Universidad Autónoma de Madrid | Escuela de Doctorado

Desarrollo de algoritmos de procesamiento de imagen para Crio-microscopía Electrónica

Mª Estrella Fernández Giménez,

Programa de Doctorado en Ingeniería Informática y de Telecomunicación

Centro realización Tesis/Institución: Centro Nacional de Biotecnología (CNB-CSIC)

Universidad Autónoma de Madrid | Escuela de Doctorado

Departamento UAM: Escuela Politécnia Superior

Madrid, 2023

Agradecimientos

Parecía que nunca iba a llegar el día de escribir los agradecimientos, pero llegó.

Gracias a mis directores y tutor de tesis, en especial a Carlos Óscar Sorzano, por ser el guía y solucionador de problemas de los proyectos realizados durante estos cuatro años de tesis, por enseñarme su forma de trabajar, por enseñarme a hacer ciencia de alto nivel y por descubrirme el mundo de la crío-microscopía electrónica, que me ha permitido “ver” proteínas con mis propios ojos, cosa que me fascina.

Gracias a todos mis copañeros de la Unidad de Biocomputación del Centro Nacional de Biotecnología o “B13”, por toda su ayuda teórica, práctica, técnica, administrativa y anímica.

Gracias a todo el apoyo que he tenido durante estos cuatro años de tesis, sin el cuál habría sido sin duda imposible llegar hasta aquí viva y cuerda. En especial, muchas gracias a mis padres, por haberme puesto en el principio del camino en las mejores condiciones posibles. Y por último y no menos importante si no más, infinitas gracias a Edgar, por su gran paciencia, ánimo, amor y aguante, porque ha conseguido que llegue al final de este camino. Deberían concederle el título de Doctor Consorte.

Development of Image Processing Algorithms for Cryo-Electron Microscopy

Author: Estrella Fernández Giménez[1,2]
Director: Carlos Óscar S. Sorzano[1]
Co-Director: José María Carazo[1]
Tutor: Roberto Marabini[1,2]

[1]Centro Nac. Biotecnología (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain
[2]Univ. Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain

September 2023

Abstract

The present thesis, structured as a compendium of articles, is focused on the development of algorithms for image processing of cryo-electron microscopy (cryo-EM), mainly in the modality of Single Particle Analysis (SPA), but also with some smaller contributions to Tomography and Subtomogram Averaging (STA) modalities.

Cryo-EM is an emergent field that is positioning itself as one of the most important tools for revealing the three-dimensional structure at near-atomic and atomic resolution of biological macromolecules and complexes, contributing considerably to diseases research such as cancer, viral infections, neuro-degenerative conditions, and drugs development. To achieve these goals, the cryo-EM field uses the most advanced technology both in hardware and software. In particular, image processing algorithms are one of the key parts of this field.

The main contribution of this thesis is the development and implementation of signal subtraction algorithms both for three-dimensional reconstructions (volume subtraction) and two-dimensional particles (projection subtraction). Besides, the thesis contributes to the validation and robustness of the field with a deep analysis of local defocus refinement. Finally, other developments have been made for specific needs of the group and for the contribution to Xmipp and Scipion in general, two of the most used software in the sector.

All the code developed in this thesis is open source and available at Xmipp and Scipion repositories on GitHub. The software developed is fully available and usable through Xmipp and Scipion.

Resumen

La presente tesis, estructurada como un compendio de artículos, se centra en el desarrollo de algoritmos para el procesamiento de imágenes de crio-microscopía electrónica (cryo-EM, por sus siglas en inglés), principalmente en la modalidad de Análisis de Partículas Individuales (SPA, por sus siglas en inglés), pero también con algunas contribuciones a las modalidades de tomografía y promedio de subtomogramas (STA, por sus siglas en inglés).

La microscopía electrónica es un campo emergente que se está posicionando como una de las herramientas más importantes para desvelar la estructura tridimensional a resolución casi atómica y atómica de macromoléculas y complejos biológicos, contribuyendo considerablemente a la investigación de enfermedades como el cáncer, infecciones víricas, condiciones neurodegenerativas y al desarrollo de fármacos. Para lograr estas metas, el campo de la cryo-EM utiliza la tecnología más avanzada tanto en hardware como en software. En particular, los algoritmos de procesamiento de imágenes son una de las partes clave de este campo.

La principal contribución de esta tesis es el desarrollo e implementación de algoritmos de sustracción de señales tanto para reconstrucciones tridimensionales (sustracción de volúmenes) como de partículas bidimensionales (sustracción de proyecciones). Además, esta tesis contribuye a la validación y robustez del campo con un análisis profundo sobre el refinamiento del desenfoque local. Finalmente, se han realizado otros desarrollos para necesidades específicas del grupo y para contribuir a Xmipp y Scipion en general, dos de los softwares más utilizados en el sector.

Todo el código desarrollado en esta tesis es código abierto y está disponible en los repositorios de Xmipp y Scipion en GitHub. El software desarrollado está completamente disponible y usable en Xmipp y Scipion.

Contents

1. Introduction

1.1. Cryo-Electron Microscopy

Cryo-electron microscopy (cryo-EM) is a revolutionary imaging technique that has transformed our understanding of the molecular world. It allows scientists to capture high-resolution three-dimensional images of biological molecules and complexes, providing unprecedented insights into their structure and function.

Traditionally, researchers have used X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy to study the molecular structures of biological samples. While these methods have been invaluable, they have certain limitations. X-ray crystallography requires crystallizing the sample, which is challenging for many molecules, and NMR spectroscopy is limited in its ability to resolve large macromolecular complexes [24].

In contrast, cryo-EM has emerged as a powerful alternative that bypasses many of these limitations [10]. It enables the visualization of molecules in their near-native states, without the need for crystallization. This is accomplished by flash-freezing the sample in a thin layer of vitreous ice, preserving its natural structure and avoiding the artifacts introduced by other techniques.

The key breakthrough in cryo-EM came with the development of direct electron detectors, which replaced traditional photographic films. These detectors are capable of capturing low-energy electrons emitted by the sample, converting them into digital signals [21]. The resulting images contain detailed information about the sample's structure, which can be reconstructed into a three-dimensional model.

The cryo-EM workflow involves several steps, including sample preparation, grid preparation, data collection (acquired images are known as micrographs), which requires sophisticated electron microscopes equipped with automated imaging systems capable of collecting thousands of images in a short period [42], image processing [35] and model building [8] (see Fig. 1).

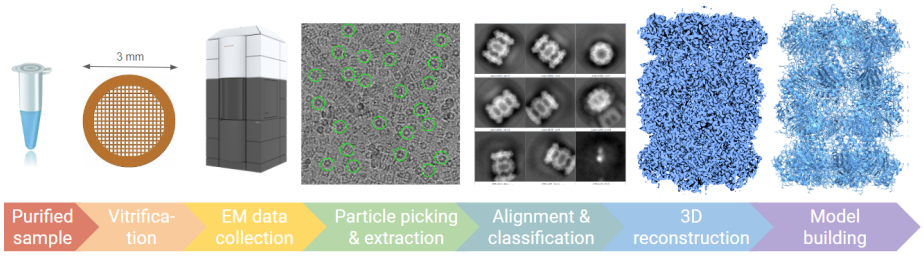


Figure 1: Typical cryo-EM workflow.

One of the critical aspects that have contributed to the success of cryo-EM is the development of advanced image processing techniques. Image processing

plays a vital role in extracting high-quality structural information from the raw data collected by cryo-EM experiments [45].

The cryo-EM imaging process involves the collection of a large number of two-dimensional projection images of the sample from various angles. Each image represents a snapshot of the sample in different orientations. However, these images are affected by various sources of noise, including electron beam-induced motion [52], beam-induced radiation damage, and detector imperfections [34]. Furthermore, the desired structural information is hidden within a low signal-to-noise ratio.

Image processing techniques aim to overcome these challenges and extract a high-resolution three-dimensional reconstruction of the sample. The process involves several steps, including image alignment, particle picking, and three-dimensional reconstruction.

The first step in image processing is *image alignment*, also known as motion correction or image registration. Since the sample undergoes random movements during data collection, the alignment process corrects for these movements and aligns the images with sub-pixel accuracy. Several algorithms, such as cross-correlation and phase correlation, are used for accurate image alignment [52].

Once the images are aligned, the next step is *particle picking*. It involves the identification and extraction of individual particle images from cryo-EM micrographs, which are often crowded with noise and contaminants. Particle-picking algorithms employ various strategies, such as template matching [14] and machine learning [36], [47] approaches, to automatically detect particles and generate a set of coordinates representing their positions.

The final step in cryo-EM image processing is a *three-dimensional reconstruction* of the sample, the so-called *density map*, which represents the structure of the sample at near-atomic resolution. It provides valuable insights into the organization and arrangement of the molecules within the sample [35].

The result of the image processing pipeline is a three-dimensional density map of the structure, which can be post-processed (sharpened) to improve its finner features [29]. If the resolution of the final map is good enough (typically below 3Å), it is possible to use the map to construct a detailed atomic model of the structure, a process known as *model building* [19].

Cryo-EM image processing has become an indispensable component of cryo-EM studies, allowing researchers to extract detailed structural information from raw image data. The continued refinement and development of image processing techniques in cryo-EM hold great promise for further accelerating our understanding of complex biological systems. It has contributed to numerous breakthroughs in understanding biological processes and has accelerated drug discovery by revealing the precise binding sites of potential therapeutics.

1.1.1. Single Particle Analysis

There are two modalities in cryo-EM: *Single Particle Analysis* (SPA) and *tomography*. In this section, we shall consider the former; the latter will be explained in the next section.

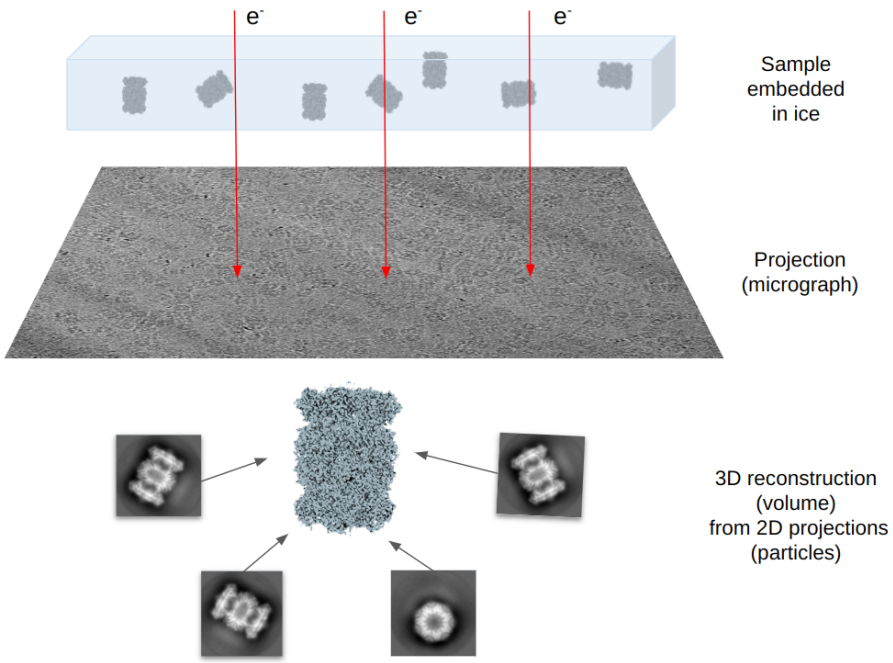


Figure 2: Schematic representation of Single Particle Analysis (SPA) technique.

SPA is a powerful technique in structural biology that allows us to determine the 3D structure of macromolecules at near-atomic resolution. The process, represented in Fig. 2, involves acquiring a large number of 2D projection images of individual particles embedded in a thin layer of vitrified ice, followed by image processing steps to reconstruct the 3D structure [5].

SPA Acquisition consists of the following steps[44]:

***Sample Preparation*:** The first step in SPA is the preparation of the sample. The macromolecule of interest is purified and then flash-frozen in a thin layer of vitrified ice to preserve its native state. This is typically achieved by plunge-freezing the sample in liquid ethane or propane at cryogenic temperatures.

***Cryo-EM Data Collection*:** The cryo-EM data collection process involves using an electron microscope to acquire a series of 2D projection images. The microscope is operated under low-dose conditions to minimize radiation damage to the sample. The sample grid containing the frozen particles in their random orientations is loaded into the microscope, and the data collection parameters, such as defocus values and microscope settings, are optimized.

