mentioned types of electrons. As the information is gathered, an image of the sample's surface is constructed. This surface needs to be conductive for SEM to work, so samples with non-conductive surfaces need to be coated with conductive layers like gold.

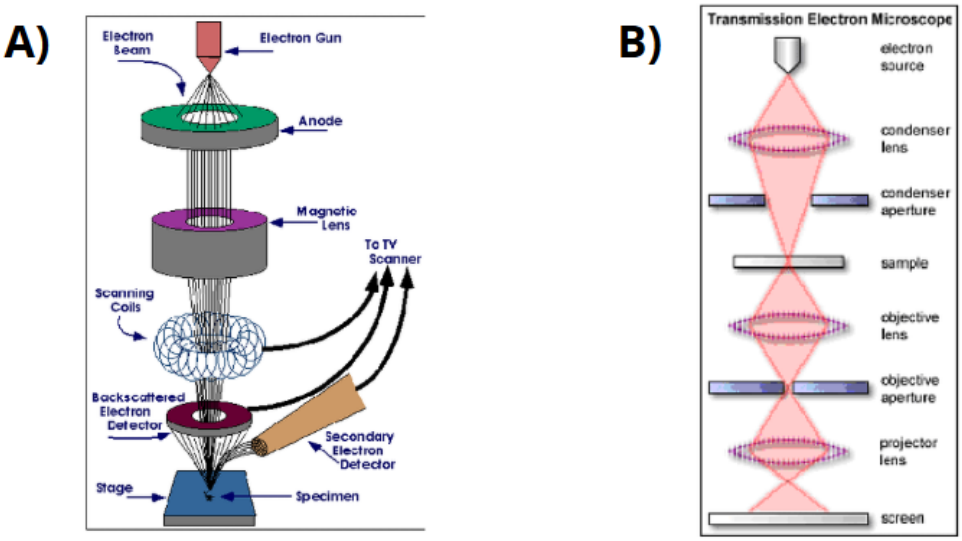
- **Transmission Electron Microscopy (TEM)**: this type of EM has detectors for transmitted electrons, the ones that do not scatter and cross through the sample; diffracted electrons, the elastically scattered ones; and the inelastically scattered electrons. The thicker areas of the sample will present fewer transmitted electrons than the thinner areas. Elastically scattered electrons (no energy loss) generate a diffraction pattern that reveals the atomic arrangement of the material of the sample. Finally, inelastically scattered electrons, the ones that lose energy after colliding,

provide some essential information about the sample bonds and crystallographic information in thicker samples.

TEM uses a focused electron beam in vacuum conditions that is directed to the sample. The transmitted electrons that go through the sample are then directed and magnified using magnetic lenses and different apertures to finally lead them to the screen or detector, where their information is gathered and a high-resolution image of the sample's composition is generated.

Figure 2.2

A) Schematic of the components of a Scanning Electron Microscope. The electron beam generated by the gun passes through magnetic lenses and scanning coils before interacting with the sample, generating secondary and backscattered electrons that are detected to form the image; B) Schematic of the components of a Transmission Electron Microscope. The electron beam passes through a series of electromagnetic lenses and apertures, interacting with the sample to produce transmitted electrons that form a high-resolution images on the screen.



Source: [73]

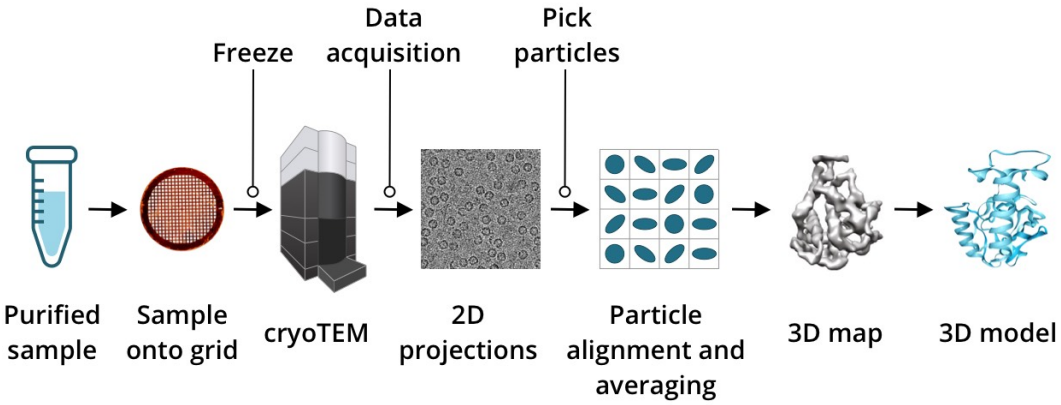
2.1.1. Cryo-electron microscopy

Although electron microscopy reaches unprecedented magnification, it still presents some limitations when working with biological samples due to the required vacuum conditions, which often disrupt the sample's structure. The solution to this problem is Cryo-electron microscopy. This novel technique consists of rapidly freezing biological samples to preserve their native structure and imaging them using an electron microscope. This process of rapid freezing is called vitrification and involves quickly cooling the sample in liquid ethane to form vitreous ice, preserving its natural structure without forming harmful

crystalline ice. The images obtained are afterwards processed and the final map of the molecule is obtained.

Figure 2.3

Typical workflow for cryo-EM data acquisition. The process begins with a purified biological sample frozen on a grid and imaged using a cryo-TEM. From the recorded 2D projections, particles are selected, aligned, and averaged to generate a 3D density map, which is then used to build the final atomic model.

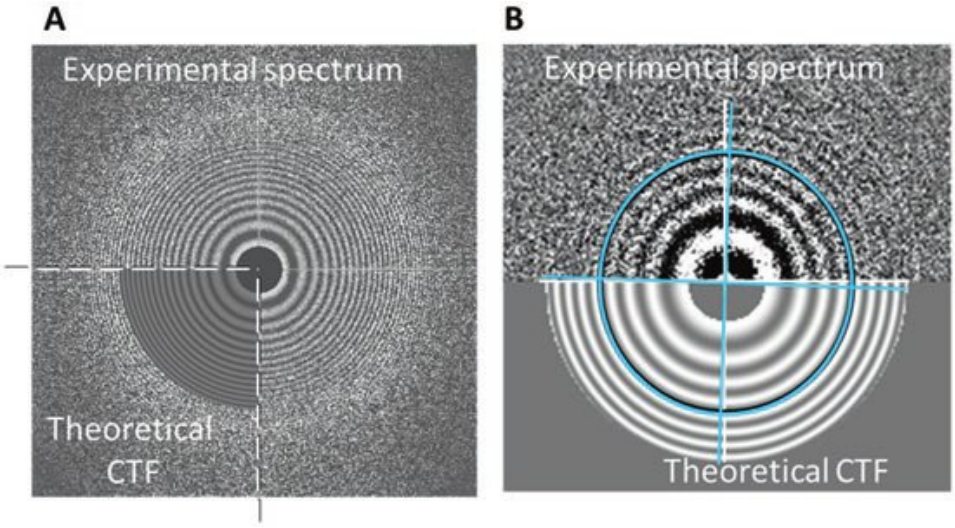


Source: [1]

When a Cryo-EM sample receives the electron beam, high-performance detectors produce movie data composed of different frames. Prolonged exposure to the electron beam produces damage to the sample. This is why during cryo-EM imaging, the detectors collect a series of frames of the same field of view of the sample, which are later combined to form a movie. This shorter exposure reduces damage but generates noisier images. The initial frames obtained after the sampling usually have poor contrast resolution due to low electron exposure, so periodic calibration of the camera gain is needed to obtain consistent images. Another noise source is the beam-induced motion, which is the blurring produced by the movement of the particles in the sample. By capturing the particles in slightly different positions in each frame and later processing this data to align and stabilize the data, the images produced have their noise reduced notably. This step is known as Beam-Induced Motion Correction. Once all the frames in a movie have been corrected in their gain and their beam-induced motion, they are all averaged into a single 2D image called a micrograph. This step increases the signal-to-noise ratio (SNR) as the random noise is averaged, which means it is reduced as the noise average tends to zero [57].

Each generated micrograph must now undergo Contrast Transfer Function (CTF), which describes how the microscope modifies the image based on spatial frequency. The main goal of this step is to correct the distortions generated by the microscope's optics, such as lens imperfections or defocus. By applying the CTF to the micrographs, different frequencies in the image are either amplified, attenuated, or inverted to account for these distortions.

Figure 2.4
Comparison of experimental and theoretical Contrast Transfer Funtion (CTF) in Cryo-EM Micrographs. The alignment between the experimental power spectrum and the theoretical CTF curves is essential for accurate image correction and resolution estimation in cryo-EM data processing.



Source: [11]

The next step of the sample processing is the Particle Picking. During this step, individual particles are extracted from the CTF-corrected micrographs. This particle selection can either be performed manually or using automatic algorithms [13]. From this last group, the algorithms applied can be differentiated into two large groups:

- Basic methods that require no training:
  This type of methods rely in particles with a generic size and shape to detect them using specific functions. Some examples are the DoG Picker [72], which applies the difference of Gaussian functions, or the LoG Picker [74], which relies on the Laplacian of a Gaussian to detect the particles. This type of methods are fast and simple to apply. On the other hand, due to their simplicity, they can generate false positives and false negatives when detecting particles on the micrographs.

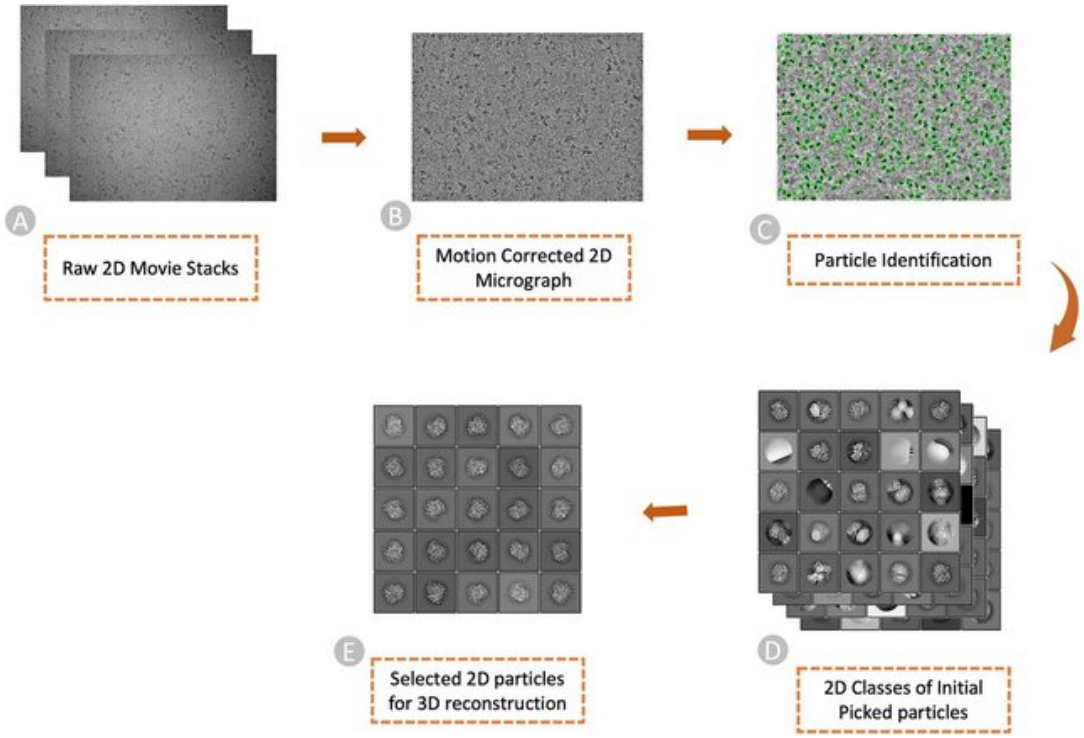
- Machine learning methods:
  These methods require the training of different models that learn using machine learning to detect the particles in the micrographs. An example of this type of methods is Xmipp Picker [12], a protocol integrated inside Scipion software that detects the particle positions using machine learning. The main advantage of this type of methods is the accuracy. However, they require previous training of the models and need more resources.

After the particle picking is finished, the next step is 2D classification, which consists of grouping the particles with similar shape and orientation, filtering artifacts such as

misaligned particles, false positives or false negatives generated during particle picking or ice contaminants. This step ensures that the image only maintains the high-quality and meaningful projections for the 3D reconstruction. Different models like CryoSPARC [37] or CL2D [70] can filter the particles and produce high-quality results.

Figure 2.5

Processing steps of cryo-EM micrographs prior to 3D reconstruction. Starting from raw 2D movie stacks (A), motion correction is applied to obtain a clean micrograph (B), followed by particle identification (C), classification of particle images (D), and selection of the best 2D classes for 3D reconstruction (E).



Source: [13]

Once the 2D classification is done, the initial 3D reconstruction is performed. This step generates a first 3D map (or volume) of the biological sample using the clean 2D-classified particles of the previous step. Algorithms like Xmipp RANSAC [71] are capable of constructing this initial volume, which is generated by taking the 2D projections acquired at unknown orientations and estimating the angles of the particles. After this initial three-dimensional reconstruction is generated, it needs to be classified into structurally homogeneous subgroups, each representing a potentially distinct 3D conformation. By doing this, the system is able to identify the differences between the different particle subsets and detect errors like incorrect particle alignment or mixed populations of particles.

The sample is now ready for the 3D refinement step, in which the accuracy of the particle alignment will be improved and a high-resolution 3D-map of it will be reconstructed. In this step, after selecting a homogeneous particle set from the 3D classification,

it is updated with new estimations of the orientation of each particle until convergence is achieved. Consequently, a high-quality map with mostly reduced noise is achieved, often offering close to atomic resolution.

After the 3D refinement, the reconstructed map might still appear blurred, with low contrast, or show some noise. This is why post-processing is usually needed. By applying some techniques such as the Fourier Shell Correlation (FSC) or the B-factor Sharpening, the interpretability and visual clarity of the final map are increased substantially.

The final step of the map processing is the validation, which consists in a series of tests that, even if they do not guarantee the map is correctly modelled, failing them would mean the failure in the reconstruction of the map. This validation process looks for several features inside the map, such as artifacts, poor angle estimation, misaligned particles, or low SNR.

After validation is finished, the map is ready to be used for model building and interpretation. Using the validated map, further analysis can be performed by building atomic models from it and studying its biological implications.

2.2. *De novo* Atomic Modelling. Generation and storage

After sampling a molecule using cryo-EM technology, a density map is obtained. This density map corresponds with the 3D visualization of electron density in the different regions of the biological sample. It contains only the surface of the biological sample represented as electron densities; it does not contain any atomic coordinates, residues, or protein information [60].

De novo atomic modelling makes use of these density maps and generates the corresponding atomic models using different techniques such as AI-based tools like ModelAngelo [23], DeepTracer [35], and AlphaFold-Multimer [16]. These tools detect the atomic positions inside the cryo-EM density maps and their residue and chemical information.

Once an atomic model has been generated, it must be stored in a format that keeps all the information generated: atomic positions, chemical identities, structural annotations, and any other created metadata. The traditional storage method is a PDB file, which is composed of text lines that store atom and sequence coordinates and other information such as the bonds between ligands [69]. However, the metadata stored inside these types of files is limited and not flexible, so they cannot store very large or complex assemblies.

2.2.1. Macromolecular Crystallographic Information files

Macromolecular Crystallographic Information File (mmCIF) is a standardized, text-based file format that solves PDB files limitations. The mmCIF files are able to store large, complex biological structures, including all the necessary metadata information [19].

CIF files contain detailed information such as atomic coordinates, chain identifiers, residue names, and experimental metadata, including resolution and software used during structure determination. This information is stored in a standardized way that is easy for humans to understand as well as computationally efficient. MmCIF format uses a key-value format. Each key is followed by its corresponding value. MmCIF files also allow the organization of information in loops that ensure that data remain organized and linked by column [4]. An example of this organization is shown in Figure 2.6. In it, the information of each atom in the model is stored following the loop information. In each atom, the columns will indicate each of the stored parameters: ATOM, number, symbol and extra information, the residue of which it is part of, coordinates and extra personalized information. These attributes stored inside mmCIF files will be particularly important for the development of this project, as they give much information that can be changed or modified inside the mmCIF without changing the 3D visualization of the biological structure.