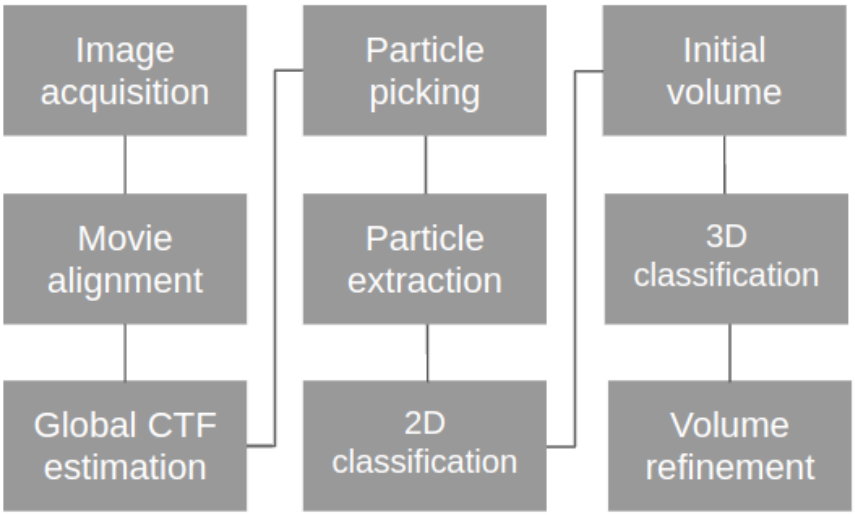


Figure 3: Standard Single Particle Analysis (SPA) workflow.

***Data Acquisition*:** The microscope captures images of the sample by illuminating it with a beam of electrons. The electrons interact with the sample and the resulting image is recorded by a detector. Modern cryo-EM instruments often use direct electron detectors, which provide high-resolution and low-noise images. The detector records the position, intensity, and defocus of each image.

SPA Image Processing consists of several steps (see a typical workflow in Fig. 3):

***Movie alignment [52]*:** Due to factors such as sample drift and beam-induced motion, the acquired images may exhibit small shifts and rotations. Movie alignment or motion correction algorithms are used to align all the images to a common reference frame, correcting for these translational and rotational motions. This step ensures that the particles in the micrographs are properly aligned, improving the accuracy of subsequent processing steps.

***Contrast Transfer Function (CTF) Correction [38]*:** The CTF describes the degradation of the electron microscope's image due to the presence of the microscope's lens system and other factors. It introduces a complex pattern of phase shifts and amplitude attenuation to the recorded images, resulting in blurred and distorted features. CTF correction aims at reversing these effects and restoring the high-resolution information in the cryo-EM images. By estimating the CTF parameters, such as defocus

and astigmatism, for each micrograph, the images can be computationally corrected to enhance contrast and improve resolution. This correction is crucial for subsequent steps in cryo-EM analysis, as it helps to achieve higher fidelity and more accurate 3D structural information from the cryo-EM data.

***Particle Picking [14][36][47]*:** In this step, algorithmic tools are employed to automatically identify and extract the individual particle images from the motion-corrected and CTF-corrected micrographs. Particle-picking algorithms use various techniques, such as template matching[36] or machine learning[47], to locate the particles and generate a set of coordinates representing their positions.

***Image Classification and Selection [49]*:** The extracted particle images often contain noise and contaminants. Image classification methods, such as reference-free 2D classification or supervised classification, are applied to group similar particles together based on their structural features. The resulting classes can be visually inspected to identify the most representative and high-quality particles for subsequent processing steps.

***3D Reconstruction [33]*:** Once a set of high-quality particles is selected, the next step is to reconstruct the 3D structure of the macromolecule. This is done by applying mathematical algorithms, such as projection matching methods, to align and combine the 2D particle images into a 3D volume. The 3D reconstruction process involves iterative refinement steps to optimize the alignment and improve the resolution of the reconstructed structure.

***Map Refinement and Validation*:** The initial 3D reconstruction can be further refined to improve the resolution and quality of the map. Techniques such as Bayesian polishing [53] or per-particle CTF refinement can be employed to reduce imaging artifacts and correct imperfections in the imaging process. The refined map is then validated using various metrics, including Fourier shell correlation (FSC) [28], to assess the quality and resolution of the structure.

1.1.2. Tomography and Subtomogram Averaging

Tomography and *subtomogram averaging* (STA) are powerful techniques in cryo-EM that enable the study of the 3D structure of macro-molecules and cellular complexes within their native environment [31]. Even though image acquisition and image processing maintain many similarities with SPA, as both are different techniques of cryo-EM, there are key differences that it is worth to remark (see Fig. 4 for a schematic representation).

Tomography Image Acquisition consists of:

***Sample Preparation [30]*:** In this modality, due to how the images are collected, is it possible to have samples in their native state (not purified),

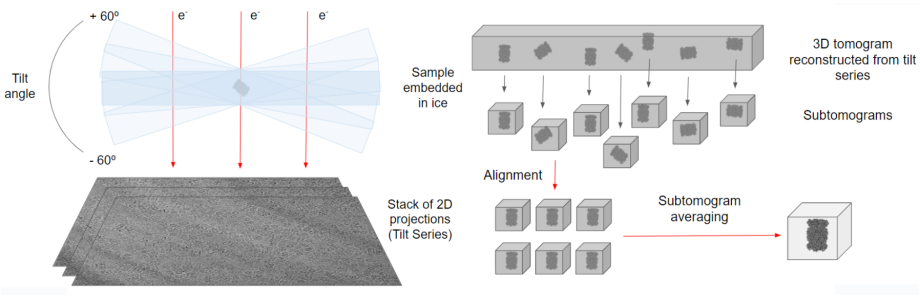


Figure 4: Schematic representation of Tomography and Subtomogram Averaging (STA) techniques.

such as intact cells or tissue sections. The sample is prepared in a suitable matrix to support it during cryo-EM imaging and then rapidly vitrified as in SPA, to preserve its native state and prevent ice crystal formation.

***Tilt-Series Data Collection [9]*:** The cryo-EM data collection process for tomography involves acquiring a series of 2D projection images at different tilt angles of the specimen. The tilting of the sample is the main difference with SPA. The grid-mounted sample is placed in the electron microscope, and the tilt series is collected by tilting the stage while acquiring images at each tilt angle. The tilt angles typically range from -60 to +60 degrees, capturing information from various perspectives. The tilt range is inherently limited due to practical constraints and the physical limitations of the microscope setup, resulting in a missing angular region around the z-axis ('missing wedge') [43]. This leads to an incomplete representation of the specimen's structure in the reconstructed volume, known as tomogram. The missing wedge can introduce artifacts, distortions, and limitations in resolution along specific orientations within the reconstructed volume.

***Data Acquisition*:** During each tilt, the electron beam illuminates the sample, and the resulting images are recorded by a detector. The images contain information about the electron scattering from the sample, as in SPA but at different tilt angles.

Tomography and STA Image Processing consist of following steps:

***Alignment and Reconstruction*:** The acquired tilt-series images undergo alignment and reconstruction to generate a 3D tomogram. Alignment algorithms are used to correct for tilt-induced image shifts and rotations, aligning the images relative to a common reference frame [37]. Subsequently, tomographic reconstruction algorithms, such as weighted back-projection (WBP)[6] or iterative methods like Simultaneous Iterative Reconstruction

Technique (SIRT)[3], are employed to reconstruct the 3D volume from the aligned tilt-series images.

***Subtomogram Picking [20]*:** Within the 3D tomogram, specific regions or subvolumes of interest containing the macromolecule or cellular complex ('subtomograms') are selected for further analysis. The selection of these regions is the same idea as the picking in SPA, however in this case they are volumes instead of 2D images. Subtomogram extraction involves cropping out these regions from the tomogram as smaller 3D subvolumes, each representing a potential instance of the structure of interest.

***Subtomogram Alignment [4]*:** The extracted subtomograms are aligned to a common reference structure to compensate for any local deviations in orientation and position. Alignment algorithms employ strategies such as cross-correlation or optimization-based approaches to iteratively adjust the subtomograms' position and orientation, aligning them with respect to a target reference structure.

***Subtomogram Averaging [2]*:** After alignment, the aligned subtomograms are averaged to enhance the signal-to-noise ratio and obtain a high-resolution representation of the macromolecule or complex. Averaging methods, such as common lines or projection matching algorithms, are applied to merge and align the subtomograms, utilizing their similar features. This results in a refined and improved 3D structure. In theory, it is possible to get similar resolutions than in SPA, however, in practice, it is usually more difficult to achieve such a high resolution in STA due to the noisy environment and the smaller amount of the macromolecule of interest.

***Map Refinement and Validation*:** Similar to SPA, the averaged 3D map can be further refined and sharpened using iterative refinement methods, including techniques like Bayesian polishing or local refinement. These methods aim to reduce imaging artifacts, correct structural heterogeneity, and enhance the resolution and quality of the final map. The refined map is then validated using metrics like Fourier shell correlation (FSC) to assess its resolution.

1.2. State of the art

In recent years, significant advancements in cryo-EM image processing have been made, thanks to increased computational power, better algorithms and new software tools. These advancements have led to improved resolution, shorter processing times and enhanced automation of the image processing pipeline.

Obtaining a high-resolution 3D reconstruction by cryo-EM is not trivial and can still be a challenging task. While cryo-EM has made significant advancements in recent years, achieving high resolution typically requires careful experimental design, optimization of sample preparation, and sophisticated data processing [22], particularly for complex samples or under certain experimental conditions.

The ease of obtaining a high-resolution cryo-EM reconstruction depends on several factors:

***Sample Quality and Homogeneity*:** High-resolution cryo-EM requires samples that are homogeneous and structurally well-defined [35]. Any sample heterogeneity or structural variability can hinder the alignment and classification of particles, resulting in lower-resolution reconstructions [10].

***Particle Numbers*:** To achieve high resolution, a sufficient number of particles is needed for accurate alignment and averaging [47]. Samples with low particle numbers can lead to lower-resolution reconstructions. Advanced data collection methods and improved detection technologies have helped overcome this challenge to some extent [47], [20].

***Imaging Parameters*:** Optimizing imaging parameters, such as defocus, electron dose, and magnification, is crucial for obtaining high-quality cryo-EM images [48]. Balancing signal-to-noise ratio (SNR) and radiation damage is important to preserve high-resolution information while minimizing noise and specimen degradation.

***Computational Methods*:** High-resolution cryo-EM relies on advanced computational methods for image processing, including particle picking, alignment, and classification [32]. These methods involve sophisticated algorithms and powerful computational resources. Choosing appropriate reconstruction algorithms and optimizing parameters can significantly impact the final resolution achieved.

***Expertise and Experience*:** Cryo-EM is a technically demanding technique that requires expertise and experience. Knowledge of sample preparation techniques, data collection strategies, and computational methods is crucial for obtaining high-resolution reconstructions [44].

While SPA has been widely used for a longer period and has a more extensive history, tomography and subtomogram averaging have also made significant progress and have become fields in their own right. Subtomogram averaging has emerged as a powerful technique to enhance the resolution and quality of specific subvolumes within tomographic reconstructions. It has been successfully applied to study various macromolecular complexes, such as viral particles, ribosomes, and membrane proteins, enabling the elucidation of their structural features and functional mechanisms.

Both tomography and subtomogram averaging have seen advancements in experimental techniques, data acquisition, image processing algorithms, and computational resources. Numerous software tools and pipelines have been developed specifically for tomography and subtomogram averaging, enabling researchers to carry out these analyses more efficiently and with greater accuracy, as they offer unique advantages in studying complex biological structures and dynamics within their native cellular context, complementing the insights provided by single particle analysis. However, software and pipelines are not as standardized as in SPA.

1.2.1. Challenging samples

Cryo-EM may face specific challenges when imaging and analyzing certain types of samples. Some of the challenging samples in cryo-EM include:

***Flexible or Dynamic Molecules [7]*:** Biological macromolecules that exhibit significant conformational flexibility or undergo dynamic structural changes pose challenges in cryo-EM. These molecules can adopt multiple conformations or states, making it difficult to align particles accurately and obtain high-resolution reconstructions. This topic is quite interesting and promising, but it is out of the scope of this thesis.