Figure 2.6

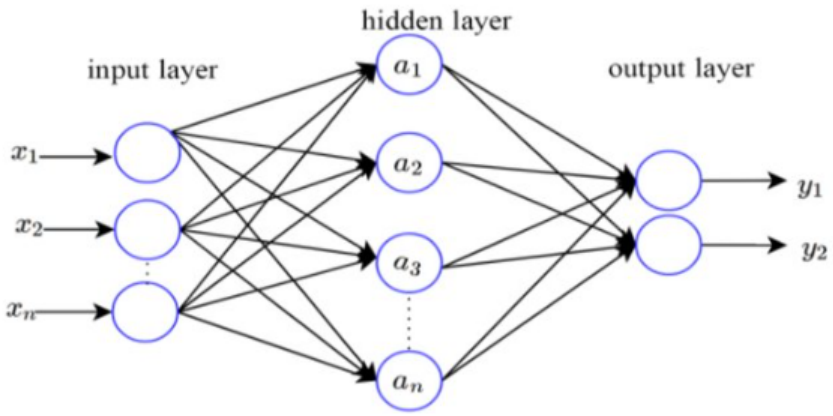
Atomic information example inside a .mmCIF file

```
1119 loop_
1120 _atom_site.group_PDB
1121 _atom_site.id
1122 _atom_site.type_symbol
1123 _atom_site.label_atom_id
1124 _atom_site.label_alt_id
1125 _atom_site.label_comp_id
1126 _atom_site.label_asym_id
1127 _atom_site.label_entity_id
1128 _atom_site.label_seq_id
1129 _atom_site.Cartn_x
1130 _atom_site.Cartn_y
1131 _atom_site.Cartn_z
1132 _atom_site.auth_asym_id
1133 _atom_site.auth_seq_id
1134 _atom_site.pdbx_PDB_ins_code
1135 _atom_site.occupancy
1136 _atom_site.B_iso_or_equiv
1137 _atom_site.pdbx_PDB_model_num
1138 ATOM 1    N N   . ASP A 1 24  132.919 121.533 160.000 A 24  ? 1.00 52.09 1
1139 ATOM 2    C CA  . ASP A 1 24  132.222 120.252 160.000 A 24  ? 1.00 50.15 1
1140 ATOM 3    C C   . ASP A 1 24  133.141 119.127 159.534 A 24  ? 1.00 46.79 1
```

2.3. Neural Networks. Convolutional Neural Networks

Neural networks are computational models that simulate the way human neurons work. They consist of a large number of nodes interconnected between them that collectively learn from a series of input data to generate the desired output. These nodes are organized in layers that extract information from the previous layer using a system of weights and bias to finally reduce the nodes to a final output. Figure 2.7 shows an example of how these nodes are organized.

Figure 2.7
Schematic of the layered organization of a Neural Network. The input is processed crossing through several layers until they converge into an output.



Source: [21]

There are several types of Neural Networks, each with its own architecture and suited for different tasks. Convolutional Neural Networks (CNN) in particular, have demonstrated their efficacy when working with image analysis, including cryo-EM image processing. Convolutional Neural Networks follow the same pattern as standard Neural Networks . The main difference is the way the data is handled. For an image of 12x12 pixels, a standard neural network would need 144 nodes, which is a viable number, but if the image size is increased to standard sizes such as 128x128 pixels, the number of nodes increases exponentially, and the neural network consumes large amounts of resources. CNNs address this issue by incorporating a unique architecture to handle image data and significantly reduce the number of parameters needed [32].

In the case of cryo-EM, Convolutional Neural Networks are a suitable solution for many of the issues that arise during sample processing. One example of the applications of CNNs in cryo-EM field is that images obtained with this microscopy technique often suffer from low SNR due to the low electron doses delivered to the sample to prevent damage on it. CNN models can be trained to reduce this noise, enhance contrast and improve the image quality [33]. Moreover, CNNs have also proved to be a suitable solution for atomic model generation. Using a trained CNN, it is possible to predict the coordinates of backbone atoms of a biological structure, providing the density map obtained in cryo-EM sampling [36].

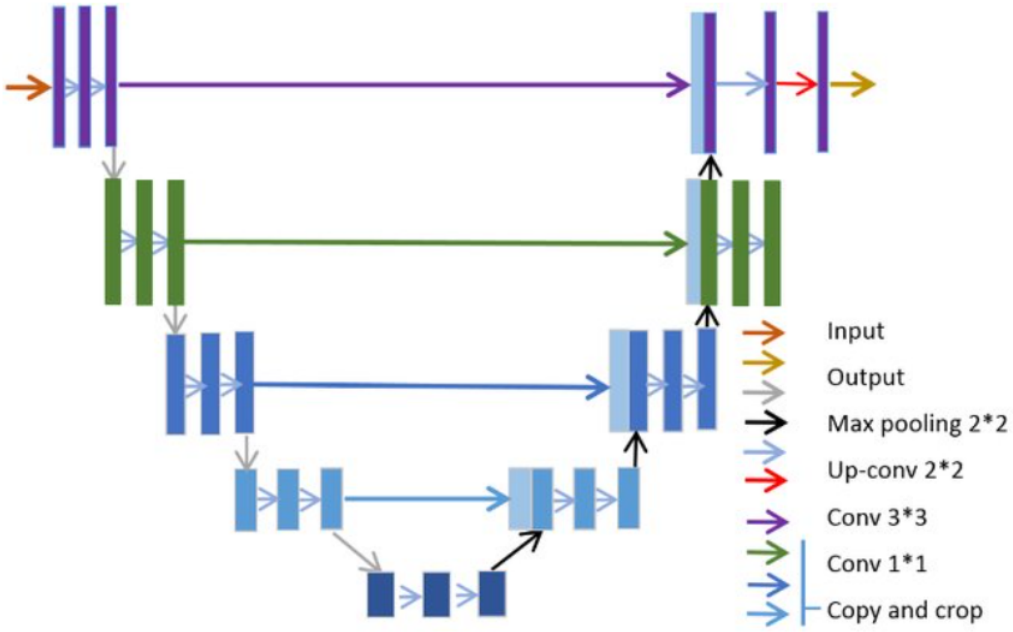
2.3.1. U-Net

A U-net is a particular type of CNN specifically designed for image segmentation. The architecture of U-Net follows a symmetrical encoder-decoder design. The encoder captures

and compresses image features using the Convolutional Neural Network layers. Meanwhile, the decoder decompresses these image features back to their original size. Moreover, the U-net model includes lateral connections between the encoder and the decoder so that high-resolution features captured in the encoder can be also used in the decoder stage, therefore improving the accuracy of the algorithm [42]. A scheme of the architecture of an U-Net can be seen in Figure 2.8.

Figure 2.8

Architecture of the U-Net model for image segmentation. The network consists of a contracting path (left) for feature extraction and an expansive path (right) for precise localization, with skip connections (blue arrows) that combine low-level features with upsampled outputs to enhance segmentation accuracy.



Source: [14]

U-Net models are particularly useful in biological applications because of their great performance even when the training data is limited [2], which is a common case in the microscopy field, as data is expensive to generate and models need to be trained with a reduced amount of samples. Moreover, U-Net approaches to cryo-EM processing have demonstrated their efficacy when working with low SNR images. Due to the skip connections the algorithm have, the fine features of the samples are preserved, minimizing loss of resolution and making it useful for enhancing structural information [50]

3. MATERIALS AND METHODS

This section is intended to explain all the tools and software necessary to understand the work done during the development of the project.

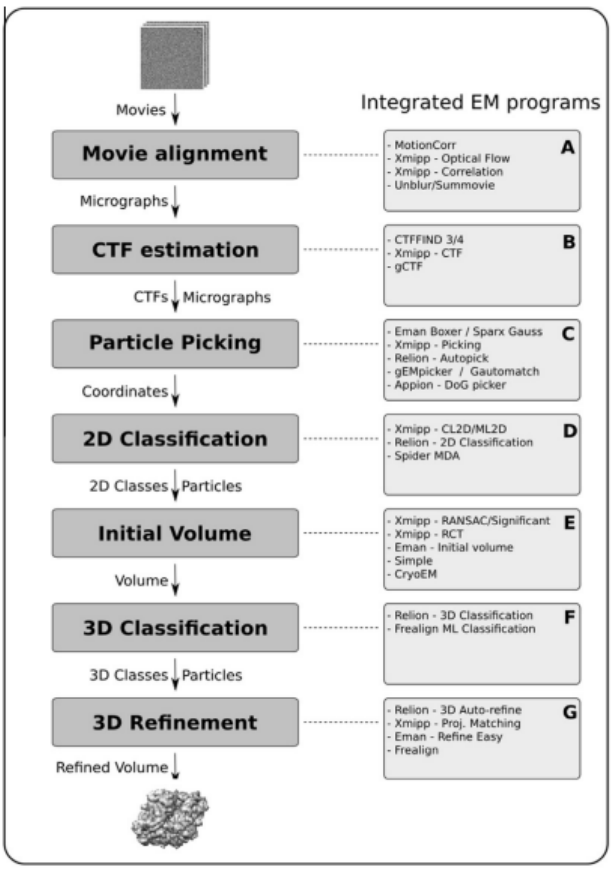
3.1. Scipion Framework

Scipion is an open-source image processing framework developed by the Biocomputing Unit of the National Centre of Biotechnology (CNB) [29]. It focuses on cryo-EM image and modelling processing and it integrates already created software into an intuitive and reproducible interface. Inside Scipion, the user can create workflows that combine different packages without having to switch between different software. Each software integrated in Scipion has its corresponding plugin, making Scipion framework a flexible environment where users can only add the packages they need. Scipion is also written in Python coding language, and so are its plugins, with very few examples. As Python is one of the most widespread programming languages in the world, it facilitates external users of Scipion to integrate their own plugins inside the framework, helping Scipion to be in constant evolution and development.

In Scipion, every software executed asks the user for a series of parameters that are stored for later recall, and a fixed output is generated [47], this output can serve as the input for the next method, creating a mesh that helps save the path of action that has been done to reach a certain result, improving traceability. Moreover, the coexistence of several packages with similar functioning helps the user to easily compare different results inside the same engine.

Figure 3.1

General image processing workflow for single particle analysis. The workflow includes sequential steps from movie alignment and CTF estimation to particle picking, classification, initial volume generation, and final 3D refinement. Each step is supported by different integrated EM software tools commonly used in cryo-EM processing.



Source: [43]

Figure 3.2

Scipion interface example, zoomed regions: A) Section where the protocols installed can be selected to be executed; B) Working region where the executed workflow is shown and the user can monitor its execution; C) Section where each individual protocol can be analyzed to see its input and output or any other details of the protocol execution.

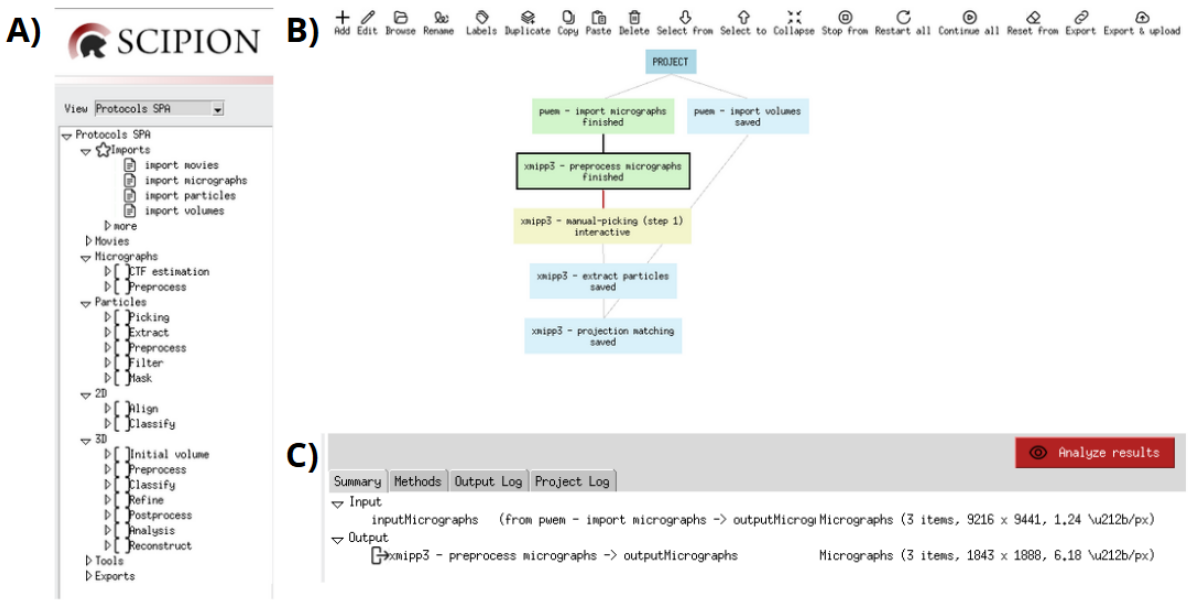


Figure 3.2 shows an example of how Scipion's interface most important sections look like. Three sections can be clearly differentiated in Scipion's interface:

- The leftmost section, corresponding with image A in Figure 3.2, contains and organizes all the available Scipion protocols (programs). They are classified in sections for the user to find them as fast as possible

- The upper-center region, corresponding with image B in Figure 3.2, is the working region of the interface. In it, the user can see and trace all the protocols they have executed and where do they come from.

  Scipion workflows are organized in modules that appear as squares or boxes that correspond to each protocol executed and are linked between them if the input of one of the programs is the output of another. When the box is yellow, it means it is being executed. If the box turns green, it means the protocol has been executed correctly. On the other side, if it turns red, it means it has failed, helping the user to easily interpret their results.

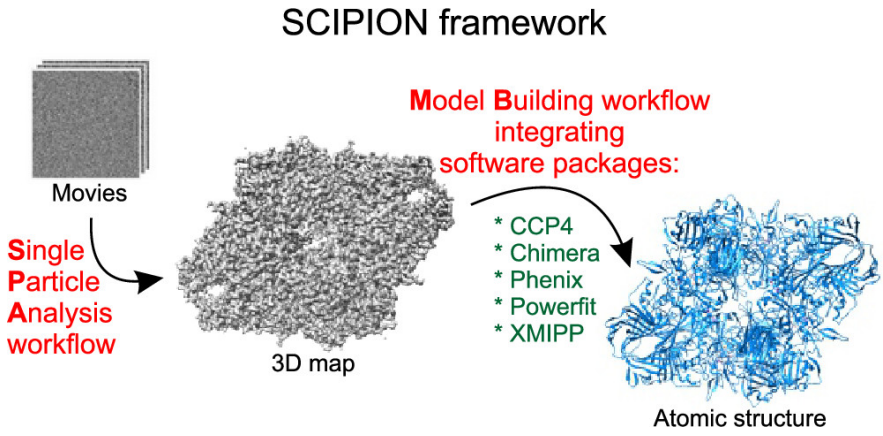
- The down-center window, corresponding with image C in Figure 3.2, is the section dedicated to each protocol execution. Inside it, the user can check all the information regarding the software performance: internal execution messages, the inputs and outputs, or the project log. In addition, there is a red section called "Analyze results" where the user can easily see the protocol's results in integrated viewers like ChimeraX or Xmipp.

3.1.1. Xmipp framework

If Scipion framework is mentioned, Xmipp must be too. Also developed by the Biocomputing Unit of the National Center for Biotechnology in 2013, it is a software package that gathers different software for cryo-electron microscopy image processing [58].

Initially, Scipion was a modification of Xmipp graphical user interface, but it evolved into a separate project in 2018. Since this happened, Xmipp provides the core functionalities inside Scipion framework, with several protocols for nearly all steps in the single particle analysis (SPA) processing pipeline

Figure 3.3
Overview of the SCIPION framework for structural biology workflows.



3.2. CryoTEN software. Effective enhancement of cryo-EM density maps

As explained before, cryo-electron microscopy enables the magnification of EM to be used in biological samples. However, the acquisition of images using this technology is often limited by noise and missing density values in the resulting density maps. This usually results in lower quality maps because of the lack of contrast, which results in worse reconstruction of the protein structures analyzed.

To address the challenges associated with cryo-EM map sharpening, several deep learning-based methodologies have been developed for fully automated enhancement. One of them is DeepEMhancer, which sharpens cryo-EM density maps using a U-net[45]. However, this software does not provide successful results when low-quality maps are presented, making it inconsistent. Other approaches like EMReady [20] or EM-GAN [28] are trained using simulated maps produced from known protein structures. CryoTEN continues this last approach to present a consistent and minimum consumption map enhancer model.

CryoTEN is a 3D transformer-based deep learning model designed to enhance cryo-EM density maps. It has demonstrated effective enhancement of the density maps, improving their quality. Moreover, built atomic models from CryoTEN-processed maps

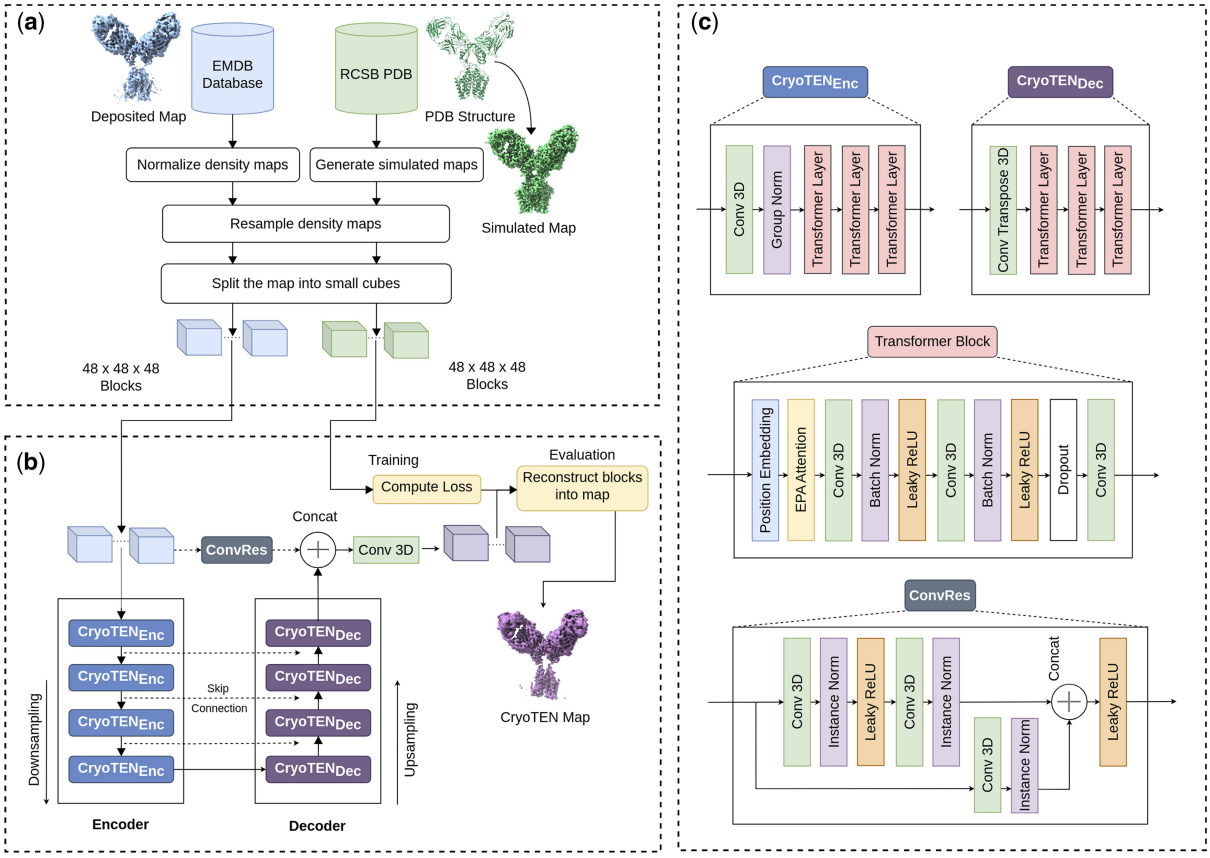
proved to have higher quality than those generated from density maps without being processed by CryoTEN [49].

To train CryoTEN, the authors used 1295 cryo-EM maps along with target maps that were simulated using a Gaussian function from known protein structures. These target maps act as a reference from which the model can learn. First, both the real and simulated maps were resized to use a consistent 1 angstrom grid and normalized so their values range from 0 to 1. Because the maps are very large, they were split into smaller overlapping 3D blocks (cubes). Only blocks that contained protein structures were used. During testing, maps were divided into evenly sized blocks with an overlap to ensure full coverage.

CryoTEN works as a transformer-based neural network built in a UNet-style shape [51], with four layers each of encoders and decoders. Encoders compress the 3D data while decoders rebuild it, and skip connections pass detailed information directly from the encoder to the decoder to avoid losing key features. Each block also goes through a ConvRes unit, a special module that helps preserve early-layer information with shortcut connections. Inside the encoder and decoder layers are transformer blocks that use a smart mechanism called Efficient Paired Attention (EPA) to focus on both spatial patterns and feature channels. This design helps CryoTEN enhance noisy cryo-EM maps efficiently, with less memory and much faster speed than other deep learning approaches. The full procedure can be seen in Figure 3.4.

Figure 3.4

Overview of CryoTEN data processing overview.(a) Training data preparation involves normalizing and resampling deposited EM maps and generating simulated maps from PDB structures, both split into 48º voxel blocks. (b) The CryoTEN model follows an encoder-decoder architecture with ConvRes blocks and skip connections, trained to reconstruct enhanced maps from noisy inputs. (c) Detailed architecture of CryoTEN components, including the encoder and decoder modules, transformer blocks, and convolutional layers

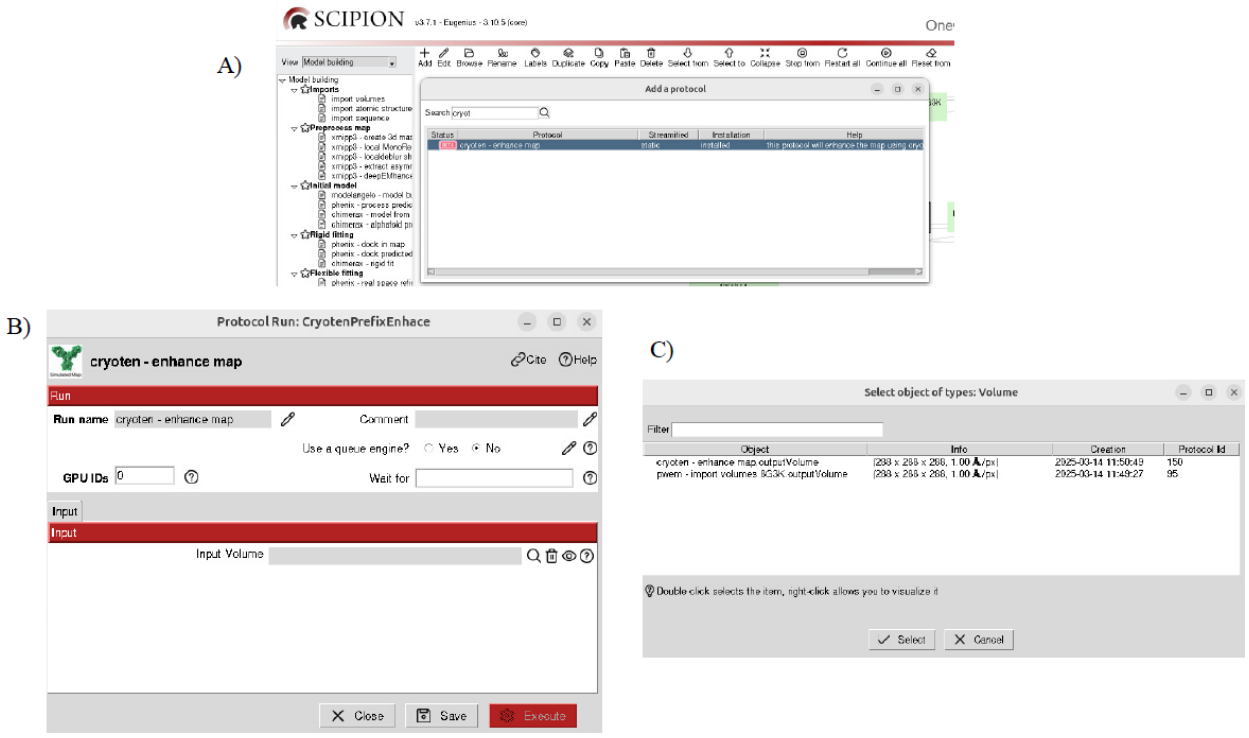


Source: [49]

CryoTEN software is open and available on GitHub [48]. It can be freely downloaded and executed locally. In this project, a plugin to implement this software inside the Scipion environment was created and it is also freely available on GitHub [41] [40]. Once the user installs the CryoTEN plugin, all the necessary software is automatically installed inside the Scipion software location. Then, the user can freely access CryoTEN software directly from the program interface, as it is shown in Figure 3.5.

Figure 3.5

A) CryoTEN installed as a plugin in Scipion ; B) Protocol "Enhance" interface ; C) Parameter volume selection interface.



As seen in Figure 3.5, the protocol only asks the user to select the volume they want to send to CryoTEN software. After the user introduces the input, the protocol internally activates the necessary environment for CryoTEN software to work and sends this input density map to the software, establishing the output path as the same as Scipion's protocol. This way, after CryoTEN software finishes the enhancement of the map, the user will automatically see it as the output of the protocol, which is ready to be used for later processing.

In addition to the development of the installation and running protocol code, a test was designed for the user to try the newly installed plugin and check that everything runs correctly. This test can be run with the shell command: *scipion3 tests cryoten.tests.tests_cryoten.TestCryoten*. When called, Scipion automatically downloads a sample biological structure, PDB-9GRD [55] [56]. Then, it calls CryoTEN's newly installed protocol and runs it. If an output result is generated, the test has been successful and CryoTEN protocol is ready to be used.

With this newly integrated plugin in the Scipion framework, the first steps of molecule processing are significantly improved. The user is provided with a simple protocol that automatically denoises the biological structure introduced, which results in better atomic model generation and more accurate measurements.

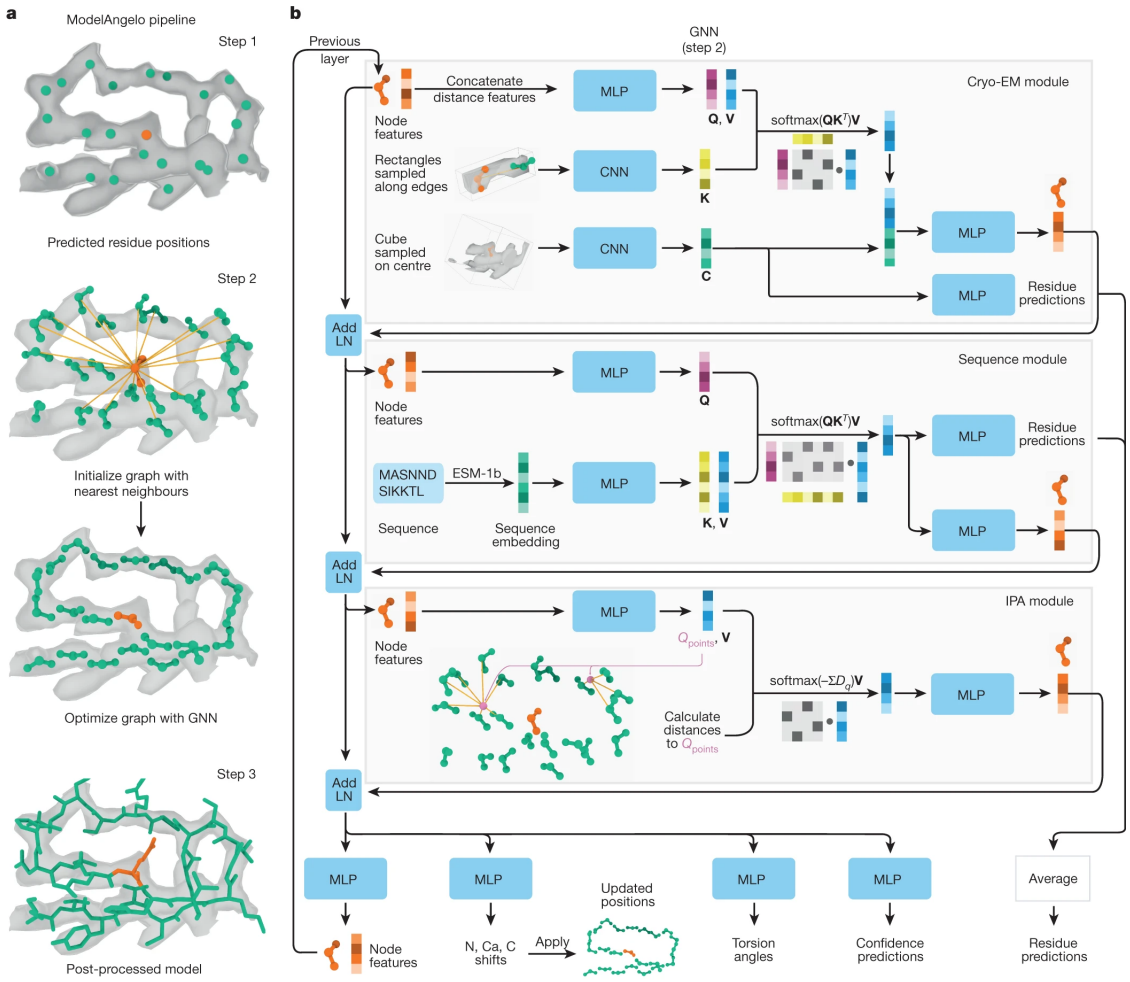
3.3. Modelangelo. Atomic model prediction

With the latest advancements in cryo-electron microscopy, the resolution of the molecules analyzed has been improved exponentially. Traditionally, atomic modelling has been done manually from the cryo-EM density maps using three-dimensional computer software. However, this procedure is time-consuming and requires high levels of expertise to generate useful and accurate atomic models.

Modelangelo is an automated machine-learning software that builds atomic models combining cryo-EM density maps, protein sequence data and structural geometry. It outperforms human experts in identifying unknown proteins and builds atomic models nearly as complete as those created manually [23].

Figure 3.6

Modelangelo data processing overview.(a) The Modelangelo pipeline begins with the prediction of residue positions from cryo-EM density maps, followed by graph initialization based on nearest neighbors and graph optimization using a Graph Neural Network (GNN). (b) The architecture combines three modules: the cryo-EM module, which processes spatial features; the sequence module, which embeds amino acid information; and the IPA module, which integrates positional information. These modules work together to update node features, predict residue identities, atomic positions, torsion angles, and associated confidence scores.



Source: [23]

As it can be seen in Figure 3.6, Modelangelo generates atomic models using a three-step procedure:

1. A convolutional neural network (CNN) identifies likely positions of protein backbone alpha-carbon atoms of the structure. These atoms are used to construct an initial graph where each residue is a node, connected to its nearest neighbours.