***Heterogeneous Complexes [15]*:** Macromolecular complexes with compositional heterogeneity, where different subunits or components adopt distinct conformations or binding states, can present challenges. This heterogeneity leads to variations in particle appearance and makes it challenging to align and classify particles accurately, resulting in lower-resolution reconstructions.

***Small Specimens [51]*:** Cryo-EM struggles to achieve high-resolution structures for small particles, typically below 100 kDa in molecular weight. The limited number of particles and their low contrast makes obtaining high-resolution reconstructions very challenging. Advanced data acquisition strategies, such as higher electron energies or phase plates, along with improved image processing methods, are being developed to address this challenge. Other techniques are applied when preparing the sample, such as the addition of antibodies in other to add molecular weight to the sample. However, these extra components may complicate the image processing.

***Membrane Proteins [1]*:** Membrane proteins, which are embedded in lipid bilayers, can pose challenges in cryo-EM due to their hydrophobic nature. Maintaining the stability and integrity of membrane proteins during sample preparation, preventing aggregation, and obtaining high-quality images with well-defined lipid environments are ongoing challenges in cryo-EM. It is usual to reconstruct the specimen in nanodiscs to stabilize it, but again, this may complicate the processing.

***Large Assemblies [25]*:** Large macromolecular complexes, such as viral capsids or large protein assemblies, can present challenges in cryo-EM due to their size and structural complexity. The sheer volume of particles, conformational heterogeneity, and the need for accurate alignment and classification pose significant computational and analytical challenges.

Addressing these sample-specific challenges often requires advanced data analysis and image processing strategies. Continued advancements in cryo-EM methodologies are aimed at overcoming these challenges and expanding the applicability of the technique to a broader range of complex biological samples.

1.3. Objectives of the Thesis

The main objective of the present thesis is the development of algorithms that help in the processing of challenging samples by means of signal subtraction, in order to be able to remove the signal in the cryo-EM data that comes from non-of-interest components, such as heterogeneous subunits, extra components (antibodies, nanodiscs, etc) or to be able to isolate small regions of the specimen to improve the resolution in difficult areas. Signal subtraction algorithms have been developed for volumes (post-process of refined volume) and for particles (bidimensional images), which allow their re-classification and the refinement of their angles. This contribution has been mainly done in the SPA field, even though it has been extended in a limited way for Tomography.

A secondary objective of the thesis is to contribute to the analysis of the robustness (both statistically and in terms of stability) of the existing methods, as this is an emergent issue in the cryo-EM field. As this is a quite broad objective, the work has focused on a specific step which is the "Local Defocus Refinement" (explained in Section 2.2). This step has been chosen as it has not yet been analyzed in the literature, even though it is at the end of the cryo-EM image processing pipeline and thus, it is used it to try to push forward the resolution of their final volume. However, that improvement it is not always obtained.

Finally, an underlying objective of this thesis is the contribution to the stabilization and standardization of image processing workflows in SPA and STA, through the contribution to the management, maintenance and testing of Xmipp and Scipion software.

2. Methodologies and Results

This thesis is presented as a compendium of articles, which collects the main contributions of the work. Three articles are already published in "Journal of Structural Biology" (JSB, impact factor of 3), with E. Fernandez-Gimenez as main author. There are also other smaller contributions to the field, including publications as co-author and others not published in the literature.

The code that implements the algorithms in this thesis has been developed and added to Xmipp Software Package for Cryo-EM, which is written in C++, and it has been integrated into the user-friendly framework for workflows in cryo-EM Scipion, written in Python.

All the software developed is open source code and it is included in `https://github.com/I2PC` (Xmipp software) and `https://github.com/scipion-em` (Scipion software).

2.1. Signal subtraction

As described in the introduction, there are challenging samples in cryo-EM with characteristics that make them need extra attention when performing the image processing pipeline. In samples with different subunits (especially if they are heterogeneous), with membranes, or with extra components added on purpose (such as antibodies or nanodiscs), it is usual that the alignment and classification be most driven by the component with the intense signal, obtaining a good reconstruction of it, but leaving the regions of the macromolecule with lower signal (due to its smaller size, bigger flexibility or other causes) poorly resolved. In other situations, we just want to focus on a specific region of a bigger macromolecule, for example, a specific protein of a virus.

In all these cases it seems that subtraction of the unwanted component may help. However, image subtraction is not trivial, because even though we can visually perceive that two images look the same, they may not be equal as they may have different ranges of intensity values. This issue is harder in cryo-EM image processing, due to the recorded images being noisy, possibly not perfectly aligned due to several causes and affected by the CTF that might have been imperfectly corrected. Hence, a previous step to subtraction, what we call *adjustment*, is required in order to obtain the real differences once we subtract.

In subsections 2.1.1 and 2.1.2, the algorithms developed in this thesis for volume and particle adjustment and subtraction will be introduced, discussing the main results achieved. Full work is in the papers included in Appendixes A and B respectively.

2.1.1. Volume Adjustment and Subtraction

Volume subtraction appeared to be a useful tool in the cryo-EM image processing pipeline. Nevertheless, the most important software packages in the field do not have an accurate implementation of it. That is the main motivation that led to the development of our own algorithm for volume subtraction in cryo-EM.

As explained before, adjustment of volumes before subtraction is crucial to obtain good results, that is, to determine the real differences in the structure of the macromolecule of interest, such as differences in specific regions due to conformational changes, ligands, presence or absence of a specific protein, etc. To perform this adjustment in the case of volumes, both volumes that will be subtracted must have the same size (box size) and same pixel size (sampling rate), they must have the same origin of coordinates and they must be aligned. The reconstructed volumes in cryo-EM have the CTF corrected, as this happened at the particle level when computing the reconstruction, and thus, CTF is not a problem.

The adjustment realized by the developed algorithm for volumes is based on the idea that the most relevant and detailed information of an image (in this case, a volume) is contained in the phase, rather than in the amplitude of the complex numerical values of the Fourier Transform of the image (the 'phase problem' [40], see Fig. 5). Thus, in the algorithm the amplitude of one of the volumes that we want to subtract is used as a reference to adjust the amplitude of the other in Fourier Space, in order to have images in the same range of intensity values for subtraction in Real Space. But then we need to recover the phase, to not lose the original differences in the shape of the adjusted volume. This idea is implemented through operations based on Projectors Onto Convex Sets (POCS)[18].

Once the 'numerical adjustment' of volumes is obtained, the subtraction is performed, revealing the differences in the structure of the volumes better than the only competitor software (ChimeraX [13]) found for volume subtraction in cryo-EM, which does not perform any numerical adjustment (see Fig. 6).

The adjustment is not only needed for subtraction, but it can be useful for other applications, such as sharpening and volume consensus. Sharpening consists in post-processing the obtained reconstruction in order to slightly improve its resolution, typically to make model building easier. There are several approaches for performing sharpening, some of which are based on prior knowledge extracted from already published structures of the same macro-molecule or from a similar one. Thus, if an atomic structure previously obtained is converted into a cryo-EM density map, it can be used as a reference volume to adjust the reconstruction, obtaining a sharpened version of it. This procedure has been compared with state-of-the-art sharpening methods, exhibiting improved results. Obviously, this method has the drawback that a previous atomic structure is needed (even though it is common in the field to have it in public databases), however, this drawback is shared with other state-of-the-art methods [41], [16], [27] (see Fig. 7).

The adjustment allows comparing objectively volumes of the same macromolecule coming from different reconstruction algorithms (whose outputs typically have different pixel value ranges despite having the same inputs), even though they come from other techniques, such as SPA (typically higher resolution) and STA (native state), as it is illustrated in [11]. Several reconstructions of the same macromolecule, where typically none of them is clearly better than the others, but some exhibits some regions better resolved than others, can

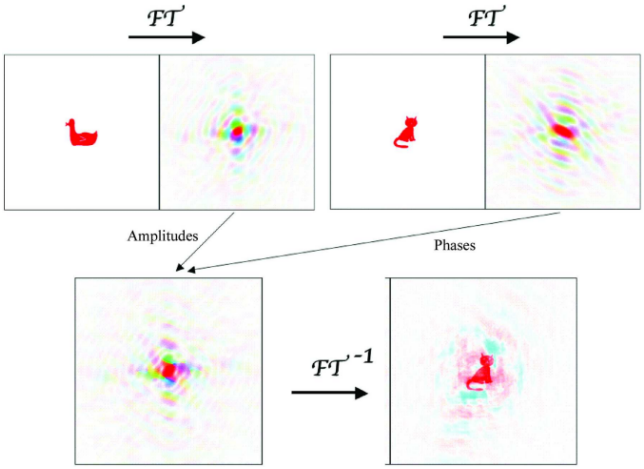


Figure 5: 'Phase problem' [40] (reproduced with permission): when the amplitude of the Fourier Transform of a duck image is combined with the phase of the Fourier Transform of a cat image and then, the Inverse Fourier Transform is computed, the image recovered is similar to the original cat image more than the duck one, illustrating that the main information about the shape of the object resides in the phase more than in the amplitude of the Fourier Transform.
.

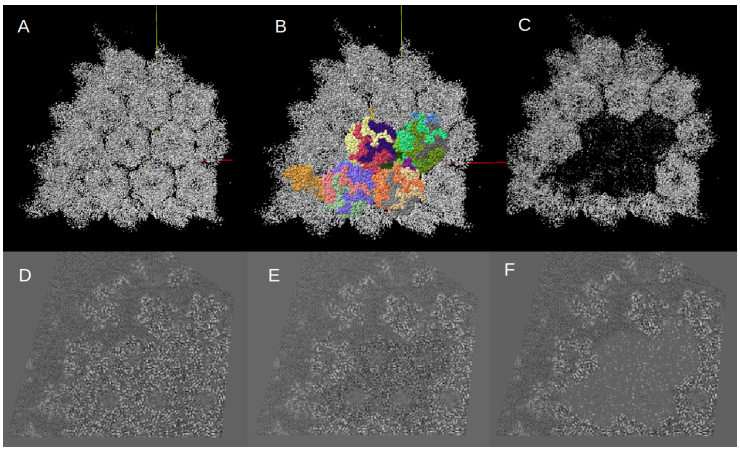


Figure 6: A) Reference map extracted from the capsid of the human adenovirus 41 (HAdV-F41) (EMD-10768). B) AdV-F41 ASU atomic structure (PDB ID6YBA) fitted to the reference map. C) Subtraction result of reference map in A minus the map derived from the conversion of the atomic model in B with the proposed algorithm. D) Central slice of the reference map. E) central slice of the subtraction computed by ChimeraX. F) Central slice of C.

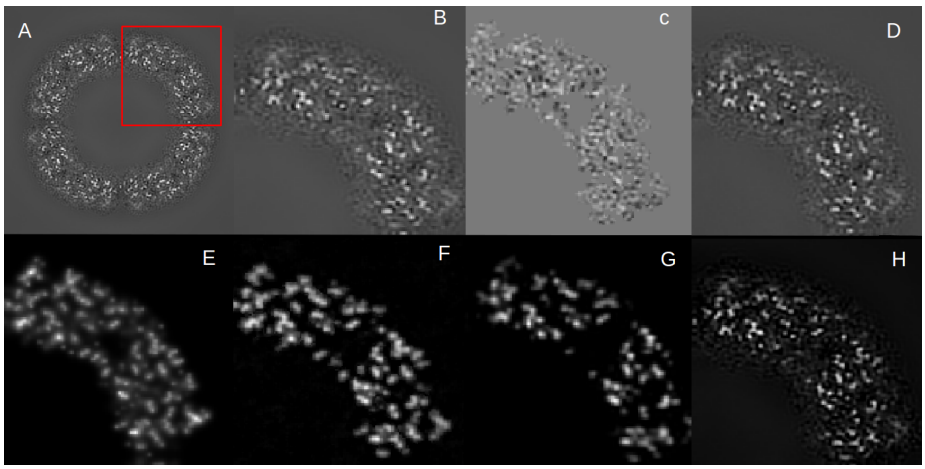


Figure 7: Central slices of A) Apoferritin density map EMD-11122. B) Detail of the region remarked in red in A. C) Result of LocScale. D) Result of Phenix sharpening. E) Apoferritin density map derived from atomic model PDB ID 6Z9F. F) Result of the proposed algorithm as sharpening method low pass filtered at the input map resolution (1.56 Å). G) Result of DeepEMHancer. H) Result of LocalDeBlur.
.

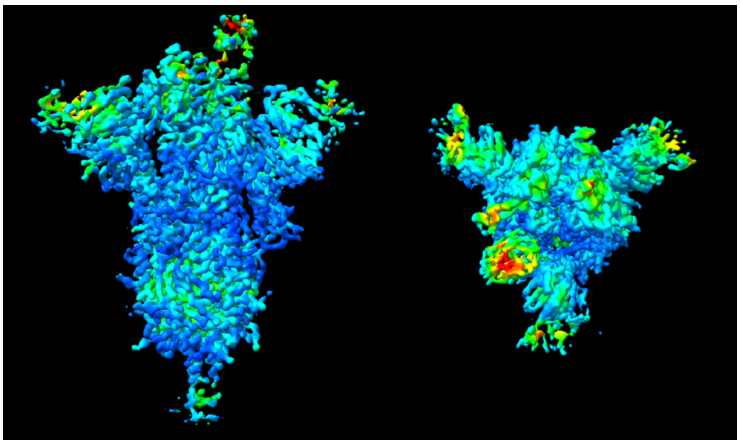


Figure 8: Result of the consensus volume from five reconstructions of the SARS-CoV-2 Spike, side (left) and top (right) views. The colors represent the degree of similarity of the input map signal, being dark blue the smallest difference between input volumes and red being the biggest difference.

be numerically adjusted to have them in a common framework, which gives us 'equivalent' volumes that can be actually combined taking the best part of each, obtaining a consensus volume.

The *consensus volume algorithm* has been developed by combining the adjusted volumes with the Wavelet Transform [50]. The result obtained demonstrates the validity and utility of this method, which to the best of our knowledge, does not have any equivalent in cryo-EM.

Finally, the adjustment and volume subtraction algorithm has been extended to be used for subtomograms, as they are also volumes. In this case, the software includes a pre-processing step to align each subtomogram to the reference volume before the alignment and subtraction.

2.1.2. Projection Subtraction

Volume subtraction appears to be a very useful tool in the cryo-EM image processing pipeline. However, it can be applied only to volumes, which appear just in two parts of the pipeline: as the initial volume (reference) for reconstruction and at the end of the pipeline as the final reconstruction, which can be post-process.

Hence, volume subtraction does not solve completely the need for subtraction in cryo-EM. On many occasions, the signal may be subtracted at the particle level, previous to the reconstruction. This is known in the field as 'projection subtraction', as cryo-EM picked particles are bidimensional and they are understood as 2D projections of the macromolecule we want to reconstruct. Projection subtraction is a must, especially for the challenging samples mentioned in the

introduction. That is why in this case, the leading software packages in the field (namely CryoSPARC, Relion and Xmipp) have their own approaches. However, we have tested their projection subtraction implementations and it turns out that their algorithms may be improved (as we show in [12]). Therefore, in this thesis, the development of a new projection subtraction algorithm in Xmipp has been undertaken.

For the projection subtraction algorithm, a first approach was implemented using a similar strategy to volume subtraction (adjustment with POCS and the recovery of the original phase previous to subtraction). However, the adjustment was not as accurate as desired as particles have issues that do not appear in volume subtraction. As a reference, we have a volume where the region to keep or subtract is designed by a three-dimensional mask, while no such thing is available for particles. Nevertheless, that volume has to be projected to have a two-dimensional reference image (projection) which will be subtracted from the particle. The alignment of each particle to the reference volume is different for each of them, as they are in random orientations (see Fig. 9). To carry out the projections we need to have a previous reconstruction and alignment of the particles before performing projection subtraction. Secondly, the two-dimensional projection of the volume has different characteristics than the particle, as the projection is not affected by the CTF (it has been already corrected when reconstructing the volume), but in the particle, the CTF has not been corrected already. Moreover, the projection coming from a reconstructed volume is not affected by noise, however, the particle that has been extracted from the micrograph is usually highly affected by noise and has a lower SNR.

All these reasons make the adjustment more challenging than in the case of volumes. Therefore, an approach based on linear regression was implemented to adjust the projections of the volume. A linear model of order zero and another of order one are computed for each particle. Then the one with the best fitting is applied to the projection, once the parameters of the CTF estimated for the particle have been applied. The projection subtraction algorithm of Relion [17], uses a similar approach to the order zero model (but it is computed in another way both conceptually and mathematically, as explained more in detail in [12]) and the same model is applied to all the particles that have been extracted from the same micrograph. Thus, theoretically, our algorithm should be more precise as each particle is adjusted individually. We can not compare the method used by CryoSPARC [26] as its software is not open source and the mathematics behind it has not been published.

In [12], we compared the results of the three algorithms by subtracting the large subunit of a ribosome to improve the resolution in the reconstruction of the small subunit. The results for Relion and the algorithm presented are quite similar and better than the results of CryoSPARC. To evaluate more in detail the performance of Relion and our algorithm, we chose a smaller target in a crowded environment, a specific protein of the capsid of Adenovirus. In this more difficult case, our algorithm gives better results than Relion, because it can remove more signals (see Fig. 10).

The algorithm has been also tested for the subtraction of a nanodisc using

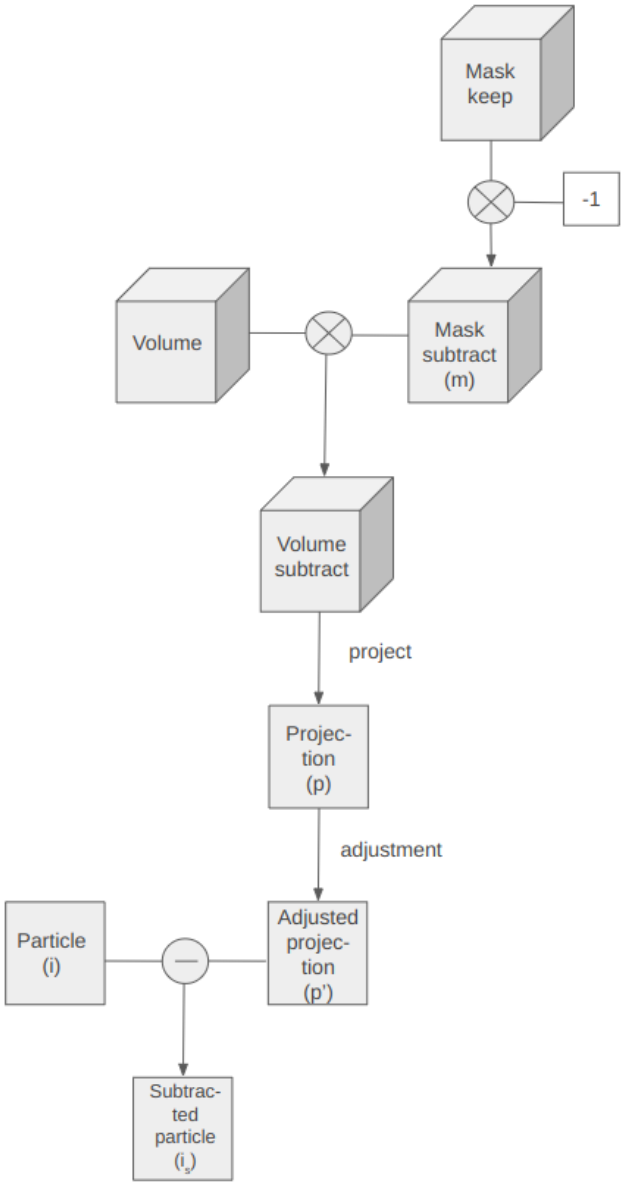


Figure 9: Subtraction process schema. The input volume is masked with a mask that defines the region to subtract (m). However, the user can input a mask of the region to keep and its inverse will be automatically computed to obtain m. The subtraction volume is then projected generating p, which will be adjusted (p') to be subtracted to the input particle i, obtaining as a result the subtracted particle i_s.

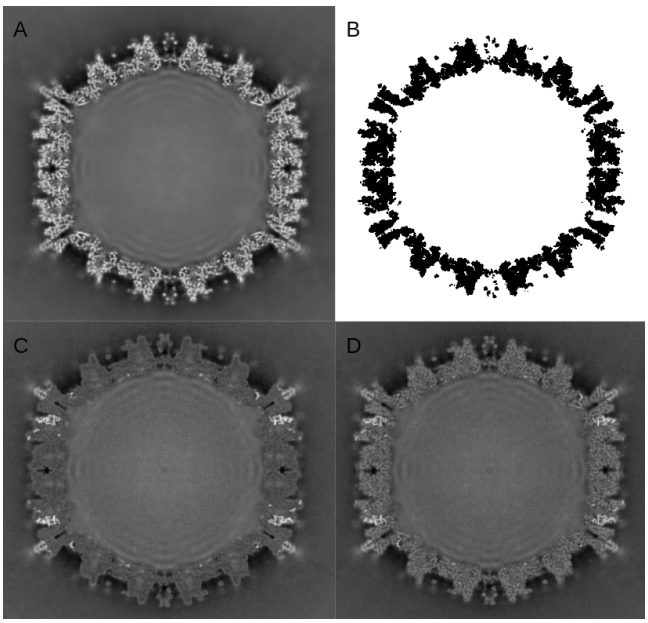


Figure 10: Human Adenovirus (central slice) (A) Reconstruction without subtraction (B) Mask of the capsid (in black is the region to subtract) (C) Map reconstruction of subtracted particles by Xmipp. (D) Map reconstruction of subtracted particles by Relion.

a sample from a public database, getting better local resolution than the one reported in the published data (see Fig. 11). In this case, Relion was not able to perform the subtraction correctly, since the reconstruction of the resulting particles has many artifacts. CryoSPARC subtraction was also not able to recover the protein properly.

Finally, the algorithm in [12] was tested to try to localize a small ligand (usually a drug) bound to a protein without prior knowledge the binding site. This entails of subtracting the reconstruction of a protein without a ligand from a reconstruction of the protein with the ligand, in order to remove the signal of the protein and leave just the small ligand. To do so, it is important to get high-resolution reconstructions as the ligand would be otherwise imperceptible. Our algorithm was able to subtract correctly the protein and keep the small density corresponding to the ligand, both with synthetic and experimental data (see Fig. 12). In this case, CryoSPARC was not able to remove correctly all the signals from the protein nor to keep the signal of the ligand. Relion cannot be used for this application as it requires as input the mask of the part to subtract and, in this application, the location of the ligand is unknown.

In conclusion, the algorithm that we have developed for projection subtraction showed improved results in comparison with the competitors' algorithms, Relion and CryoSPARC, and it is usable in more applications than the others, where it has demonstrated better performance.

2.2. Local Defocus Refinement Analysis

Acquiring images out of focus is a method broadly used in cryo-EM. It allows increasing tne phase contrast of the images, to the detriment of the resolution. However, defocus can be corrected by estimating it when estimating the CTF. Typically, the CTF (and thus, the defocus) is estimated and corrected for each micrograph, supposing that all the particles extracted from the same micrograph have the same CTF and defocus parameters. The acquired images (micrographs) are two-dimensional. However, the sample has a certain ice thickness, which is bigger than the diameter of the purified specimen. This means that the individual particles can be placed at different 'heights' inside the sample, having a slightly different defocus, which is called the local defocus (see Fig. 13).

Once we have obtained a high-resolution reconstruction, we can compute the defocus per particle in order to correct more precisely that defocus and, hopefully, increase the resolution of our reconstruction. We observed that local defocus is a tool used to try to push forward the resolution of an already high-resolution reconstruction, but it does not seem to work all the time. There are four state-of-the-art methods capable of estimating the local defocus included in the most used software packages fro cryo-EM, namely Relion [17], CryoSPARC [26], GCTF [54] and Xmipp [39].