2. This initial graph is refined using a Graph Neural Network (GNN) and compared with the input protein sequence to improve the accuracy of residue positions and

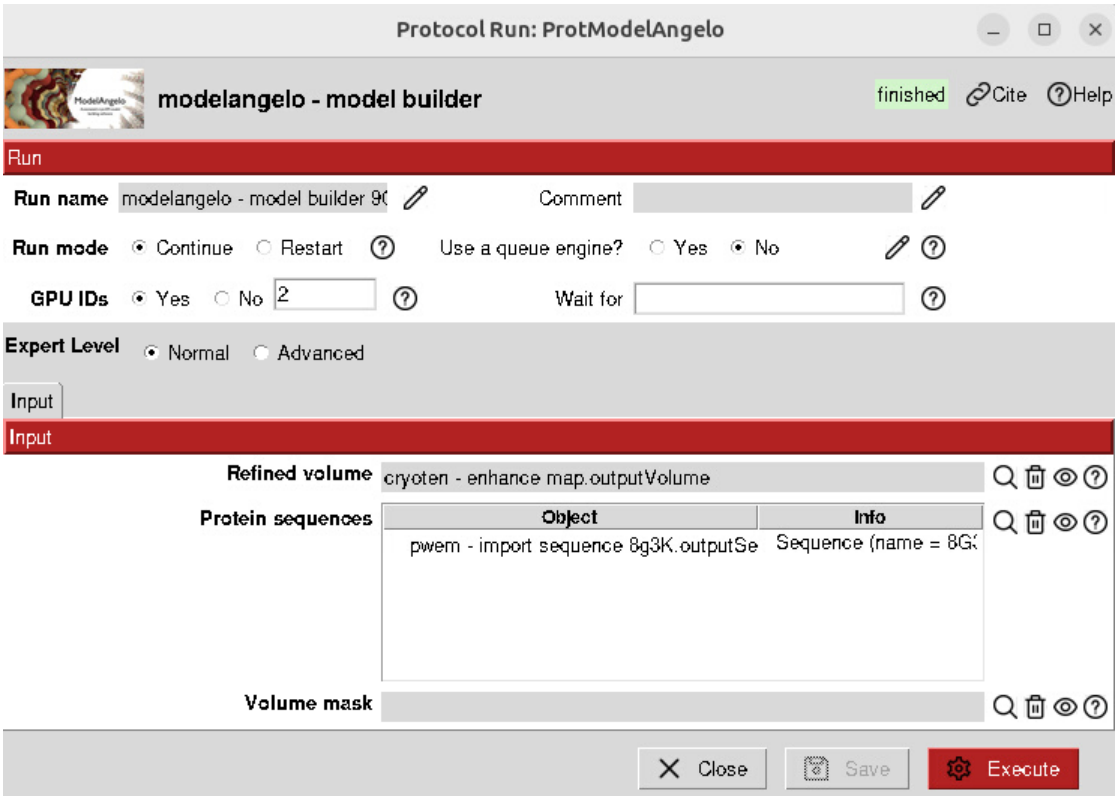
orientations. This refinement is performed using three modules: cryo-EM module, sequence module and IPA module.

3. In the final step, the updated residue information is used to predict full atomic details, such as angles, confidence scores, and residue types.

Modelangelo is fully integrated inside the Scipion environment, and it has its corresponding plugin, which can be found in GitHub [67]. The interface of Modelangelo's protocol in Scipion can be seen in Figure 3.7. It requires the user to introduce the following parameters:

- Refined volume: the cryo-EM density map to be modelled. In this case, the models used will come from cryoTEN plugin
- Protein Sequences: optional parameter for the sequences that correspond to that density map. They are used to match predicted residue positions to the actual amino acid sequence.
- Volume mask: optional parameter that helps the user focus modelling on a specific region of the map

Figure 3.7
Modelangelo Protocol interface in Scipion.



3.4. Kiharalab. DeepMainMast protocol to find Atomic Models Backbones

KiharaLab, based at Purdue University and led by Professor Daisuke Kihara, is a research group specializing in computational biology and bioinformatics, with a strong focus on structural biology [3]. The lab is known for developing innovative algorithms and software tools that assist in the interpretation of biomolecular structures, particularly in the context of cryo-EM. By combining techniques from machine learning, image processing, and structural modelling, KiharaLab aims to overcome challenges in analyzing complex or low-resolution cryo-EM data. Most of their projects have been implemented in the Scipion environment under the plugin scipion-em-kiharalab [17]. In this work, the protocol DeepMainMast, developed by Kiharalab and implemented inside Scipion, will be used to generate atomic models.

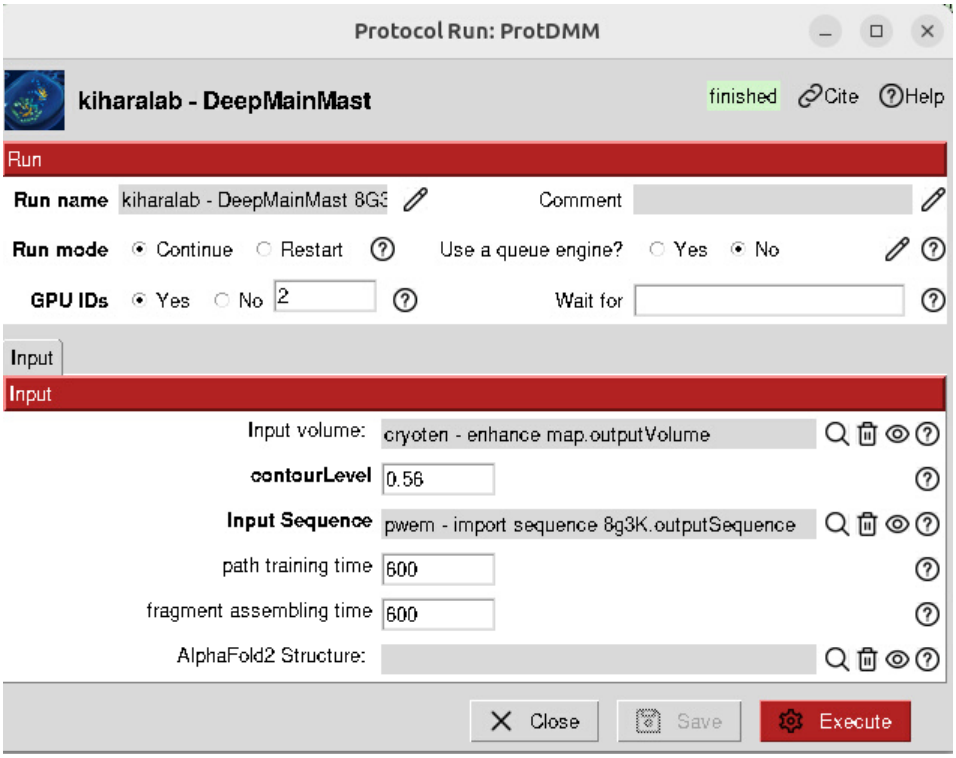
3.4.1. DeepMainMast

DeepMainMast (DMM) is a method that builds a full 3D model of a protein's alpha carbons directly from a high-resolution cryo-EM map, without needing a known structure. It uses deep learning to recognize patterns in the map that match amino acids and atoms, helping to trace the protein's backbone. It can also work together with AlphaFold2 to improve accuracy. This protocol can be summarized in 6 steps [59]:

1. Detect atom and amino acid types using a deep learning model.
2. Tracing the alpha carbon backbone path and matching it to the protein sequence.
3. Assembling alpha carbon fragments and resolving chain identities.
4. Combining models created with different parameters to improve accuracy.
5. Building and refining full-atom models.
6. Scoring the models based on structure quality using DAQ(AA) and DOT scores.

The interface of the protocol DMM inside Scipion environment can be seen in Figure 3.8.

Figure 3.8
DeepMainMast Protocol interface in Scipion.



The parameters that the protocol asks the user are:

- Input volume: the density map the user wants to use to generate the atomic model.
- Contour Level: it determines the threshold at which the density is considered significant.
- Input sequence: protein sequence file that will be used.
- Path training time: parameter controlling the duration of path training.
- Fragment assembling time: duration or iterations for the fragment assembly step.
- AlphaFold2 Structure: AlphaFold2-predicted structure to assist with model building.

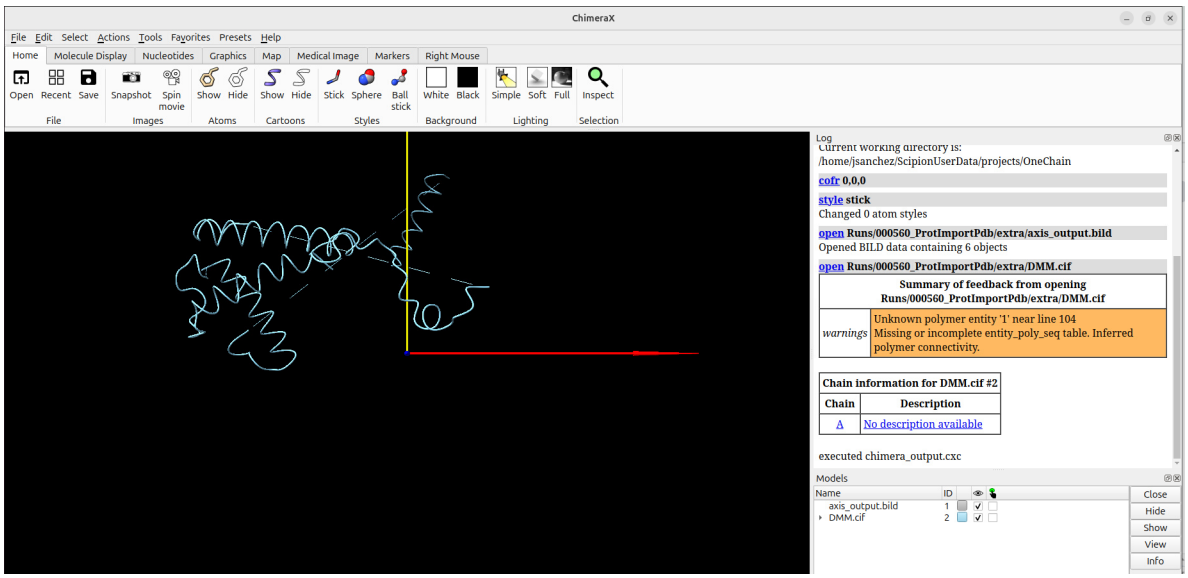
3.5. ChimeraX

ChimeraX is a next-generation molecular visualization tool developed by the UCSF Resource for Biocomputing, Visualization, and Informatics. It is widely used to analyze and

display atomic models, protein structures, and cryo-EM density maps in an interactive and high-resolution environment [18].

Figure 3.9

ChimeraX interface inside Scipion environment.



ChimeraX is free for noncommercial use and can be installed in Mac, Windows, and Linux environments [34]. It is fully integrated inside Scipion software and can be used to visualize 3D structures quickly and intuitively, as shown in Figure 3.9. In this work, ChimeraX was used to visualize and validate the output structures generated by cryo-EM processing tools (e.g., DeepMainMast, and Modelangelo).

Moreover, ChimeraX has its own plugin in Scipion (scipion-em-chimeraX), where the user can use several protocols to perform several operations inside ChimeraX software automatically. Some of these protocols are "chimeraX - operate", which allows the user to access ChimeraX and save the result in the Scipion framework; or "chimeraX - contacts", which identifies atomic contacts based on van der Waals radii [66].

3.5.1. ChimeraX Scipion plugin - Protocol Find Discrepancies

In this project, a new protocol was coded and added to the list of available protocols: chimeraX - find discrepancies. This protocol aims to process different atomic models, whether they are AI-generated or directly downloaded from a database and generate a visual result using ChimeraX software, where the user can visualize the regions of highest confidence in all introduced models [66] [39]. This new protocol makes use of the following ChimeraX internal commands:

- *matchmaker*: this command performs a pairwise alignment of two protein structures and then fit these two structures by adjusting the residues' location so that they coincide similarly to the alignment [62].

- *sequence header n attribute save \path*: by using this command, the "attribute" selected of the model "n" will be saved in the path specified [63]. In this work, the attribute saved will always be RMSD (root mean square difference) of the alignment, which will be explained now.
- *setattr x*: this command helps to set the attribute "x" to any specified value, which will be very useful for the visualization of the results [64].
- *color*: this command colors the desired target (e.g, residues) in a certain palette so that the user can visually differentiate them based on a specific attribute or grouping [61].

The chimeraX - find discrepancies protocol is fully coded in Python and it can be found on the corresponding GitHub repository [66] [39]. It can be explained by following these steps:

1. As seen in Figure 3.11, the user can introduce any number of atomic models to be processed. The first step of the protocol is to establish pairs between them to perform later alignment so that all the models are aligned with all the rest.
2. A ChimeraX script is automatically created. In this script, all the models are pairwise aligned using *matchmaker* and their aligned models are stored. In addition, all the residues' attribute occupancy is set to an arbitrary value of 1111.11 to later process it. Finally, the RMSD values of the residues in each alignment are also stored in another text file.
    a. Root Mean Square difference (RMSD): it is a common measure used to quantify the average distance between residues of aligned protein structures. It is based on a simple distance formula:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\vec{r}_i^{\,\text{model1}} - \vec{r}_i^{\,\text{model2}}\right\|^2} \tag{3.1}$$

    N is the number of atoms in each residue, and $\vec{r}_i^{\,\text{model1}}$ and $\vec{r}_i^{\,\text{model2}}$ are the 3D coordinate vector of each atom in both models. Because of how the *matchmaker* command works, this N will always be one, as the alignment will be performed taking only the alpha-carbon of each residue. As a consequence, the complete RMSD of every atom in the atomic model is not computed and, thus, some information is lost. However, this approach significantly reduces the operations performed and its computational complexity while still maintaining a good overall fit between the structures.

    Because RMSD measures how close the residues are, if the value between two aligned residues from different models is close to zero, it indicates that both models predicted the same atom at the same location, implying high

confidence. In contrast, a high RMSD value suggests a significant discrepancy between the two models' predictions. Therefore, this attribute can be used as a quantification of the discrepancy regions in each atomic model

3. This ChimeraX script is then executed internally, so that the alignments between atomic model pairs are done and the corresponding results are generated.

4. The next step is to join the RMSD values obtained in a separate file into each of the models. An example of this file can be seen in image A of Figure 3.10. To do this, the plugin adds the RMSD value of each residue as a list of scipion attributes at the end of the .cif file of each model. Moreover, the arbitrary number in the occupancy slot of each atom is also changed to the real RMSD value that corresponds to their residue. To do this, a FASTA file is also downloaded during the execution of the ChimeraX script that provides the residue position in each alignment. An example of this file can be seen in image B of Figure 3.10. This file helps the code to interpret the relative positions of the residues once the alignment has been done. For instance, the first residue of an atomic model can occupy the third position in the alignment if there are two gaps before it.
   In the example alignment that can be seen in Figure 3.10, the alignment RMSD values start at residue 24 of the alignment, which corresponds with the residues "PI" that are aligned in both atomic models, which demonstrates the correct functioning of the code.

5. At this point, there is a total number of n! alignment results, where n is the number of atomic models introduced as input by the user. It is needed to reduce this amount of information to give a useful result. To do this, each unique model is sent to a folder called "FINAL-OUTPUTS". In this folder, all models corresponding to the same input atomic model have their RMSD values averaged, and a final output model for each input one is generated.

6. The output obtained consists of the same number of models introduced by the user but with their RMSD values updated. Consequently, these models have the information about the discrepancy regions between them.

7. When going to "Analyze Results", ChimeraX software is opened and the generated models are shown and their residues are colored in *paegreen* [61] palette using *color* command. This way, the user can easily see and compare the different models and the regions of highest confidence (lower RMSD - darker green) with the regions that might not be correct, as their confidence is lower (higher RMSD - whiter color)

Figure 3.10
Example files of the RMSD values and FASTA sequences after the ChimeraX script has been executed.