In order to contribute to this emerging topic in the field of validation and robustness of the methods, we have analyzed the local defocus estimations computed by the four different methods, revealing that they do not agree (however we cannot know which of them, if any, is correct, as we do not know the ground

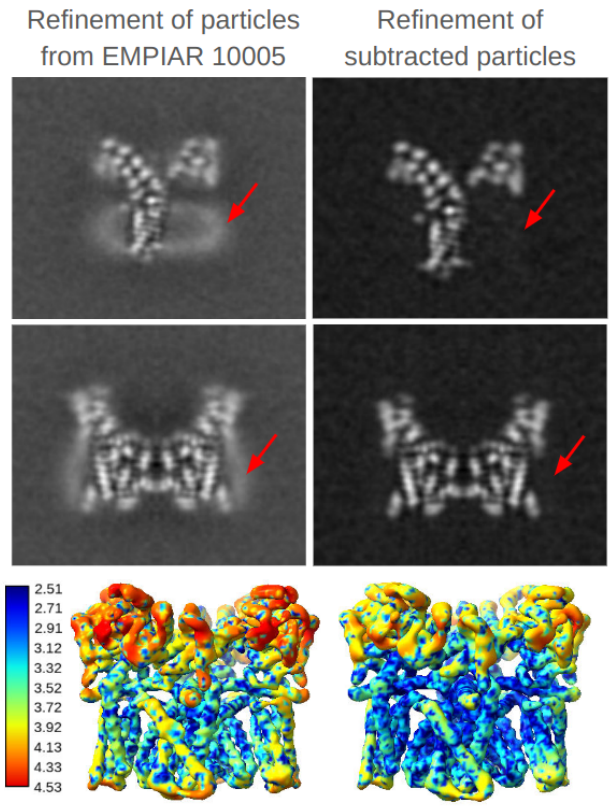


Figure 11: (Top) The image displays two different slices from the volumes obtained through the refinement of particles in EMPIAR 10005. The left side represents the particles before undergoing subtraction by Xmipp, while the right side shows the particles after. The same refinement parameters have been applied. The red arrows highlight the signal produced by the nanodisc, which has been eliminated in the subtracted case. (Bottom) Local resolution measures with MonoRes [46] of the refined volumes for (left) particles in EMPIAR 10005 and (right) the same particles once Xmipp has subtracted them.

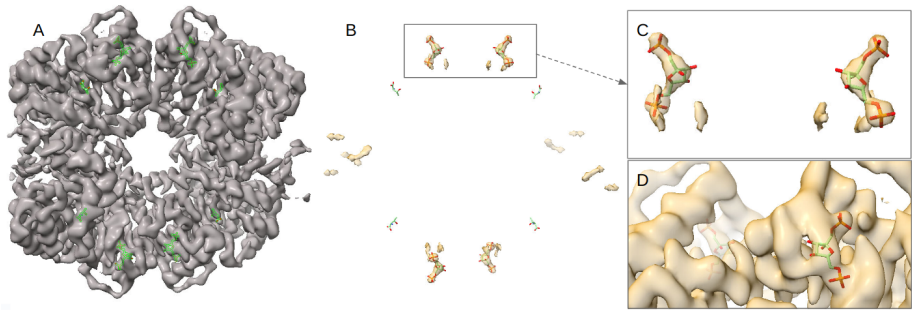


Figure 12: (Top) The image displays two different slices from the volumes obtained through the refinement of particles in EMPIAR 10005. The left side represents the particles before undergoing subtraction by Xmipp, while the right side shows the particles after. The same refinement parameters have been applied. The red arrows highlight the signal produced by the nanodisc, which has been eliminated in the subtracted case. (Bottom) Local resolution measures with MonoRes [46] of the refined volumes for (left) particles in EMPIAR 10005 and (right) the same particles once Xmipp has subtracted them.

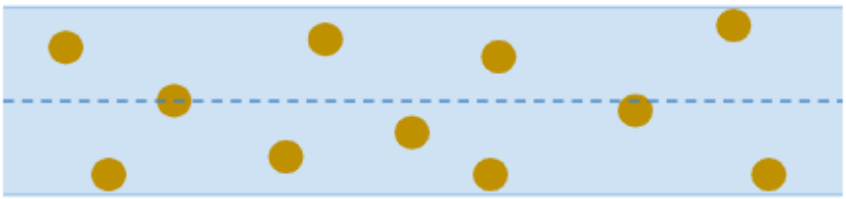


Figure 13: Distribution of proteins along the ice thickness of the sample [23]. The dotted line represents the global defocus estimation for the micrograph. However, the height of each particle (schematically represented by yellow dots) in the sample does not agree in many cases with the height corresponding to the global defocus.

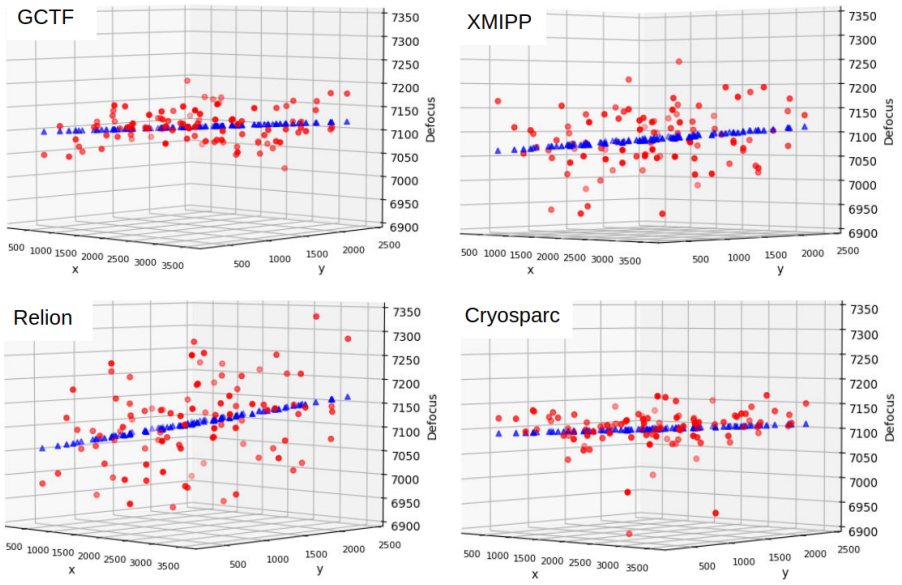


Figure 14: Plots computed by the developed analysis protocol of the distribution of all the particles in red (coordinates X and Y are the positions of the particle in the micrograph, while Z coordinate corresponds to the local defocus estimated for the particle in Å) along the thickness of the sample for a particular micrograph of the data set according to the different local defocus estimation software. Note that small differences in the rotation of the axes in each subplot have been made for convenience to better visualize the adjustment plane (blue) in each case.

truth of local defocus). To perform this study, an ad hoc analysis protocol was developed. The methods tend to report similar values, but with considerable differences that are big enough in a refinement step like this that try to increase precision (see Fig. 14 and the complete work in Appendix C). Moreover, the stability of the methods has been evaluated showing that they are not quite stable, meaning that the values reported have noisy fluctuations.

2.3. Other Contributions

In addition to the major contributions listed above, this work has contributed to the general development, maintenance and testing of the Xmipp software package and Scipion software framework, both in its SPA and tomography parts. Moreover, several software tools for specific tasks required by the scientists in the group have been developed, mainly related to coordinates management both in SPA (program for shifting coordinates concerning a volume) and tomography-STA (fit coordinates in vesicles to ellipsoids, filter coordinates by its normal

direction and/or tilt angle, visualization of subtomogram average in original coordinates: 'Map-back').

These tools are not published specifically in any publication as they do not have enough entities individually, however, they are somehow collected in the papers mentioned in Section 3.2. All of them have been done on purpose for real and specific needs in the image processing pipeline both in SPA and tomography-STA. The tools have been used by scientists in real projects and they are freely available and published at Xmipp and Scipion GitHub repositories.

3. List of Publications

3.1. Authored Publications

In the following publications, I am the author of the developed algorithms, I have run the experiments and analyzed the results. I have also prepared the text and participated in the evaluation.

1. **E. Fernández-Giménez**, M. Martínez, R. Sánchez-García, R. Marabini, E. Ramírez-Aportela, P. Conesa, J.M. Carazo and C.O.S. Sorzano. ***Cryo-EM density maps adjustment for subtraction, consensus and sharpening.*** Journal of Structural Biology, Volume 213, Issue 4, 2021, 107780, ISSN 1047-8477, https://doi.org/10.1016/j.jsb.2021.107780.

2. **E. Fernández-Giménez**, M. Martínez, R. Marabini, D. Strelak, R. Sánchez-García, J.M. Carazo and C.O.S. Sorzano. ***A new algorithm for particle weighted subtraction to eliminate signals from unwanted components in Single Particle Analysis*** Journal of Structural Biology, Volume 215, Issue 4, 2023, 108024, ISSN 1047-8477, https://doi.org/10.1016/j.jsb.2023.108024.

3. **E. Fernández-Giménez**, J.M. Carazo and C.O.S. Sorzano. ***Local defocus estimation in Single Particle Analysis in Cryo-Electron Microscopy*** Journal of Structural Biology, Volume 215, Issue 4, 2023, 108030, ISSN 1047-8477, https://doi.org/10.1016/j.jsb.2023.108030.

3.2. Other Contributions

1. Strelak, D.; Jiménez-Moreno, A.; Vilas, J.L.; Ramírez-Aportela, E.; Sánchez -García, R.; Maluenda, D.; Vargas, J.; Herreros, D.; **Fernández-Giménez, E.**; de Isidro-Gómez, F.P.; Horacek, J.; Myska, D.; Horacek, M.; Conesa, P.; Fonseca-Reyna, Y.C.; Jiménez, J.; Martínez, M.; Harastani, M.; Jonić, S.; Filipovic, J.; Marabini, R.; Carazo, J.M. and Sorzano, C.O.S. ***Advances in Xmipp for Cryo–Electron Microscopy: From Xmipp to Scipion.*** Molecules 2021, 26, 6224. https://doi.org/10.3390/molecules26206224

   - I have participated in the general development, maintenance and testing of Xmipp Software Package and its corresponding plugin in Scipion Software Framework.

2. Sorzano, C. O. S., Vilas, J. L., Ramírez-Aportela, E., Krieger, J., del Hoyo, D., Herreros, D., **Fernandez-Giménez, E.**, Marchán, D., Macías, J. R., Sánchez, I., del Caño, L., Fonseca-Reyna, Y., Conesa, P., García-Mena, A., Burguet, J., García Condado, J., Méndez García, J, Martínez, M., Muñoz-Barrutia, A., Marabini, R., Vargas, J. and Carazo, J. M. ***Image processing tools for the validation of CryoEM maps, Faraday***

Discuss (The Royal Society of Chemistry) 2022, volume 240, issue 0, pages 210-227, doi 10.1039/D2FD00059H,
url: http://dx.doi.org/10.1039/D2FD00059H

- I have contributed to validation tools through the update of the reprojection comparison method by integrating on it the Projection Subtraction algorithm explained in this thesis.

3. J. Jiménez de la Morena, P. Conesa, Y.C. Fonseca, F.P. de Isidro-Gómez, D. Herreros, **E. Fernández-Giménez**, D. Strelak, E. Moebel, T.O. Buchholz, F. Jug, A. Martinez-Sanchez, M. Harastani, S. Jonic, J.J. Conesa, A. Cuervo, P. Losana, I. Sánchez, M. Iceta, L. del Cano, M. Gragera, R. Melero, G. Sharov, D. Castaño-Díez, A. Koster, J.G. Piccirillo, J.L. Vilas, J. Otón, R. Marabini, C.O.S. Sorzano and J.M. Carazo. ***ScipionTomo: Towards cryo-electron tomography software integration, reproducibility, and validation.*** Journal of Structural Biology, Volume 214, Issue 3, 2022, 107872, ISSN 1047-8477,
url: https://doi.org/10.1016/j.jsb.2022.107872.

   - I have participated in the general development, maintenance and testing of the tomography part of Scipion Software ('ScipionTomo'). I have developed several tomography tools for coordinates management, filtering and visualization. I have also integrated third-party software in different ScipionTomo plugins.

4. Sorrentino, S., Conesa, J. J., Cuervo, A., Melero, R., Martins, B., **Fernández-Giménez, E.**, de Isidro-Gomez, F. P., de la Morena, J., Studt, J. D., Sorzano, C. O. S., Eibauer, M., Carazo, J. M., and Medalia, O. (2021). ***Structural analysis of receptors and actin polarity in platelet protrusions.*** Proceedings of the National Academy of Sciences of the United States of America, 118(37), e2105004118.
https://doi.org/10.1073/pnas.2105004118

   - I have contributed to this work through the integration and support of tomography software in Scipion used to carry out this study and I have developed part of the software ('Map-back') used for rendering the results.

4. Conclusions

Firstly, I have developed two algorithms for signal subtraction (one for volumes and another for particles) and both have demonstrated improved results compared to the state-of-the-art methods.