A)

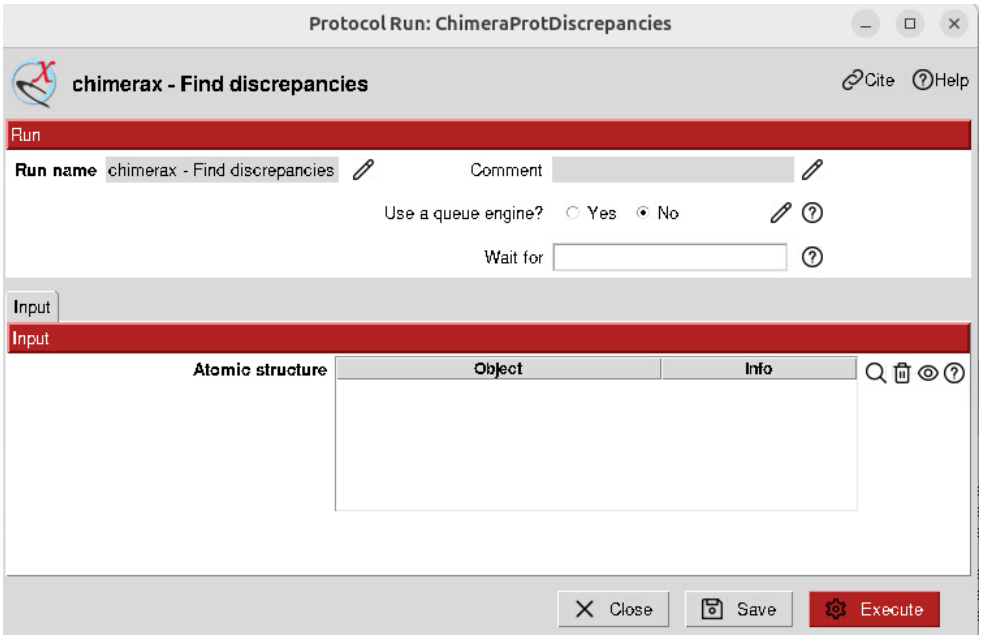
```
Cα RMSD header for 1
1: None
2: None
3: None
4: None
5: None
6: None
7: None
8: None
9: None
10: None
11: None
12: None
13: None
14: None
15: None
16: None
17: None
18: None
19: None
20: None
21: None
22: None
23: 54.1136247134619
24: 50.75074098863813
```

B)

```
>extra.cif, chain Aa
.....................PIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDI
YKIMGEIGSKGKIEGISEERIKWLAGLIERYSEMVNKLSFVLPGGTLESAKLDVCRTIAR
RAERKVATVLREFGIGTLAAIYLALLSRLLFLLARVIEIEKNK............

>DMM.cif, chain A
TTKVGDKGSTRLFGGEEVWKDDPI.............IGEAKHYVDEEMKGILEEIQNDI
YKIMGEIGSKGKIEGISEERIKWLAGLIERYSEMVNKLSFVLPGGTLES..LDVCRTIAR
RAERKVATVLREFGIGTLAA..............VIEIEKNKLKEVRSHHHHHH
```

Figure 3.11
ChimeraX - find discrepancies Protocol interface in Scipion.



With the development of this plugin, the process of comparing two different atomic models of the same biological structure is significantly simplified in Scipion. By using "ChimeraX - find discrepancies", users can perform this process in a reproducible and automated way without executing any code inside external software like ChimeraX. Everything is executed internally, the visual output is displayed, and the atomic models are modified to quantify these results.

3.6. Databases and Structural Repositories

This section is intended to explain the Databases used during the development of the project and their functionalities.

3.6.1. Protein Data Bank

The Protein Data Bank (PDB) is the largest global database dedicated to store the 3D structures of biological molecules like proteins or DNA [5]. This database is freely available online globally, which makes it the leading global repository of experimental data essential to scientific discovery. There are several Protein Data Banks around the world that, together, form the Worldwide Protein Data Bank (wwPDB) [75]. This organism was formed to maintain a single PDB archive of macromolecular structural data. This way, each structure published in any PDB around the world has its unique PDB ID that corresponds to a Digital Object Identifier (DOI) that is shared between all the PDBs in the world.

In each PDB entry, the user can obtain all the information about the corresponding biological structure, such as the sequence or the experimental data (e.g, technology used to capture the structure, resolution). During this work, PDB Database was widely used to obtain different protein structures and download the needed files, like the FASTA file that contains the sequence of the protein needed for protocols like DeepMainMast or Modelangelo [38].

3.6.2. Electron Microscopy Data Bank

The Electron Microscopy Data Bank (EMDB) is a public repository that stores 3D cryo-EM structures of biological structures [10]. In this online database, the user can search for specific techniques inside cryo-EM tomography, such as single-particle analysis or electron crystallography. As of 7 May 2025, the EMDB holds 45,446 entries. In this database, the user can directly download the raw maps of the 3D structures before biological interpretation, so they are just the "surface" of the molecule without any information about their components.

EMDB entries include not just the 3D maps, but also metadata like resolution, sample preparation details, imaging conditions, and software used. Many entries are linked to PDB atomic models, allowing users to explore both the raw density and the interpreted structure [15].

4. RESULTS AND DISCUSSION

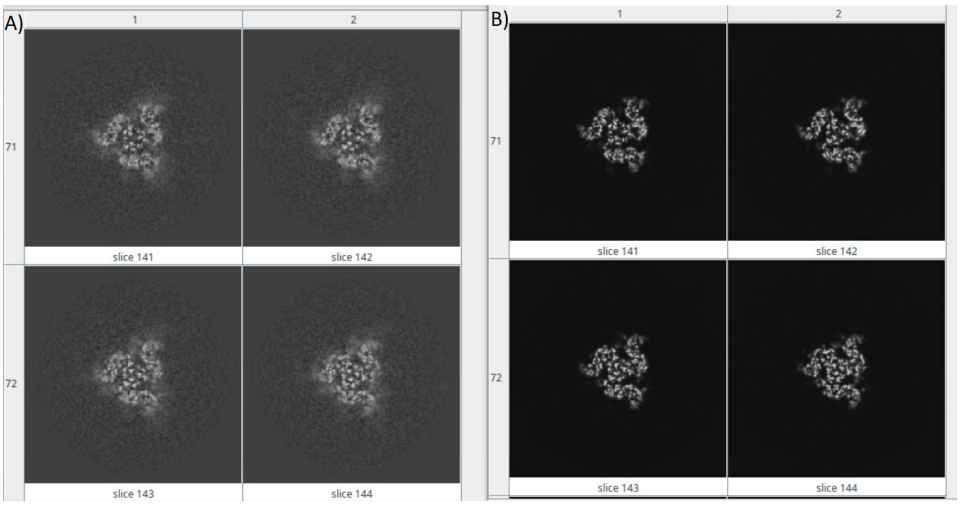
4.1. CryoTEN

4.1.1. Density Map enhancement

After the coding of the new plugin for CryoTEN in Scipion, the user can download the plugin from GitHub and start using it inside their Scipion workflows. The results after processing a map CryoTEN can clearly be seen at plain sight. To illustrate this, it will be used as an example the molecule PDB-9BEA [54], which corresponds to the structure of the SARS-CoV-2 spike protein. To use it, its density map was imported from EMD-44475 [53].In Figure 4.1, the improvement performed by CryoTEN software is clearly visible. The original volume (A) displays a high level of background noise and less-defined contrast, while the CryoTEN output (B) shows significant denoising and sharper features, reducing the salt-and-pepper noise in the original density map and enhancing the contrast difference between the molecule and the background, helping the user to differentiate them. In this figure, some sample slices are shown that can be compared, and the following conclusions can be reached:

- CryoTEN effectively removes low-intensity background while preserving core structural information.
- This improvement can most likely lead to better model fitting.

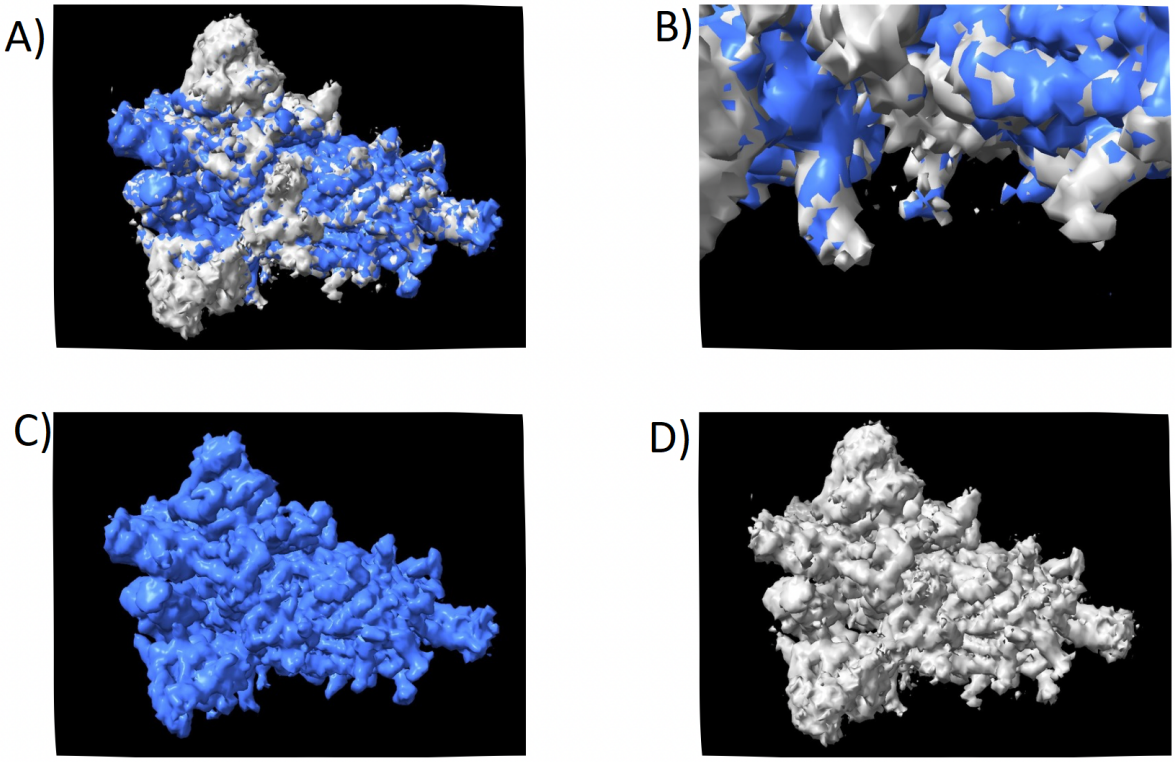
Figure 4.1
Comparison of original and enhanced density maps for PDB-9BEA. (A) Raw density map downloaded from the Protein Data Bank (PDB). (B) Same map after enhancement using CryoTEN software integrated into Scipion, showing improved contrast and reduced background noise.



The results are also visible in 3D visualization, as can be seen in Figure 4.2. From this figure, it can be stated:

- The CryoTEN-enhanced map (blue) shows sharper and more connected densities, especially in peripheral regions. The original map (white) appears more fragmented in these areas
- The original map (C) exhibits a noisier texture, with irregular surface detail. In contrast, the CryoTEN map (D) appears smoother, with fewer background artifacts.

Figure 4.2
3D visualization of the results of CryoTEN software: A) Overlap between the original and CryoTEN density maps; B) Detail of the overlap; C) 3D density map resulting from CryoTEN software; D) Original density map downloaded from EMDB.



These results demonstrate the effectiveness of CryoTEN software as it improves the quality and resolution of the input EM maps. The enhanced maps show clearer structures in both 2D and 3D, which make them more viable for later processing such as atomic model generators (e.g, Modelangelo)

4.1.2. Atomic model generation

After comparing density maps, atomic models should also be compared. To test if CryoTEN software results in better atomic model generation, the biological structure PDB-

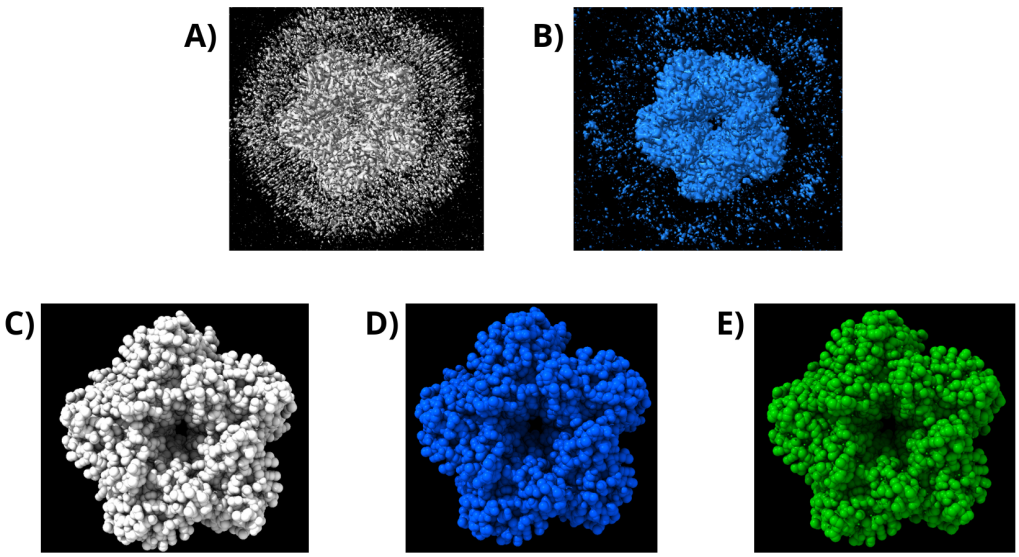
8R0O [25] will be used. This structure was specially selected as its EMDB density map [24] contains large amounts of noise as it can be seen in images A-B on Figure 4.3. After having it processed by CryoTEN software, both the original density map and the resulting one from CryoTEN were sent to Modelangelo to generate the corresponding atomic model. These models were each compared with the database atomic model directly downloaded from PDB as it can be seen in images C-E on Figure 4.3. By using *matchmaker* command in ChimeraX, the average RMSD of the alignment was obtained:

- Cryoten atomic model yielded an overall RMSD of 0.189 angstroms.
- The model generated from the database density map resulted in an overall RMSD of 0.199.

As a result, CryoTEN demonstrated to improve cryo-EM density maps so that the atomic models generated from them are more accurate. This improvement, although small in absolute value, proves a measurable enhancement in the atomic model, improving its accuracy and effectiveness to remove noise in cryo-EM data.

Figure 4.3

Visualization of enhancing structure PDB-8R0O in ChimeraX: A) Density map downloaded from EMDB database; B) Density map enhanced by CryoTEN plugin; C) Atomic model generated by Modelangelo from the database density map; D) Atomic model generated by Modelangelo from the CryoTEN density map; E) Atomic model downloaded from PDB database.



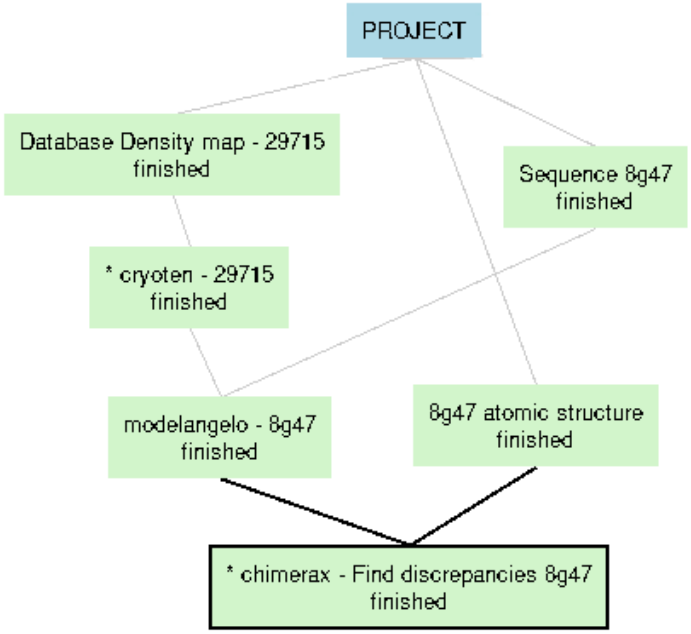
4.2. ChimeraX - find discrepancies

To test the results of the new protocol ChimeraX - find discrepancies, a simple workflow was designed in Scipion using the structure PDB-8G47 [6]. To better observe and analyze

the performance of this protocol, only chain A of the two chains (A and B) that form the PDB-8G47 structure was sent. This way, chain B in the database model will not be modelled by Modelangelo and it will appear white in the output results following the color palette. The full Scipion workflow used can be seen in Figure 4.4 and performs the following operations:

- **Import the density map** from the EMDB database with the ID 29715.
- **Import the FASTA sequence** from the PDB database with the ID 8G47.
- Use **CryoTEN** protocol to enhance the database density map.
- Generate the atomic model from this enhanced density map using **Modelangelo** protocol.
- **Import the database atomic model** from the PDB database to have it as a reference structure to see how Modelangelo performed.
- Use the developed protocol **ChimeraX - find discrepancies** to obtain a visual result of the regions of higher confidence generated by Modelangelo software.

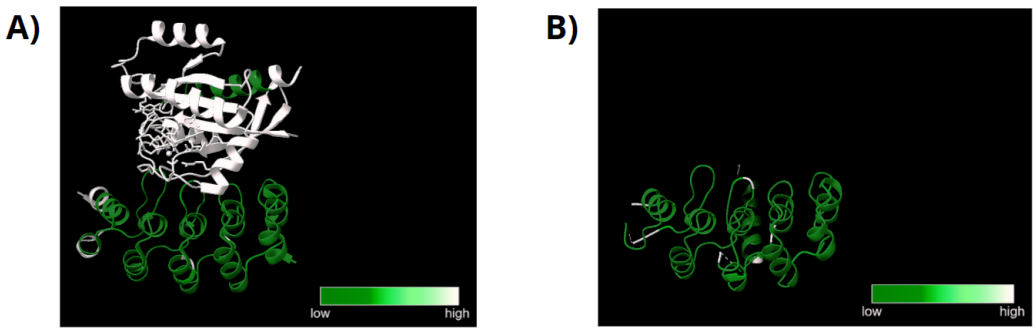
Figure 4.4
Scipion Workflow used to analyze ChimeraX - find discrepancies results for the 8G47 structure. The workflow includes importing the experimental density map and sequence, enhancing the map with CryoTEN, generating an atomic model with Modelangelo, and comparing it with the reference structure using the ChimeraX protocol.



After all the protocols have finished, the results obtained can be seen in Figure 4.5. In this protocol, the ideal expected result is a fully dark green alignment. Following the color palette *paegreen* [65], the residues with occupancies close to 0 will have the darkest green, and this color will become whiter as the value 30 is approached. Residues with values equal or higher than 30 will be pure white. With this color palette, the user can easily visualize the areas that were best modelled or that align the most. This is why the residues in the alignment that do not obtain a value of RMSD are given an arbitrary number of 250, leading to pure white colouring when visualization, showing that region as a low-confidence one. As this protocol bases the confidence in the RMSD, a lower value will mean that the residues in the alignment are closer, meaning they are more similar, ideally having an RMSD of 0 and being exactly the same residue.

In the example shown in Figure 4.5, the biological structure used was "KRAS G12C complex with GDP and AMG 510", which is a molecule composed of two chains: chain A (RCG-33) and chain B (GTPase KRas). Chain A corresponds with the higher confidence region that Modelangelo could easily model and is colored in darker green in both results. Meanwhile, chain B corresponds with the white region that can be seen in image A (Database model) of Figure 4.5. Following the input sequences introduced, chain B was not modelled by Modelangelo software. This is why it does not appear in image B of Figure 4.9. Meanwhile, it does appear in the database model (image A in Figure 4.5), and it is colored in pure white, which means that it couldn't be aligned with the Modelangelo structure, resulting in a high RMSD. These were the expected results, proving the consistency of the protocol and validating this way the Modelangelo output, which shows a dark green color, meaning high confidence.

Figure 4.5
ChimeraX - find discrepancies results: A) Database 5M1S atomic model; B) Modelangelo-generated atomic model.



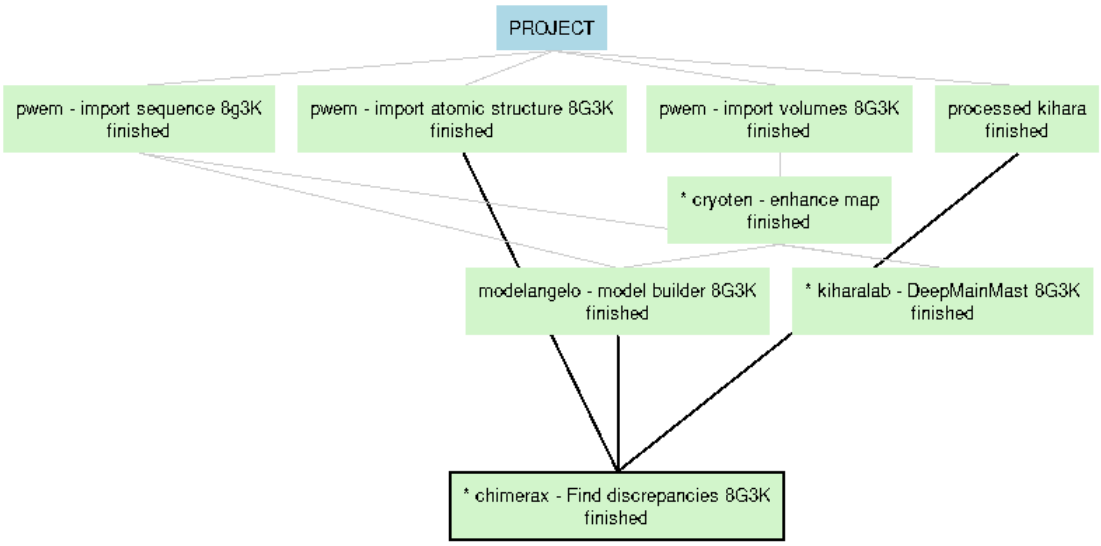
Overall, it can be stated that the protocol "ChimeraX - find discrepancies" satisfactory compared to the two biological structures sent and provided the user with a visual result to quickly analyze the quality of the atomic models. Moreover, further analysis can be done with the output atomic models of the protocol, which store in their metadata the numerical values for each residue's RMSD, allowing for quantitative measurements.

4.3. Full workflow for atomic model generation

After having tested both implemented plugins and seeing their successful performance, a design of a workflow is proposed in Figure 4.6. The stages of this workflow are:

- **pwem - import volumes 8G3K** and **pwem - import sequence 8G3K**: Import both the database density map from EMDB and the corresponding sequence: this can be done with Base Scipion plugin (pwem) [68].
- **cryoten - enhance map**: Apply CryoTEN protocol to enhance the density map imported from the database
- **modelangelo - model builder 8G3K** and **kiharalab - DeepMainMast 8G3K**: Use this enhanced density map and the imported sequence to generate atomic models using Kiharalab DeepMainMast protocol, modelangelo, or any other atomic model generator software.
- **pwem - import atomic structure 8G3K**: Import also the atomic model from the database to establish a reference: also done with Base Scipion plugin.
- **ChimeraX - Find discrepancies 8G3K**: Generate a visual comparison between these models and see the regions of highest confidence.

Figure 4.6
Scipion Workflow used to evaluate the performance results of Modelangelo and DeepMainMast on 8G3K structure. The process includes importing the sequence, atomic model, and density map, enhancing the map using CryoTEN, generating atomic models with both Modelangelo and DeepMainMast, and comparing the results using the ChimeraX – Find discrepancies protocol.



This workflow can be applied to any atomic structure available in PDB and EMDB. For this example, the selected model was 8G3K [8] [9]. This biological structure is composed of two chains (A and B) that form a complex. For this example, only the chain A sequence was sent to the atomic model generators Modelangelo and DeepMainMast to have a simpler output. All the results obtained during the execution of this workflow are shown in Figures 4.7 and 4.8.

Figure 4.7

Visualization of the protocols results in ChimeraX: A) 8G3K density map downloaded from EMDB; B) 8G3K density map after being processed by CryoTEN protocol; C) 8G3K atomic model generated by Modelangelo protocol; D) 8G3K atomic model generated by DeepMainMast protocol; E) 8G3K atomic model downloaded from PDB.

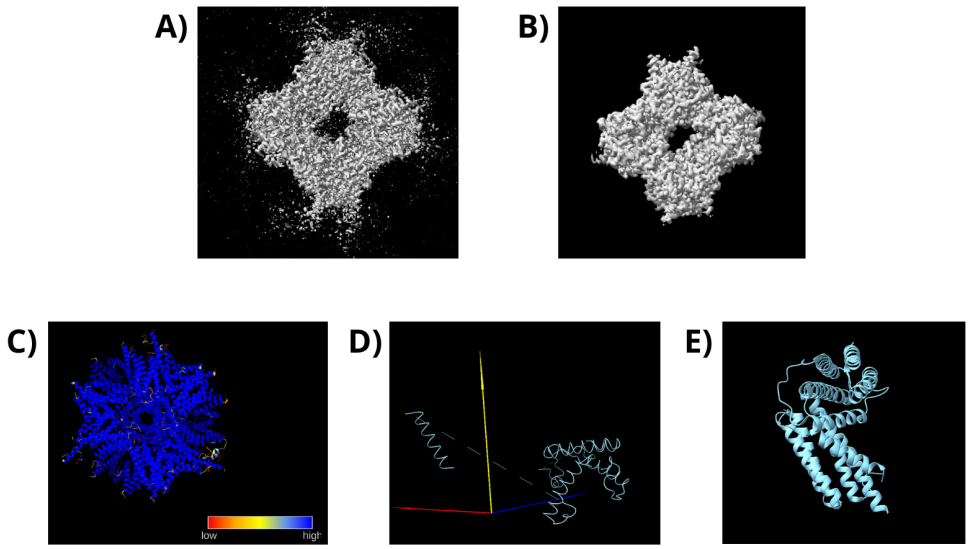
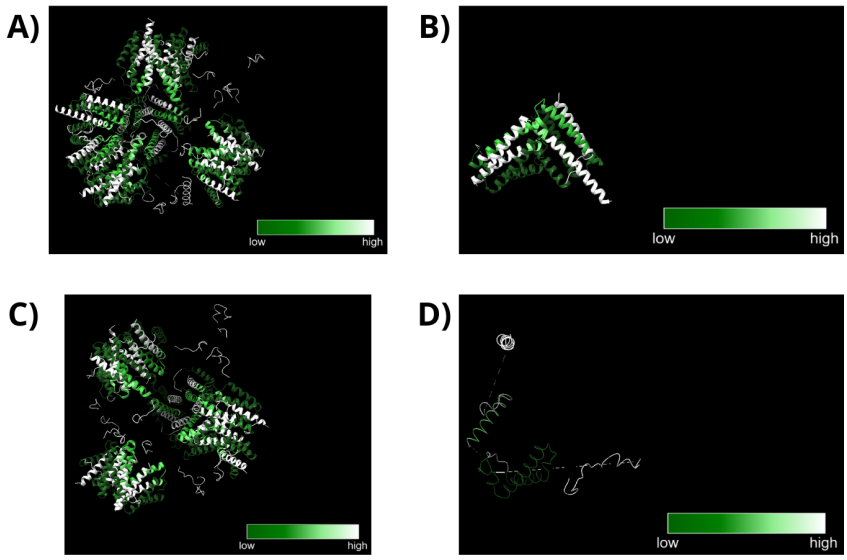


Figure 4.8

Visualization of the protocol "ChimeraX - Find discrepancies" results: A) Viewer with the three models introduced as input; B) Visualization of discrepancy regions of PDB Database 8G3K atomic model; C) Visualization of discrepancy regions of Modelangelo 8G3K atomic model; D) Visualization of discrepancy regions of DeepMainMast 8G3K atomic model.



As it can be seen in Figures 4.7 and 4.8, Modelangelo generates a larger atomic structure than Kiharalab - DeepMainMast protocol. This can be explained with Figure 4.9. In it, it can be seen that the density map EMDB-29700 corresponds with the whole KRAS G12C complex with GDP. However, the PDB-8G3K structure features only subunits A and B. The output shown in image B of Figure 4.8 corresponds to the whole PDB-8G3K complex composed of both chain A and B as it is directly downloaded from the database. Meanwhile, DeepMainMast results shown on image D of Figure 4.8 represent only subunit A of the structure, as it is the one corresponding to the sequence that was sent. DMM protocol adjusts its output to the information given, generating subunit A of the density map. However, Modelangelo takes as a reference the whole density map sent to it and generates the corresponding output, generating 12 subunits A of the structure PDB-8G3K following the geometry of the map EMDB-29700. If the original publication of how structure EMDB-29700 was characterized is read [7], it can be seen that the sample has been processed as 12 symmetrical units grouped in groups of three identical subunits following a tetrahedral symmetry, which corresponds exactly with the output that Modelangelo gives. This can explain why, even if Modelangelo provides several subunits that do not have any match with the DeepMainMast and Database models, their residues appear in a dark green in the "KiharaX - Find Discrepancies" protocol, as they are the same subunits but replicated 12 times to adapt to the shape of the density map provided as an input.

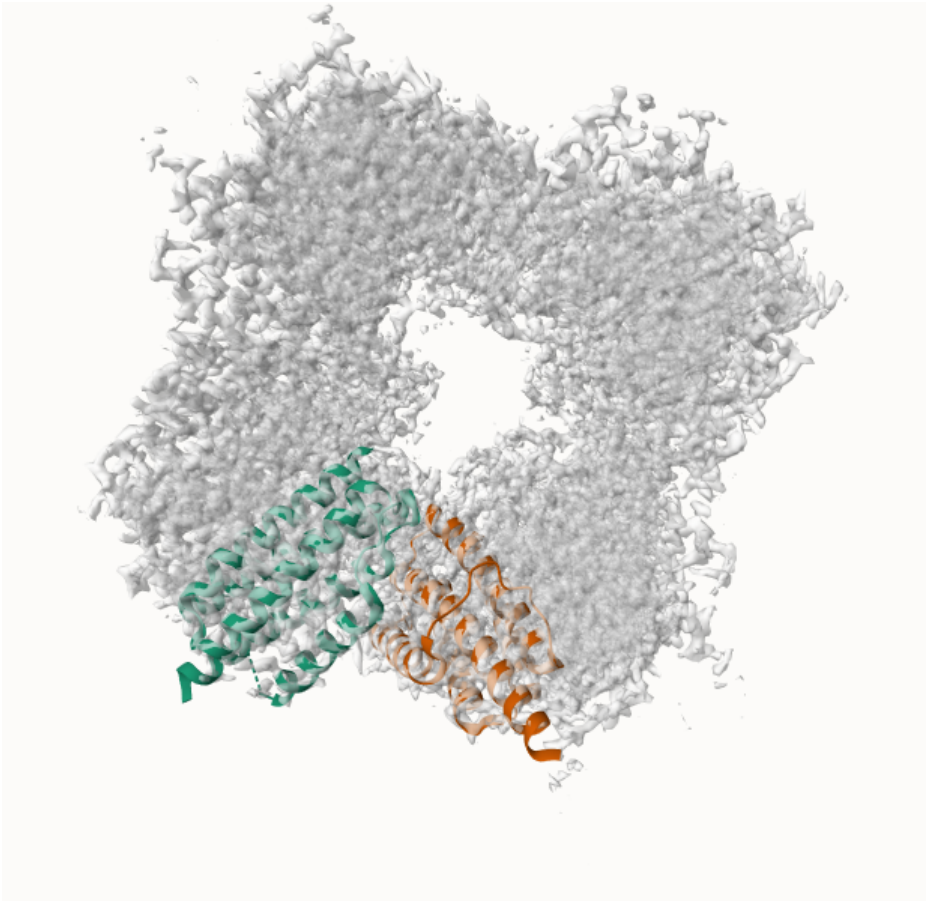
Having explained the difference in geometry of the different outputs, it can be seen

that both atomic generation models, Modelangelo and DeepMainMast, performed satisfactory in the generation of the atomic structure, as the final output provided by "KiharaX - Find Discrepancies" protocol shows mostly green regions, although some regions were not modelled, showing a white color. If the database model is looked (image B in Figure 4.8), approximately half of the structure appears white. This can be explained with the missing Chain B of the model that was not sent neither to Modelangelo nor DeepMainMast, therefore not being modelled and being colored white in the output.