Volume subtraction has been also extended to be used with subtomograms. Moreover, the adjustment strategy developed as a previous step to volume subtraction turned out to be used also as a 'sharpening' algorithm (achieving results at the state-of-the-art level) and has allowed the development of another algorithm for volume combining in a way that makes sense, producing a consensus volume, which is a tool that it was not developed in the field before.

The projection subtraction algorithm developed in this thesis considerably improves the results of the state-of-the-art algorithms in the field, especially in the most complex cases. Besides, it can be used in applications that the competitors' algorithms can not be used because of their limitations or because they did not perform well, such as nanodisc subtraction and small ligand detection.

On the other hand, I have contributed to the rising topic of validation and robustness of the methods in the field with a deep study of local defocus refinement methods, by an exhaustive comparative for which I have developed a tool to analyze their outputs, as the results from different algorithms differ considerably in their estimated values.

4.1. Future Work

As future work, I consider that it can be worth deep into the following topics:

- Explore with more examples the ability of volume and projection subtraction to detect small ligands, as it is a very useful tool for cryo-EM scientists that works with binding ligands, such as drugs. Actually, an approach for adjusting volumes locally is under development.
- Deep testing and improvement of subtomogram subtraction.
- Accelerate volume consensus algorithm as it takes a long time with large volumes and it consumes a high amount of RAM memory.

5. Conclusiones

En primer lugar, he desarrollado dos algoritmos para la sustracción de señales (uno para volúmenes y otro para partículas 2D) y ambos han demostrado mejores resultados en comparación con los métodos actuales.

Tambien, la resta de volúmenes se ha ampliado para poder usarse con subtomogramas. Además, la estrategia de ajuste desarrollada como paso previo a la sustracción de volúmenes, ha resultado ser útil también como un algoritmo de 'sharpening' (logrando resultados al nivel del estado del arte) y ha permitido desarrollar otro algoritmo para la combinación de volúmenes de manera que ésta tenga sentido, produciendo un volumen de consenso, herramienta que no se había desarrollado en el campo antes.

El algoritmo de sustracción de proyecciones desarrollado en esta tesis mejora considerablemente los resultados obtenidos por los otros algoritmos del campo, especialmente en los casos más complejos. Además, se puede usar en aplicaciones en que los algoritmos competidores no pueden ser usados por sus limitaciones o no funcionan bien, como la sustracción de nanodiscos y la detección de pequeños ligandos.

Por otro lado, he contribuido al creciente tema de la validación y robustez de los métodos en el campo con un estudio profundo de los métodos de refinamiento de desenfoque local, mediante una comparativa exhaustiva para la cual he desarrollado una herramienta para analizar sus resultados, ya queue los resultados de diferentes algoritmos difieren considerablemente en sus valores estimados.

5.1. Trabajo Futuro

Como trabajo futuro considero que puede valer la pena profundizar en los siguientes temas:

- Explorar con más ejemplos la capacidad de la resta de proyecciones para detectar pequeños ligandos, ya que es una herramienta muy útil para los científicos de crio-EM que trabajan con fármacos.
- Probar en mayor detalle y mejorar la sustracción de subtomogramas.
- Acelerar el algoritmo de consenso de volúmenes, ya que tarda mucho tiempo y consume gran cantidad de memoria RAM con volúmenes grandes.

References

[1] S Akbar, S Mozumder, and J Sengupta. «Retrospect and Prospect of Single Particle Cryo-Electron Microscopy: The Class of Integral Membrane Proteins as an Example.» In: *Journal of chemical information and modeling* 60 (5 2020), pp. 2448–2457. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.9b01015. ppublish.

[2] B Basanta et al. «A guided approach for subtomogram averaging of challenging macromolecular assemblies.» In: *Journal of structural biology: X* 4 (2020), p. 100041. ISSN: 2590-1524. DOI: 10.1016/j.yjsbx.2020.100041. epublish.

[3] Y Censor. «Binary steering in discrete tomography reconstruction with sequential and simultaneous iterative algorithms». In: *Linear Algebra and its Applications* 339 (2001), pp. 111–124.

[4] Y. Chen et al. «Fast and accurate reference-free alignment of subtomograms.» eng. In: *Journal of Structural Biology* 182 (2013), pp. 235–245.

[5] Y Cheng et al. «A primer to single-particle cryo-electron microscopy.» eng. In: *Cell* 161.3 (2015), pp. 438–449. DOI: 10.1016/j.cell.2015.03.050.

[6] M. E. Davison and F. A. Grunbaum. «Tomographic reconstruction with arbitrary directions». In: *Communications on Pure and Applied Mathematics* XXXIV (1981), pp. 77–120.

[7] M T Degiacomi. «Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space.» In: *Structure (London, England : 1993)* 27 (6 2019), 1034–1040.e3. ISSN: 1878-4186. DOI: 10.1016/j.str.2019.03.018.

[8] T Dodd, C Yan, and I Ivanov. «Simulation-Based Methods for Model Building and Refinement in Cryoelectron Microscopy.» In: *Journal of chemical information and modeling* 60 (5 2020), pp. 2470–2483. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.0c00087. ppublish.

[9] F Eisenstein, R Danev, and M Pilhofer. «Improved applicability and robustness of fast cryo-electron tomography data acquisition». In: *Journal of Structural Biology* 208.2 (2019), pp. 107–114. DOI: 10.1016/j.jsb.2019.08.006.

[10] D. Elmlund, S. N. Le, and H. Elmlund. «High-resolution cryo-EM: the nuts and bolts.» In: *Current opinion in structural biology* 46 (2017), pp. 1–6. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2017.03.003.

[11] E. Fernandez-Gimenez et al. «Cryo-EM density maps adjustment for subtraction, consensus and sharpening». In: *Journal of Structural Biology* 213.4 (2021), p. 107780.

[12] E. Fernández-Giménez et al. «A new algorithm for particle weighted subtraction to decrease signals from unwanted components in single particle analysis». In: *Journal of Structural Biology* 215.4 (2023), p. 108024. ISSN: 1047-8477.

[13] Thomas D. Goddard et al. «UCSF ChimeraX: Meeting modern challenges in visualization and analysis». In: *Protein Science* 27.1 (2017), pp. 14–25. DOI: 10.1002/pro.3235.

[14] Z. Huang and P. A. Penczek. «Application of template matching technique to particle detection in electron micrographs». In: *Journal of Structural Biology* 145 (2004), pp. 29–40.

[15] S L Ilca et al. «Localized reconstruction of subunits from electron cryomicroscopy images of macromolecular complexes.» In: *Nature communications* 6 (2015), p. 8843. ISSN: 2041-1723. DOI: 10.1038/ncomms9843.

[16] Arjen J Jakobi, Matthias Wilmanns, and Carsten Sachse. «Model-based local density sharpening of cryo-EM maps». In: *eLife* 6 (2017), e27131. ISSN: 2050-084X.

[17] D. Kimanius et al. «New tools for automated cryo-EM single-particle analysis in RELION-4.0». In: *Biochemical Journal* 478.24 (2021), pp. 4169–4185.

[18] V. J. Madisetti and D. Williams. *Digital signal processing handbook*. CRC Press, 1999.

[19] M. Martinez et al. «Integration of Cryo-EM Model Building Software in Scipion». In: *J. Chemical Information and Modelling* 60 (2020), pp. 2533–2540.

[20] A Martinez-Sanchez et al. «Template-free detection and classification of membrane-bound complexes in cryo-electron tomograms.» In: *Nature methods* 17 (2 2020), pp. 209–216. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0675-5.

[21] G. McMullan, A. R. Faruqi, and R. Henderson. «Methods in Enzymology. The Resolution Revolution: Recent Advances In cryoEM». In: Academic Press, 2016. Chap. Direct Electron Detectors, pp. 1–18.

[22] P. Neumann, A. Dickmanns, and R. Ficner. «Validating Resolution Revolution.» In: *Structure* (2018), pp. 785–795.

[23] A. J. Noble et al. «Routine single particle cryoEM sample and grid characterization by tomography.» In: *eLife* (2018).

[24] J. Pierson et al. «Toward visualization of nanomachines in their native cellular environment.» eng. In: *Histochem Cell Biol* 132.3 (2009), pp. 253–262. DOI: 10.1007/s00418-009-0622-0.

[25] G Pintilie et al. «Resolution and Probabilistic Models of Components in CryoEM Maps of Mature P22 Bacteriophage.» eng. In: *Biophys J* 110.4 (2016), pp. 827–839. DOI: 10.1016/j.bpj.2015.11.3522.

[26] A. Punjani et al. «cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.» In: *Nature methods* 14 (2017), pp. 290–296.

[27] Erney Ramírez-Aportela et al. «Automatic local resolution-based sharpening of cryo-EM maps». In: *Bioinformatics* 36.3 (Aug. 2019), pp. 765–772. ISSN: 1367-4803.

[28] P B Rosenthal and J L Rubinstein. «Validating maps from single particle electron cryomicroscopy.» In: *Current opinion in structural biology* 34 (2015), pp. 135–144. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2015.07.002.

[29] 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.

[30] M Schaffer et al. «Optimized cryo-focused ion beam sample preparation aimed at in situ structural studies of membrane proteins.» In: *Journal of structural biology* 197 (2 2017), pp. 73–82. ISSN: 1095-8657. DOI: 10.1016/j.jsb.2016.07.010.

[31] F K Schur. «Toward high-resolution in situ structural biology with cryo-electron tomography and subtomogram averaging.» In: *Current opinion in structural biology* 58 (2019), pp. 1–9. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2019.03.018.

[32] Justin T. Seffernick and Steffen Lindert. «Hybrid methods for combined experimental and computational determination of protein structure.» In: *The Journal of chemical physics* 153 (Dec. 2020), p. 240901. ISSN: 1089-7690.

[33] C O S Sorzano et al. «A new algorithm for high-resolution reconstruction of single particles by electron microscopy.» In: *Journal of structural biology* 204 (2018), pp. 329–337.

[34] C O S Sorzano et al. «Blind estimation of DED camera gain in Electron Microscopy.» In: *Journal of Structural Biology* 203 (2018), pp. 90–93. ISSN: 1095-8657.

[35] C O S Sorzano et al. «Semiautomatic, high-throughput, high-resolution protocol for three-dimensional reconstruction of single particles in electron microscopy.» In: *Methods in molecular biology (Clifton, N.J.)* 950 (2013), pp. 171–193. ISSN: 1940-6029. DOI: 10.1007/978-1-62703-137-0_11.

[36] C. O S Sorzano et al. «Automatic particle selection from electron micrographs using machine learning techniques.» In: *Journal of Structural Biology* 167.3 (2009), pp. 252–260. DOI: 10.1016/j.jsb.2009.06.011.

[37] C. O. S. Sorzano et al. «Improvements on marker-free images alignment for electron tomography». In: *J. Structural Biology X* 4 (2020), p. 100037.

[38] C. O. S. Sorzano et al. «Transfer function restoration in 3D electron microscopy via iterative data refinement». In: *Proc. of the Sixth Intl. Meeting on Fully Three-dimensional Image Reconstruction in Radiology and Nuclear Medicine*. 2001, pp. 133–136.

[39] D. Strelak et al. «Advances in Xmipp for Cryo-Electron Microscopy: From Xmipp to Scipion». In: *Molecules* 26.20 (2021).

[40] G Taylor. «The phase problem». In: *Acta Crystallographica Section D* 59.11 (2003), pp. 1881–1890. DOI: 10.1107/S0907444903017815.

[41] Thomas C Terwilliger et al. «Automated map sharpening by maximization of detail and connectivity.» In: *Acta crystallographica. Section D, Structural biology* 74 (Pt 6 June 2018), pp. 545–559. ISSN: 2059-7983.

[42] R F Thompson et al. «An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology.» In: *Methods (San Diego, Calif.)* 100 (2016), pp. 3–15. ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2016.02.017.

[43] B Turonova, L Marsalek, and P Slusallek. «On geometric artifacts in cryo electron tomography.» In: *Ultramicroscopy* 163 (2016), pp. 48–61. ISSN: 1879-2723. DOI: 10.1016/j.ultramic.2016.01.002.

[44] J Vargas et al. «Foil-hole and data image quality assessment in 3DEM: Towards high-throughput image acquisition in the electron microscope.» In: *Journal of structural biology* 196 (3 2016), pp. 515–524. ISSN: 1095-8657. DOI: 10.1016/j.jsb.2016.10.006.