Overall, the results obtained were satisfactory and provide an easy and intuitive way for users to quickly visualize how different atomic model generators performed.

Figure 4.9

3D visualization of the EMDB entry 29700. The density map is shown in transparent gray, with two fitted atomic models (chains A and B) highlighted in green and organce, respectively, illustrating the structural composition and organization within the cryo-EM volume.



Source: [9]

This workflow provides a systematic method for analyzing the performance of atomic model generation tools against established reference structures. In this example, Modelangelo and Kiharalab - DeepMainMast were used, but they could be any other *de novo* atomic generation models like the mentioned DeepEMhancer. The execution of the work-

flow validates this approach and establishes this Scipion workflow as a useful tool to evaluate the quality and confidence levels of different generated atomic models.

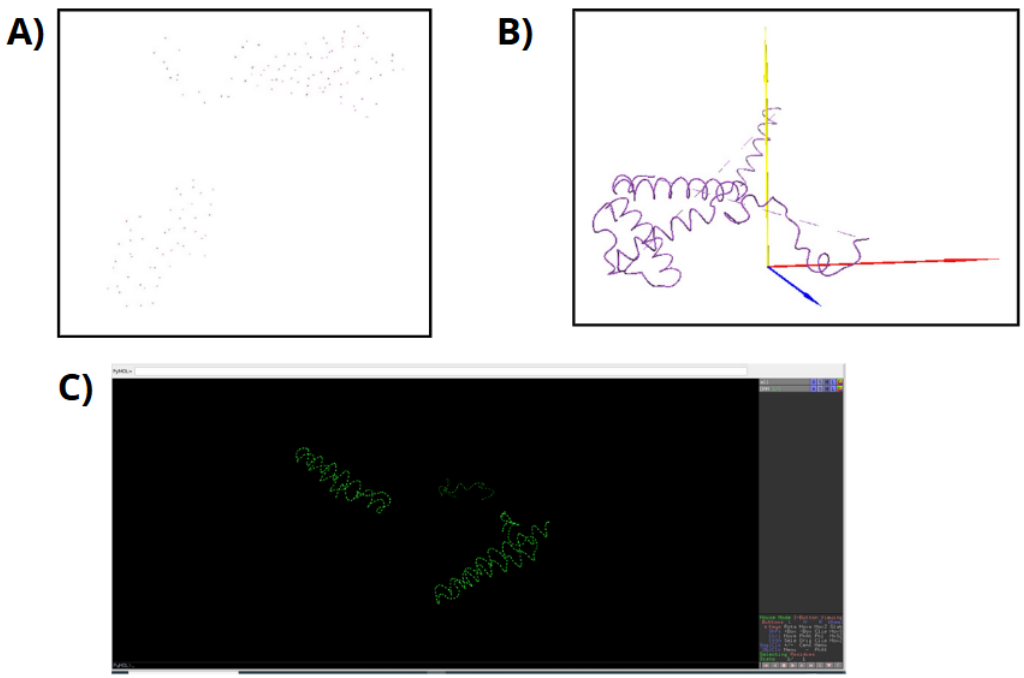
4.4. Challenges

This section of the results is intended to expose the issues encountered during the development of the project that could not be solved and other potential limitations of the workflow proposed.

The first main issue is the result obtained using the Kiharalab DeepMainMast protocol. When running this protocol, a PDB file should be generated with the full atomic structure predicted by DMM model. However, it only generates a CIF file that, when opened with ChimeraX, raises the error: "missing entity information". This error means that this CIF file lacks some metadata information, such as whether it is DNA, RNA, protein, ligand, etc., relationships between sequences and structures, or which residues belong to which molecules. As a result, ChimeraX does not generate chains between the predicted alpha-carbon atoms. Instead, they are just represented as dots in the coordinates they were predicted, as can be seen in image A of Figure 4.10.

Figure 4.10

Visualization of Pymol processing applied to the DeepMainMast output. (A) Original atomic model generated by DeepMainMast, visualized in ChimeraX, showing disordered and misaligned particles. (B) Final corrected model re-imported into Scipion and visualized again in ChimeraX, displaying an improved and coherent structure. (C) PyMOL interface used to manually adjust and correct the atomic coordinates before re-integration into the workflow.



Time constraints did not allow this issue to be resolved before the submission deadline. As a temporary solution, an intermediate software was introduced to recover the missing entity information in the CIF file. This software was PyMOL [46], a molecular visualization tool that enables the modification and analysis of biological structures. Using an educational license, the outputs from the DeepMainMast protocol were imported into PyMOL, which reinterpreted the entity information and generated a uniform molecular structure, consistent with the expected output of the DMM protocol (see image C in Figure 4.10). The final corrected molecule was then sent back to Scipion, as shown in image B of Figure 4.10

The second main issue comes with the implemented protocol ChimeraX - find discrepancies. By the way it is coded, when the user introduces any number of input atomic models, the protocol internally aligns them pairwise and use ChimeraX *matchmaker* command to obtain the results. This pairwise alignment is stored in folders called: "{model 1}_{model 2}". Once all alignments are performed, to generate the final output models, the code search for this folders and obtain the final atomic models by reading the names "{model 1}" and "{model 2} separated by the sign "_". Therefore, if any of the models already contain a "_" sign in its name, the code would not recognize the atomic model names properly and will raise an error. This issue was solved in the latest stages of development of the protocol, allowing the protocol to change the name of the input parameters to avoid conflict. However, the original file name of the input model is changed.

Apart from this, this protocol also poses some challenges for the virtual machine it is run on. By the way it was coded, once the protocol is finished and the user clicks on "Analyze results", a viewer is loaded and all the output molecules are opened in a ChimeraX window and coloured using *paegreen* palette. ChimeraX is a GPU-demanding visualization tool, so the higher the number of atomic models introduced by the user as input, the higher the GPU consumption will be. This can lead to ChimeraX program breaking due to this limitation if there are more atomic models than the virtual machine can handle. Solutions to this issue will be purposed in the section *Future Work*.

5. CONCLUSION AND FUTURE WORK

5.1. Conclusion

In the current state of scientific development, working at the cellular level has become the standard practice and reaching the atomic level is no longer a distant milestone. To achieve this goal, it is essential to study the distribution, interaction and behavior of atoms within molecules. Simulated molecular biology enables such studies to be done faster, more cost-effectively, and with greater efficiency.

By leveraging the spatial resolution potential of cryo-EM and applying appropriate post-processing techniques, the analysis of biological samples at atomic resolution is significantly simplified. Scipion is under continuous development, ensuring it remains aligned with the latest advances in molecular processing to maximize its usefulness. In this context, two new protocols were developed and integrated into the Scipion framework during this project, addressing key challenges encountered in sample processing.

CryoTEN implementation makes the enhancement of a density map, which is usually noisy, as easy as introducing the map the user wants to enhance. By implementing CryoTEN as a Scipion plugin, the density maps can notably improve their resolution and make it easier for atomic model generation algorithms to reconstruct the atomic structure with higher fidelity and confidence. This protocol can be applied to any molecule processing workflow and it will improve the results obtained.

The **"ChimeraX - Find Discrepancies"** protocol was developed to simplify and accelerate the comparison of different atomic models. Rather than requiring users to leave the Scipion environment and rely on external software to align and compare structural regions, this newly integrated protocol enables users to perform the comparison directly within Scipion. It provides a fast and intuitive visual representation of discrepancy regions between models. Additionally, the computed information is stored within the structure's metadata, allowing for further quantitative analysis.

With the development of these two new protocols, the study of *De novo* atomic models is greatly simplified, allowing for an easier, better, faster and more automated way for analyzing the confidence and quality of them without having to use external software models. Everything can be done inside the Scipion framework and obtain reproducible results.

Overall, the integration of these two protocols significantly enhances the capabilities of the Scipion framework, aligning with its core objective of unifying all cryo-EM processing tools within a single, streamlined environment.

5.2. Future Work

The Scipion framework is a software in constant evolution, as new post-processing cryo-EM tools appear almost every week. This microscopy technique is still young and much work is yet to be done to improve the quality of the samples acquired and, therefore, facilitate the analysis of them at the atomic level. In the case of the development of this bachelor thesis, several implementations remain to be done to better organise or visualize the results of the two purposed protocols: "CryoTEN - enhance map" and "ChimeraX - find discrepancies".

CryoTEN is fully implemented in Scipion and functional to the date of this work. However, it would be useful to implement a viewer inside the protocol to easily visualize the results of the map enhancement performed by CryoTEN software. For instance, this viewer could show in the ChimeraX interface both the input density map sent as input to the protocol and the output volume generated by the protocol. By colouring both of them at different tones, the user can easily see which noisy sections of the original map were removed and analyze the topography of both density maps.

In the case of "ChimeraX - find discrepancies" protocol, it is also fully functional inside the Scipion framework. However, several implementations could be useful to tackle the different issues that its implementation carries:

- The main atomic model the user wants to analyze should be introduced as a unique parameter, and the rest of the input atomic models would be sent in another parameter. This way, the software will have the information of which structure is the one the user is interested on and adapt the output visualization in ChimeraX by showing only the atomic model the user was interested in. This way, the GPU consumption that the current viewer carries when showing several atomic models at the same time can be significantly reduced by showing only the important model.

- The arbitrary RMSD value currently assigned when the alignment does not yield any result should be allowed to be introduced as a parameter by the user. By doing this, the user selects the worst RMSD value possible to be coloured and, consequently, the code of the protocol could be adapted so that it adapts the color palette to the value the user has established.
  For instance, if the user needs to analyze the result of an atomic model generation for an application that requires very high confidence, they could select the arbitrary value for misaligned residues at 15 angstroms, meaning that any RMSD value over 15 Å will be coloured white in the output visualization, instead of the value of 30 Å that is currently used.

Further work could also be done in both protocols in terms of optimization to reduce executing times and GPU consumption. Moreover, new functionalities can always be added to them if they are required in the following years.

6. REGULATORY FRAMEWORK

The techniques and methodologies employed in this study are based on both simulated and experimental cryo-EM data, all of which are publicly available and utilized in accordance with scientific, legal, and ethical standards. This project does not involve any proprietary or sensitive information, and all software tools applied are distributed under open-source licenses suitable for academic research. All cryo-EM density maps were obtained from the Electron Microscopy Data Bank (EMDB), and the corresponding atomic models were sourced from the Protein Data Bank (PDB). Both repositories provide open access to structural biology data, facilitating transparency and reproducibility in research.

The computational analyses were conducted using Python, an open-source programming language, within the PyCharm 2024.3 Integrated Development Environment (IDE) [44], accessed through a free student license. Molecular analysis were also performed using PyMOL, under a free student license.

The Scipion framework is based on the Linux operating system. This work was fully performed in Ubuntu 20.04.6 (Linux environment) [27], which is also open-source and freely available for download. Additional software tools employed include:

- UCSF ChimeraX: available free of charge for academic, government, nonprofit, and personal use.
- Scipion: open-source image processing framework for cryo-EM, distributed under the GNU General Public License (GPL).
- CryoTEN: released under the Creative Commons Attribution License (CC BY 4.0).
- ModelAngelo: available for free use under the MIT License.
- Kihara Lab software: provided for free use under the GNU General Public License v3 (GPLv3).

All software tools were utilized in accordance with their respective licenses, ensuring legal and ethical compliance.

7. SOCIO-ECONOMIC ENVIRONMENT

The introduction of cryoEM technology has lead to a revolution were the analysis of biological structures at the atomic level in their natural state is finally possible. With this microscopy technique, diseases can be better understood, treatments can be faithfully tested, and new molecular interactions with potential to develop new drugs can be studied. However, the processing of acquired images with cryoEM microscopy still remains available only for those academics with sufficient programming skills to use different software and scripts that improve the image quality.

This is where Scipion comes tothe stage. An open-source framework that makes cryoEM image processing more accessible and intuitive. With Scipion, any user can easily perform the operations to improve a cryoEM acquired image without the need to use various algorithms. Moreover, Scipion integrates all of them into one place, so even the professionals who can operate the different software models can find the Scipion framework useful due to its convenience. Scipion helps lower the cost and skill barrier to using advanced cryo-EM software and it is in constant development to keep up to date with the constant developments and discoveries of the biomedical area.

The introduction of two new protocols: "CryoTEN - enhance map" and "ChimeraX - find discrepancies"; contributes to Scipion project and, along with the work of the rest of the team, maintains it as a relevant and updated framework. Throughout the years, Scipion users have more options to choose from when deciding the workflow for processing a particular biological structure, allowing them to perform better, more automated and easier analysis.

7.1. Budget

Although the project is mainly open-source and it is freely available for any user to directly download any part and use it, there are several associated costs of the project that are shown in the following tables.

Figure 7.1
Associated costs of human resources.

Human Resources	Hours invested	Cost / Hour (€)	Total Cost (€)
Student	385	20	7700
Tutor	60	55	3300

Figure 7.2

Associated costs of technical equipment.

Technical equipment	Cost / Month (€)	Months used	Amortization (€)
Personal Computer	25	8	200
CSIC server Virtual Machine	200	3	600

Figure 7.3

Total associated costs of the project..

Final Costs	Total (€)
Human Resources	11000
Technical equipment	800
TOTAL ASSOCIATED COST	**11800**

BIBLIOGRAPHY

[1] Microscopy Australia. *Cryo-Electron Microscopy - MyScope*. (access: May 5th, 2025). URL: https://myscope.training/CRYO_Introducing_Single_Particle_Analysis.

[2] Reza Azad et al. "Medical Image Segmentation Review: The Success of U-Net". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (2024), pp. 10076–10095. DOI: 10.1109/TPAMI.2024.3435571.

[3] Javad Baghirov et al. "BPS2025 - Kihara Lab EM Webserver: A platform for automated cryo-EM structure analysis and modeling". In: *Biophysical Journal* 124.3, Supplement 1 (2025), 313a. ISSN: 0006-3495. DOI: https://doi.org/10.1016/j.bpj.2024.11.1737. URL: https://www.sciencedirect.com/science/article/pii/S0006349524024652.

[4] I. David Brown and Brian McMahon. "CIF: the computer language of crystallography". In: *Acta Crystallographica Section B* 58.3 Part 1 (June 2002), pp. 317–324. DOI: 10.1107/S0108768102003464. URL: https://doi.org/10.1107/S0108768102003464.

[5] Stephen K. Burley et al. "Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive". In: *Protein Crystallography: Methods and Protocols*. Ed. by Alexander Wlodawer, Zbigniew Dauter, and Mariusz Jaskolski. New York, NY: Springer New York, 2017, pp. 627–641. ISBN: 978-1-4939-7000-1. DOI: 10.1007/978-1-4939-7000-1_26. URL: https://doi.org/10.1007/978-1-4939-7000-1_26.

[6] Roger Castells-Graells et al. "Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold". In: *Proceedings of the National Academy of Sciences* 120.37 (2023), e2305494120. DOI: 10.1073/pnas.2305494120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2305494120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2305494120.

[7] Roger Castells-Graells et al. "Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold". In: *Proceedings of the National Academy of Sciences* 120.37 (2023), e2305494120. DOI: 10.1073/pnas.2305494120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2305494120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2305494120.

[8] Yeates T.O. Castells-Graells R. Sawaya M.R. *Cryo-EM imaging scaffold subunits A and B used to display KRAS G12C complex with GDP*. (access: May 18th, 2025). URL: https://www.wwpdb.org/pdb?id=pdb_00008g3k.

[9] Yeates T.O. Castells-Graells R. Sawaya M.R. *Cryo-EM imaging scaffold subunits A and B used to display KRAS G12C complex with GDP*. (access: May 18th, 2025). URL: https://www.ebi.ac.uk/emdb/EMD-29700.