[45] J L Vilas et al. «Advances in image processing for single-particle analysis by electron cryomicroscopy and challenges ahead.» In: *Current opinion in structural biology* 52 (2018), pp. 127–145. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2018.11.004.

[46] J. L. Vilas et al. «MonoRes: automatic and unbiased estimation of Local Resolution for electron microscopy Maps». In: *Structure* 26 (2018), pp. 337–344.

[47] F Wang et al. «DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM.» eng. In: *Journal of Structural Biology* 195.3 (2016), pp. 325–336. DOI: 10.1016/j.jsb.2016.07.006.

[48] Felix Weis and Wim J. H. Hagen. «Combining high throughput and high quality for cryo-electron microscopy data collection.» In: *Acta crystallographica. Section D, Structural biology* 76 (Aug. 2020), pp. 724–728. ISSN: 2059-7983.

[49] L Yu et al. «Projection-based volume alignment.» In: *Journal of structural biology* 182 (2 2013), pp. 93–105. ISSN: 1095-8657. DOI: 10.1016/j.jsb.2013.01.011.

[50] Dengsheng Zhang. «Wavelet Transform». In: *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval*. Cham: Springer International Publishing, 2019, pp. 35–44. ISBN: 978-3-030-17989-2.

[51] Y. Zhang et al. «Could Egg White Lysozyme be Solved by Single Particle Cryo-EM?» In: *Journal of chemical information and modeling* 60 (5 2020), pp. 2605–2613. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.9b01176. ppublish.

[52] S. Q. Zheng et al. «MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy.» In: *Nature methods* 14 (2017), pp. 331–332.

[53] J Zivanov, T Nakane, and S H W Scheres. «A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis.» In: *IUCrJ* 6 (Pt 1 2019), pp. 5–17. ISSN: 2052-2525. DOI: 10.1107/S205225251801463X.

[54] J. Zivanov, T. Nakane, and S.H.W. Scheres. «Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in Relion-3.1.» In: *IUCrJ* 7 (2020), pp. 253–267.

A. Cryo-EM density maps adjustment for subtraction, consensus and sharpening.

Journal of Structural Biology 213 (2021) 107780

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Cryo-EM density maps adjustment for subtraction, consensus and sharpening

E. Fernández-Giménez [a,b], M. Martínez [a], R. Sánchez-García [a], R. Marabini [b], E. Ramírez-Aportela [a], P. Conesa [a], J.M. Carazo [a], C.O.S. Sorzano [a,c,*]

[a] Centro Nac. Biotecnología (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain
[b] Univ. Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[c] Univ. San Pablo – CEU, Campus Urb. Montepríncipe, 28668 Boadilla del Monte, Madrid, Spain

ARTICLE INFO

Keywords:
Subtraction
Sharpening
Map fusion
SPA
Subtomogram averaging
Cryo-EM

ABSTRACT

Electron cryomicroscopy (cryo-EM) has emerged as a powerful structural biology instrument to solve near-atomic three-dimensional structures. Despite the fast growth in the number of density maps generated from cryo-EM data, comparison tools among these reconstructions are still lacking. Current proposals to compare cryo-EM data derived volumes perform map subtraction based on adjustment of each volume grey level to the same scale. We present here a more sophisticated way of adjusting the volumes before comparing, which implies adjustment of grey level scale and spectrum energy, but keeping phases intact inside a mask and imposing the results to be strictly positive. The adjustment that we propose leaves the volumes in the same numeric frame, allowing to perform operations among the adjusted volumes in a more reliable way. This adjustment can be a preliminary step for several applications such as comparison through subtraction, map sharpening, or combination of volumes through a consensus that selects the best resolved parts of each input map. Our development might also be used as a sharpening method using an atomic model as a reference. We illustrate the applicability of this algorithm with the reconstructions derived of several experimental examples. This algorithm is implemented in Xmipp software package and its applications are user-friendly accessible through the cryo-EM image processing framework Scipion.

1. Introduction

Cryo-EM is becoming a widely used technique for the determination of the atomic structure of proteins and macromolecular complexes. The number of density maps reconstructed from cryo-EM data is increasing both in 'Single Particle Analysis' (SPA) and 'Subtomogram Averaging' (StA). However, the comparison between these reconstructions is still an open problem in the field and usually implies volume subtraction.

To compare reconstructions, the volumes must be the same size and aligned, but also they must be in a common numerical frame and they should have comparable energy, both in real and Fourier space. There are recent proposals for comparing reconstructed volumes which perform volume subtraction as TemPy:DiffMap (Joseph et al., 2020) by making an amplitude scaling in Fourier space moving a small window along the two maps to be compared. This amplitude scaling can be performed globally, as was suggested by Terwilliger et al. (2020). Structure subtraction is also a key step for some angular alignment approaches, notably focused classification (Bai et al., 2015; Punjani et al., 2017).

In this article, we propose an enhancement of this approach to adjust the numerical values of the two volumes before subtracting and thus, we expect better subtraction results, and consequently, more accurate differences in densities. This difference can be performed between two cryoEM maps, a cryoEM map and an atomic model which will be converted internally into a density map, or between an SPA and an StA maps. Moreover, the proposed adjustment can be used for other applications. First, if we invert the roles of the reference (now the atomic model) and adjusted volume (now the SPA map), the adjustment operator acts as a map sharpener. Second, if we have several reconstructions of the same macromolecule (for example obtained from different reconstruction algorithms), we have observed that usually each of these reconstructions have parts better resolved than others, while other parts are worse. We may use the adjustment procedure to make sure that all maps are in the same numerical framework and then apply a map fusion

* Corresponding author at: Centro Nac. Biotecnología (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain.
E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

https://doi.org/10.1016/j.jsb.2021.107780
Received 13 April 2021; Received in revised form 4 August 2021; Accepted 9 August 2021
Available online 29 August 2021

technique based on wavelets. In this way, we have a volume that is a consensus of the input reconstructions, trying to keep the best quality parts of the different inputs. To the best of our knowledge, this is the first attempt to combine multiple reconstructions of the same structure obtained by different methods into a single map.

2. Methods

2.1. Numerical adjustment of volumes

We define the set of operations necessary to assimilate the values of one volume to another without losing the structural information as numerical adjustment of volumes. If this adjustment is not performed, the different gray scales and energies of the two volumes make a quantitative comparison impossible. Thus, to get comparable maps, we have developed an adjustment algorithm based on projectors onto sets, specifically being all of them, except for the first one, projectors onto convex sets (POCS) (Madisetti and Williams, 1999).

Lets note the two volumes to be adjusted as V_1 and V_2. V_1 will be chosen as the "reference" volume, while V_2 will be modified to be numerically adjusted to the reference volume. To perform this adjustment, we use five projectors in an iterative way, concentrating their effects to a given region in real space, to transform a map V_2 into another map whose numerical values are as close as possible to those of a reference map V_1, previously restricted by a non-negative constraint.

The first projector applies the Fourier amplitude of the reference volume to the input volume (which will be adjusted), in order to have both volumes with comparable energies. The amplitudes of the reference volume are not applied directly, but they are modulated by the quotient of the radial averages (denoted by a over line) of the Fourier magnitudes of the reference volume and the input volume. Given the Fourier transform of a map, $\widehat{V}(\boldsymbol{\omega})$, and the Fourier transform of the reference, $\widehat{V}_1(\boldsymbol{\omega})$, the proposed projector is

$$\mathscr{P}_1\left(\widehat{V}\left(\boldsymbol{\omega}\right)\right) = \widehat{V}\left(\boldsymbol{\omega}\right)\frac{\overline{\left|\widehat{V}_1\right|}\left(\omega\right)}{\overline{\left|\widehat{V}\right|}\left(\omega\right)} \tag{1}$$

This projector is similar to the approaches used in Jakobi et al. (2017) and Joseph et al. (2020).

The second projector limits the minimum and maximum gray value of the input volume in real space, $V(\mathbf{r})$, to take the minimum, m_1, and maximum, M_1 values of the reference volume.

$$\mathscr{P}_2(V(\mathbf{r})) = \min(\max(V(\mathbf{r}), m_1), M_1) \tag{2}$$

This second projector ensures that the range of gray of the adjusted map does not exceed that of the reference map.

The third projector considers two binary masks in real space, one for the reference map and another for the input map ($W_1(\mathbf{r})$ and $W_2(\mathbf{r})$), so masks must be the same size as the maps:

$$\mathscr{P}_3(V(\mathbf{r})) = W_1(\mathbf{r})W_2(\mathbf{r})V(\mathbf{r}) \tag{3}$$

This projector computes the intersection of the masks and apply the resultant mask to the modified volume. Thus, it concentrates the action of the rest of the projectors to a particular region in space, i. e. the region composed by the intersection of the masks.

The fourth projector takes the Fourier phase of the original V_2 and applies it to the current estimate of the adjusted map:

$$\mathscr{P}_4\left(\widehat{V}\left(\boldsymbol{\omega}\right)\right) = |\widehat{V}\left(\boldsymbol{\omega}\right)|\frac{\widehat{V}_2\left(\boldsymbol{\omega}\right)}{\left|\widehat{V}_2\left(\boldsymbol{\omega}\right)\right|} \tag{4}$$

The goal of this projector is to preserve the structural information of V_2 as faithfully as possible, as the phase of a map contains most of the three-dimensional information of the macromolecule.

Finally, the last projector imposes non-negativity to the adjusted map:

$$\mathscr{P}_5(V(\mathbf{r})) = \max(V(\mathbf{r}), 0) \tag{5}$$

The reason is that macromolecules should not have negative density values, and all the operations performed in Fourier space may induce some artificial ringing that should be eliminated. Although it is true that negative values can also be caused by other sources (as actual densities lower than the solvent density or imperfect CTF correction), at the point of making two structures as similar as possible while still preserving the structural details conveyed by the second structure, we cannot identify all these effects. Thus, we decided to clip negative values as they normally do not correspond to structural details of the macromolecule of interest if the background has been normalized to zero. This latter choice is not physically correct (as the Coulomb potential of the ice is not zero), but it is generally adopted as a way to prevent 3D reconstruction artifacts due to a non-zero background.

We start the iterations restricting the analysis to the common region defined by the masks W_1 and W_2 $V^{(0)} = \mathscr{P}_3(V_2(\mathbf{r}))$. Note that these masks are optional and if they are not provided, they are assumed to cover the whole input maps. Then, we sequentially apply the projectors described above. When needed we also include the Fourier and inverse Fourier transform operators ($\mathscr{F}$ and $\mathscr{F}^{-1}$). Given the current adjusted map at iteration k, $V^{(k)}(\mathbf{r})$ we produce the $k+1$-th iteration as

$$V^{(k+1)}\left(\mathbf{r}\right) = \left(\mathscr{P}_5\circ\mathscr{F}^{-1}\circ\mathscr{P}_4\circ\mathscr{F}\circ\mathscr{P}_3\circ\mathscr{P}_2\circ\mathscr{F}^{-1}\circ\mathscr{P}_1\circ\mathscr{F}\right)\left(V^{(k)}\left(\mathbf{r}\right)\right) \tag{6}$$

where $\circ$ denotes operator composition.

In summary, this is an iterative method which try to look for a non-negative volume that has the energy of V_1 in Fourier space, the phases of V_2, and whose minimum and maximum do not exceed the ones of V_1. All this search is performed in a region constrained by the masks W_1 and W_2.

Since the method usually converges after five iterations, this number of iterations has been selected by default. However, the number of iterations can be modified by the user if it is observed that convergence is not reached or if it is reached in less iterations. In order to know the degree of convergence after every projector, as well as after each iteration, the difference in terms of energy is estimated between the previous result and the current one. Since this value is reported right away, the user can estimate whether the process has converged or it needs some more iterations. At the end of the iterations, let us refer to the numerically adjusted volume as $\widetilde{V}_2(\mathbf{r})$.

2.2. Applications

In the following, we illustrate three possible uses of the numerical adjustment procedure described above. Each one differs in the problem to solve and the inputs to the procedure. In all cases, it is extremely important that the two maps to adjust have been spatially registered so that both are at the same location. The numerical adjustment procedure described above has been implemented in Xmipp (de la Rosa-Trevín et al., 2013), and the different applications described below are available through the cryo-EM image processing framework Scipion (de la Rosa-Trevín et al., 2016).