[10] The wwPDB Consortium. "EMDB—the Electron Microscopy Data Bank". In: *Nucleic Acids Research* 52.D1 (Nov. 2023), pp. D456–D465. ISSN: 0305-1048. DOI: 10.1093/nar/gkad1019. eprint: https://academic.oup.com/nar/article-pdf/52/D1/D456/55039450/gkad1019.pdf. URL: https://doi.org/10.1093/nar/gkad1019.

[11] Tiago R. D. Costa, Athanasios Ignatiou, and Elena V. Orlova. "Structural Analysis of Protein Complexes by Cryo Electron Microscopy". In: *Bacterial Protein Secretion Systems: Methods and Protocols*. Ed. by Laure Journet and Eric Cascales. New York, NY: Springer New York, 2017, pp. 377–413. ISBN: 978-1-4939-7033-9. DOI: 10.1007/978-1-4939-7033-9_28. URL: https://doi.org/10.1007/978-1-4939-7033-9_28.

[12] J.M. de la Rosa-Trevín et al. "Xmipp 3.0: An improved software suite for image processing in electron microscopy". In: *Journal of Structural Biology* 184.2 (2013), pp. 321–328. ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2013.09.015. URL: https://www.sciencedirect.com/science/article/pii/S1047847713002566.

[13] Ashwin Dhakal et al. "Artificial intelligence in cryo-EM protein particle picking: recent advances and remaining challenges". In: *Briefings in Bioinformatics* 26.1 (Jan. 2025), bbaf011. ISSN: 1477-4054. DOI: 10.1093/bib/bbaf011. eprint: https://academic.oup.com/bib/article-pdf/26/1/bbaf011/61458053/bbaf011.pdf. URL: https://doi.org/10.1093/bib/bbaf011.

[14] Yi Ding et al. "A Stacked Multi-Connection Simple Reducing Net for Brain Tumor Segmentation". In: *IEEE Access* 7 (2019), pp. 104011–104024. DOI: 10.1109/ACCESS.2019.2926448.

[15] Electron Microscopy Data Bank (EMDB). *EMDB: Electron Microscopy Data Bank*. (access: May 10th, 2025). 2025. URL: https://www.ebi.ac.uk/emdb/.

[16] Richard Evans et al. "Protein complex prediction with AlphaFold-Multimer". In: *bioRxiv* (2022). DOI: 10.1101/2021.10.04.463034. eprint: https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034.full.pdf. URL: https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034.

[17] Yunior C. Fonseca et al. *Kiharalab plugin repository*. (access: May 9th, 2025). URL: https://github.com/scipion-em/scipion-em-kiharalab.

[18] Thomas D. Goddard et al. “UCSF ChimeraX: Meeting modern challenges in visualization and analysis”. In: *Protein Science* 27.1 (), pp. 14–25. DOI: https://doi.org/10.1002/pro.3235. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.3235. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3235.

[19] S. R. Hall, F. H. Allen, and I. D. Brown. “The Crystallographic Information File (CIF): a new standard archive file for crystallography”. In: *Acta Crystallographica Section A: Foundations of Crystallography* 47.6 (1991), pp. 655–685. DOI: 10.1107/S010876739101067X. URL: https://doi.org/10.1107/S010876739101067X.

[20] J. He, T. Li, and S. Y. Huang. “Improvement of cryo-EM maps by simultaneous local and non-local deep learning”. In: *Nature Communications* 14 (2023), p. 3217. DOI: 10.1038/s41467-023-39031-1. URL: https://doi.org/10.1038/s41467-023-39031-1.

[21] Gregory Hill and Xavier Bellekens. “CryptoKnight: Generating and Modelling Compiled Cryptographic Primitives”. In: *Information* 9.9 (Sept. 2018), p. 231. ISSN: 2078-2489. DOI: 10.3390/info9090231. URL: http://dx.doi.org/10.3390/info9090231.

[22] Alex Ilitchev. *How Does an Electron Microscope Work?* (access: May 5th, 2025). 2024. URL: https://www.thermofisher.com/blog/materials/how-does-an-electron-microscope-work/.

[23] Kiarash Jamali et al. “Automated model building and protein identification in cryo-EM maps”. In: *Nature* 628.8007 (Apr. 2024), pp. 450–457. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07215-4. URL: https://doi.org/10.1038/s41586-024-07215-4.

[24] G.H.U. Lamm et al. *CryoRhodopsins: a new clade of microbial rhodopsins from cold environments - EMDB*. (access: May 16th, 2025). URL: https://www.ebi.ac.uk/emdb/EMD-18799.

[25] G.H.U. Lamm et al. *CryoRhodopsins: a new clade of microbial rhodopsins from cold environments - PDB*. (access: May 16th, 2025). URL: https://doi.org/10.2210/pdb8R0O/pdb.

[26] Marcel A Lauterbach. “Finding, defining and breaking the diffraction barrier in microscopy – a historical perspective”. In: *Optical Nanoscopy* 1.1 (Nov. 2012), p. 8.

[27] Canonical Ltd. *Ubuntu 20.04 software documentation*. (access: May 25th, 2025). URL: https://help.ubuntu.com/20.04/ubuntu-help/index.html.

[28] Sai Raghavendra Maddhuri Venkata Subramaniya, Genki Terashi, and Daisuke Kihara. “Enhancing cryo-EM maps with 3D deep generative networks for assisting protein structure modeling”. In: *Bioinformatics* 39.8 (Aug. 2023), btad494. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad494. eprint: https://academic.oup.com/bioinformatics/article-pdf/39/8/btad494/51225792/btad494.pdf. URL: https://doi.org/10.1093/bioinformatics/btad494.

[29] M. Martínez et al. “Integration of Cryo-EM Model Building Software in Scipion”. In: *Journal of Chemical Information and Modeling* 60.5 (2020), pp. 2533–2540. DOI: 10.1021/acs.jcim.9b01032. URL: https://doi.org/10.1021/acs.jcim.9b01032.

[30] E. McLeod and A. Ozcan. “Nano-imaging enabled via self-assembly”. In: *Nano Today* (2014). DOI: 10.1016/j.nantod.2014.08.005.

[31] Microscope.com. *History of Microscopes*. (access: May 5th, 2025). URL: https://www.microscope.com/education-center/articles/history-of-microscopes.

[32] Keiron O’Shea and Ryan Nash. *An Introduction to Convolutional Neural Networks*. 2015. arXiv: 1511.08458 [cs.NE]. URL: https://arxiv.org/abs/1511.08458.

[33] Eugene Palovcak et al. “Enhancing the signal-to-noise ratio and generating contrast for cryo-EM images with convolutional neural networks”. en. In: *IUCrJ* 7.Pt 6 (Oct. 2020), pp. 1142–1150.

[34] Eric F. Pettersen et al. “UCSF ChimeraX: Structure visualization for researchers, educators, and developers”. In: *Protein Science* 30.1 (2021), pp. 70–82. DOI: 10.1002/pro.3943. URL: https://doi.org/10.1002/pro.3943.

[35] Jonas Pfab, Nhut Minh Phan, and Dong Si. “DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes”. In: *Proceedings of the National Academy of Sciences* 118.2 (2021), e2017525118. DOI: 10.1073/pnas.2017525118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2017525118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2017525118.

[36] Jonas Pfab, Nhut Minh Phan, and Dong Si. “DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes”. en. In: *Proc Natl Acad Sci U S A* 118.2 (Jan. 2021).

[37] Ali Punjani et al. “cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination”. In: *Nature Methods* 14.3 (Mar. 2017), pp. 290–296.

[38] RCSB Protein Data Bank. *RCSB PDB: The Research Collaboratory for Structural Bioinformatics*. (access: May 10th, 2025). 2025. URL: https://www.rcsb.org/.

[39] Javier Sanchez del Rio. *ChimeraX plugin backup repository*. (access: May 10th, 2025). URL: https://github.com/JavierSanchez-bio/ChimeraX.

[40] Javier Sanchez del Rio. *Scipion-EM-CryoTEN backup repository*. (access: May 9th, 2025). URL: https://github.com/scipion-em/scipion-em-cryoten.

[41] Javier Sanchez del Rio and Carlos Oscar S. Sorzano. *Scipion-EM-CryoTEN official repository*. (access: May 9th, 2025). URL: https://github.com/scipion-em/scipion-em-cryoten.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[43] J.M. de la Rosa-Trevín et al. "Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy". In: *Journal of Structural Biology* 195.1 (2016), pp. 93–99. ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2016.04.010. URL: http://www.sciencedirect.com/science/article/pii/S104784771630079X.

[44] JetBrains s.r.o. *Pycharm software documentation*. (access: May 25th, 2025). URL: https://www.jetbrains.com/help/pycharm/getting-started.html.

[45] R Sanchez-Garcia et al. "DeepEMhancer: a deep learning solution for cryo-EM volume post-processing". In: *bioRxiv* (2020). DOI: 10.1101/2020.06.12.148296. eprint: https://www.biorxiv.org/content/early/2020/08/17/2020.06.12.148296.full.pdf. URL: https://www.biorxiv.org/content/early/2020/08/17/2020.06.12.148296.

[46] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.8". Nov. 2015. URL: https://www.pymol.org/.

[47] Scipion Team. *Scipion: A Workflow Framework for Electron Microscopy*. (access: May 5th, 2025). 2024. URL: https://scipion.i2pc.es/.

[48] Joel Selvaraj, Liguo Wang, and Jianlin Cheng. *CryoTEN software, github page*. (access: May 9th, 2025). URL: https://github.com/jianlin-cheng/cryoten.

[49] Joel Selvaraj, Liguo Wang, and Jianlin Cheng. "CryoTEN: efficiently enhancing cryo-EM density maps using transformers". In: *Bioinformatics* 41.3 (Feb. 2025), btaf092. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaf092. eprint: https://academic.oup.com/bioinformatics/article-pdf/41/3/btaf092/62412169/btaf092_supplementary_data.pdf. URL: https://doi.org/10.1093/bioinformatics/btaf092.

[50] Nahian Siddique et al. "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications". In: *IEEE Access* 9 (2021), pp. 82031–82057. DOI: 10.1109/ACCESS.2021.3086020.

[51] Nahian Siddique et al. “U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications”. In: *IEEE Access* 9 (2021), pp. 82031–82057. DOI: 10.1109/ACCESS.2021.3086020.

[52] M. Simon. *Cryo-electron microscopy wins the Nobel Prize in Chemistry*. (access: May 5th, 2025). 2017. URL: https://www.wired.com/story/cryo-electron-microscopy-wins-the-nobel-prize-in-chemistry/.

[53] S. Singh and S.S. Hasan. *Structure of SARS CoV-2 full-length spike protein with internal tag, 2RBD-up conformation - EMDB*. (access: May 12th, 2025). URL: https://www.ebi.ac.uk/emdb/EMD-44475.

[54] S. Singh and S.S. Hasan. *Structure of SARS CoV-2 full-length spike protein with internal tag, 2RBD-up conformation - PDB*. (access: May 12th, 2025). URL: https://doi.org/10.2210/pdb9BEA/pdb.

[55] J. Slawek et al. *Cryo-electron microscopy structure of glucose/xylose isomerase from Streptomyces rubiginosus with cobalt ions in the active site*. (access: May 16th, 2025). URL: https://www.rcsb.org/structure/9GRD.

[56] J. Slawek et al. *Cryo-electron microscopy structure of glucose/xylose isomerase from Streptomyces rubiginosus with cobalt ions in the active site*. (access: May 16th, 2025). URL: https://www.ebi.ac.uk/emdb/EMD-51521.

[57] Carlos Oscar S. Sorzano et al. “Image Processing in Cryo-Electron Microscopy of Single Particles: The Power of Combining Methods”. In: *Structural Proteomics: High-Throughput Methods*. Ed. by Raymond J. Owens. New York, NY: Springer US, 2021, pp. 257–289. ISBN: 978-1-0716-1406-8. DOI: 10.1007/978-1-0716-1406-8_13. URL: https://doi.org/10.1007/978-1-0716-1406-8_13.

[58] D. Strelak et al. “Advances in Xmipp for Cryo-Electron Microscopy: From Xmipp to Scipion”. In: *Molecules (Basel, Switzerland)* 26.20 (2021), p. 6224. DOI: 10.3390/molecules26206224. URL: https://doi.org/10.3390/molecules26206224.

[59] Genki Terashi et al. “DeepMainmast: integrated protocol of protein structure modeling for cryo-EM with deep learning and structure prediction”. In: *Nature Methods* 21.1 (Jan. 2024), pp. 122–131.

[60] Thomas C. Terwilliger et al. “Cryo-EM map interpretation and protein model-building using iterative map segmentation”. In: *Protein Science* 29.1 (2020), pp. 87–99. DOI: 10.1002/pro.3740. URL: https://doi.org/10.1002/pro.3740.

[61] Visualization UCSF Resource for Biocomputing and Informatics. *ChimeraX: command color*. (access: May 10th, 2025). URL: https://www.cgl.ucsf.edu/chimerax/docs/user/commands/color.html.

[62] Visualization UCSF Resource for Biocomputing and Informatics. *ChimeraX: command matchmaker*. (access: May 10th, 2025). URL: https://www.cgl.ucsf.edu/chimerax/docs/user/commands/matchmaker.html.

[63] Visualization UCSF Resource for Biocomputing and Informatics. *ChimeraX: command sequence*. (access: May 10th, 2025). URL: https://www.cgl.ucsf.edu/chimerax/docs/user/commands/sequence.html.

[64] Visualization UCSF Resource for Biocomputing and Informatics. *ChimeraX: command setattr*. (access: May 10th, 2025). URL: https://www.cgl.ucsf.edu/chimerax/docs/user/commands/setattr.html.

[65] Visualization UCSF Resource for Biocomputing and Informatics. *ChimeraX: palettes options*. (access: May 15th, 2025). URL: https://www.rbvi.ucsf.edu/chimerax/docs/user/commands/palettes.html.

[66] CNB - Biocomputing Unit. *ChimeraX plugin repository*. (access: May 10th, 2025). URL: https://github.com/scipion-em/scipion-em-chimera.

[67] CNB - Biocomputing Unit. *Modelangelo plugin repository*. (access: May 10th, 2025). URL: https://github.com/scipion-em/scipion-em-modelangelo.

[68] CNB - Biocomputing Unit. *Scipion Base plugin repository*. (access: May 16th, 2025). URL: https://github.com/scipion-em/scipion-em.

[69] The Regents of the University of California. *Introduction to Protein Data Bank Format*. (access: May 15th, 2025). URL: https://www.rbvi.ucsf.edu/chimera/1.2065/docs/UsersGuide/tutorials/pdbintro.pdf.

[70] Javier Vargas et al. "Efficient initial volume determination from electron microscopy images of single particles". In: *Bioinformatics* 30.20 (June 2014), pp. 2891–2898. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu404. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/20/2891/48929966/bioinformatics_30_20_2891.pdf. URL: https://doi.org/10.1093/bioinformatics/btu404.

[71] Javier Vargas et al. "Efficient initial volume determination from electron microscopy images of single particles". In: *Bioinformatics* 30.20 (June 2014), pp. 2891–2898. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu404. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/20/2891/48929966/bioinformatics_30_20_2891.pdf. URL: https://doi.org/10.1093/bioinformatics/btu404.

[72] N R Voss et al. "DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy". en. In: *J Struct Biol* 166.2 (May 2009), pp. 205–213.

[73] Bettina Voutou, Eleni Stefanaki, and Konstantinos Giannakopoulos. "Electron Microscopy: The Basics". In: 2008. URL: https://api.semanticscholar.org/CorpusID:100771920.

[74] David Woolford, Ben Hankamer, and Geoffery Ericksson. “The Laplacian of Gaussian and arbitrary z-crossings approach applied to automated single particle reconstruction”. In: *Journal of Structural Biology* 159.1 (2007), pp. 122–134. ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2007.03.003. URL: https://www.sciencedirect.com/science/article/pii/S1047847707000767.

[75] Worldwide Protein Data Bank (wwPDB). *wwPDB: Worldwide Protein Data Bank*. (access: May 10th, 2025). 2025. URL: https://www.wwpdb.org/.