2.2.1. Volume subtraction

Volume subtraction is the first of the applications. In this problem we have two volumes V_1 and V_2 and we want to see where the differences between the two volumes are. Volume subtraction has been extensively used in the field to identify small proteins in viral capsids, factors, or ligands bound to a given macromolecule. They can also be used to describe conformational changes between two different states of the same macromolecule or the same structure solved by different techniques like single particle analysis and subtomogram averaging, or single particle analysis and X-ray diffraction.

We may perform the numerical adjustment at some desired resolution (for instance, we may limit the operations to the resolution of V_1, V_2, the minimum of both, or any other resolution of interest). Let us refer as $\widetilde{V}_1$ to the lowpass filtered V_1 volume, and as $\widetilde{V}_2$ to the numerically adjusted V_2 volume filtered to the same resolution as V_1. Then, the volume subtraction is performed as

$$\Delta V(\mathbf{r}) = V_1(\mathbf{r})\left(1 - W_1(\mathbf{r})W_2(\mathbf{r})\right) + \max\left(\widetilde{V}_1(\mathbf{r}) - \widetilde{V}_2(\mathbf{r}), 0\right)W_1(\mathbf{r})W_2(\mathbf{r}) \tag{7}$$

whose interpretation is "keep the original V_1 in those regions in which the two volumes have not been adjusted $(1 - W_1(\mathbf{r})W_2(\mathbf{r}))$, and compute the difference between the adjusted volumes in the remaining regions $(W_1(\mathbf{r})W_2(\mathbf{r}))$". Note that the masks W_1 and W_2 do not need to be hard, and soft masks make perfect sense.

2.2.2. Map sharpening

The numerical adjustment procedure described above can be used as sharpening method. If V_1 is an atomic model converted to a density map (Sorzano et al., 2015), and V_2 is the cryoEM map, then we may use the numerical adjustment to push the cryoEM map to have the same amplitude spectrum as the reference volume V_1, while keeping its original information, which is mostly stored in the phases of the Fourier coefficients of V_2. The result can be low pass filtered to the input resolution if the user suspect that over-sharpening is being produced. This approach of amplitude scaling, although similar to the one proposed by Terwilliger et al. (2020) and Jakobi et al. (2017), add some restraints like constraining the range of the sharpened map and the spatial region in which the adjustment is performed.

2.2.3. Map consensus

Very often, we have multiple 3D reconstructions of the same structure obtained from the same data, but using different 3D reconstruction methods like CryoSparc (Punjani et al., 2017), Relion (Scheres, 2012), or Xmipp (de la Rosa-Trevín et al., 2013; Sorzano et al., 2018). In general, none of the maps is superior to all others in all regions. Normally, we observe that some parts of the maps are better reconstructed in one of the maps, while some other parts are better preserved in some other map. A common upgraded map could include the best domains selected from all maps. In this work, we have addressed the fusion of the optimal parts of the maps into a single one based on the measure of their local quality. The local quality of the map may be estimated by local resolution (Vilas et al., 2018), the local similarity between the map and its atomic model (Ramírez-Aportela et al., 2021), or the local energy of its wavelet coefficients (Pajares and de la Cruz, 2004).

In any case, we may fuse the different maps into a single one using any measure of their local quality. In this work, we propose to do that by cherry-picking the coefficients of the wavelet transform of the volume depending on the local quality measure. This is a well-known image fusion technique (Pajares and de la Cruz, 2004). Let us consider a set of input volumes $V_1(\mathbf{r}), V_2(\mathbf{r}), \ldots, V_N(\mathbf{r})$. We refer to the wavelet transform at scale s of the input volume $V_n(\mathbf{r})$ as $\widehat{V}_n^s(\mathbf{r})$. Then, the fusion is performed by constructing a new wavelet transform that, at every location and scale, takes the wavelet coefficient from the volume with better local properties

$$\widehat{V}^s_{\text{consensus}}(\mathbf{r}) = \widehat{V}^s_{\Phi(s,\mathbf{r})}(\mathbf{r}) \tag{8}$$

where $\Phi(s,\mathbf{r})$ is a function that considers the local quality of the N input volumes at that scale, and returns the index (1, 2, …, N) from which the wavelet coefficient must be taken from. A typical operator used in image fusion, and the one used here for our experiments, is simply

$$\Phi(s,\mathbf{r}) = \underset{n}{argmax}\left|\widehat{V}_n^s(\mathbf{r})\right| \tag{9}$$

Finally the consensus map is obtained by merely inverting the 3D wavelet transform using the consensus coefficients.

In addition to being able to keep the best coefficients from all the input volumes, we may also compute the local disagreement of the different wavelet transforms so that we have an estimate of those regions where most maps agree as well as the regions in which they disagree. We propose to do so by the following qualifier:

$$\Delta(\mathbf{r}) = \max_s\left(\max_n\left|\widehat{V}_n^s(\mathbf{r})\right| - \min_n\left|\widehat{V}_n^s(\mathbf{r})\right|\right) \tag{10}$$

Note that for this fusion to work and obtain as result a quantitative fusion map, we need that all the input maps have similar numerical values, otherwise their wavelet coefficients cannot be freely combined into a single consensus wavelet transform as they are not comparable. The numerical adjustment procedure proposed in this paper has allowed us to successfully devise this map consensus operator as illustrated in Section 3.3.

3. Results and discussion

We note that our algorithm tries to match two input signals to have similar numerical values. For this reason, along this section we make a special emphasis on showing the slices of the resulting volumes in order to see the details in the signal itself (values of the pixels). In the isosurface representation of the volumes the signal is thresholded and, consequently, we cannot appreciate its internal details.

3.1. Subtraction

We have selected three examples of subtraction to illustrate the utility and versatility of the method, being the first one the subtraction between a cryo-EM map and a converted atomic model, the second example the subtraction between two different density maps, and the third example the difference between a converted atomic model and a map. In all cases, the second volume is adjusted to the first one, as described in Section 2.2.3 previously to the subtraction.

3.1.1. Map – Model: Asymmetric unit of human adenovirus 41

To generate the starting reference map, we extracted the map fraction used to model the asymmetric unit (ASU) of the atomic structure of the capsid of the human adenovirus 41 (HAdV-F41) (Pérez-Illana et al., 2021) (see Fig. 1A). Then, we fitted the ASU atomic model to the reference map (Fig. 1(B)). This fitted atomic model was then converted into a density map using electron atomic scattering factors (Sorzano et al., 2015). To have both maps in the same position, we assigned the origin of the reference map to the model-derived map which will be adjusted and subtracted from the reference map. After that, a binary mask was computed for each of the maps.

Once we have both maps in registration and the corresponding masks, we have applied our algorithm for adjustment and subtraction (described in Sections 2.1 and 2.2.1). The adjustment has been performed with five iterations. The reference and adjusted volumes have been low pass filtered at 4 Å of resolution previously to the subtraction, as this is the resolution of the input map and smallest differences will not be reliable. The low pass filter was implemented as a raised-cosine filter whose amplitude was 1 up to 4 Å and smoothly decays (as a cosine) to 0, at 2.9 Å.

The subtraction result is showed in Fig. 1C. As can be appreciated, most of the map region where the ASU was fitted has been removed. Fig. 1D shows the central slice of the reference volume, while Fig. 1F shows the central slice of the volume resulting from the subtraction performed by the proposed algorithm. The central slice of the result obtained with UCSF ChimeraX v1.1 (Goddard et al., 2017) is showed in Fig. 1E. It can be seen that the result of the proposed algorithm in the subtraction region is less noisy and the remaining density is clearly

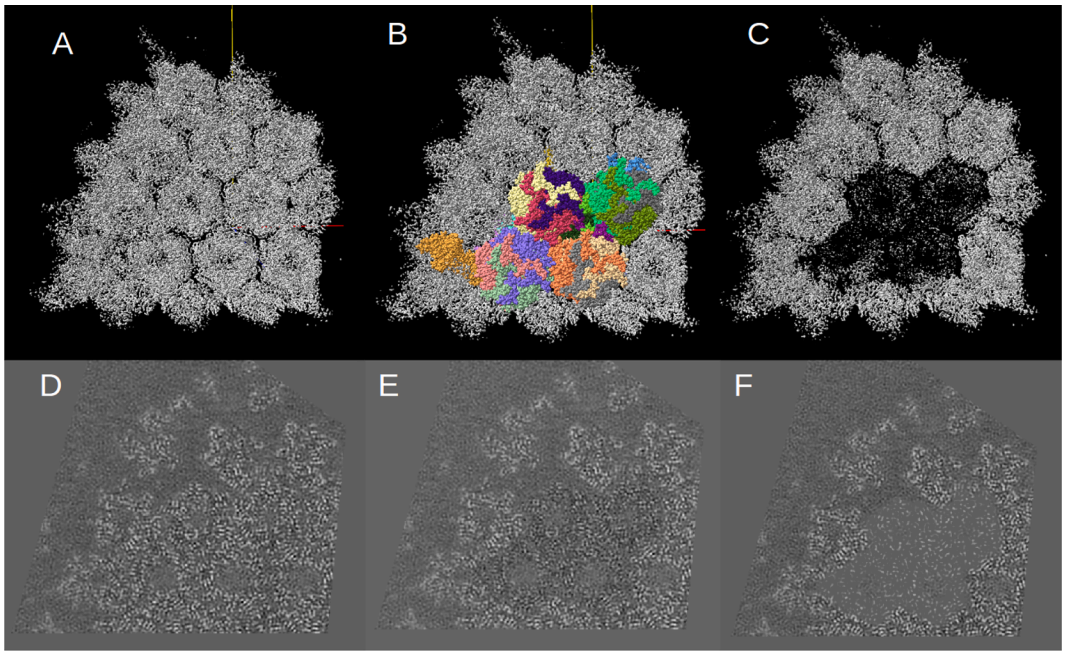


Fig. 1. A) Reference map extracted from the capsid of the human adenovirus 41 (HAdV-F41) (EMD-10768). B) AdV-F41 ASU atomic structure (PDB ID6YBA) fitted to the reference map. C) Subtraction result of reference map in A minus the map derived from the conversion of atomic model in B with the proposed algorithm. D) Central slice of reference map. E) central slice of the subtraction computed by ChimeraX. F) Central slice of C.

visible. In contrast, the result of ChimeraX is more noisy in the subtracted area and it is difficult to distinguish between remaining density and the noise.

In Fig. 2 it can also be observed that the result of ChimeraX subtraction is noisier (left column) as more small unconnected densities remains than in the result of the proposed algorithm (left column), even though all maps have been dust filtered. In the bottom part of Fig. 2 a low pass filter have been applied in order to strengthen the signal. To check if the signal is enough in the results, we focus in protein core V of the adenovirus (Rafie et al., 2021.), which is present in the reference map, but it is not traced in the atomic model of the ASU from which we derive the map to subtract. We have remarked this protein with a red rectangle in Fig. 2. As can be seen in the figure, signal is a bit higher in the case of ChimeraX low pass filtered map (C) in comparison with low pass filtered version of the difference map computed by proposed algorithm (D). Nevertheless, in the case of proposed algorithm the signal is enough to identify the lack of this protein in the reference map, easily in the low pass filtered version, but also in the original result. However, in the original result of ChimeraX even though there is more density in the region where the protein is, it is very difficult to know which densities correspond really to the protein and which are noise.

3.1.2. *Map – Map: Hepatitis-B viral capsid*

In this second experiment, we apply the subtraction between two reconstructed density maps. In order to highlight the versatility of the adjustment and subtraction algorithm, we have chosen a density map that comes from SPA reconstruction of the Hepatitis-B viral capsid as reference map and a second map of the same structure to subtract but, in this case, the map is a subtomogram average (StA) reconstruction.

For the SPA map we have chosen entry 21653 from EMDB, which has a box of 640 × 640 × 640, a voxel size of 0.65 Å and a resolution of 4.6 Å (see Fig. 3 left). The StA map comes from EMDB entry 3015, with a box of 240 × 240 × 240, a voxel size of 2.17 Å, and a resolution of 8.1 Å (see Fig. 3 right).

We have adjusted the StA map to the SPA map, as the SPA map has better resolution. The StA map was resized to have same box and pixel size than the SPA map. Then, the resized StA map was aligned to the SPA map. Once the maps were aligned, a binary tight mask was computed for each one. The adjustment was performed with five iterations. The subtraction was performed up to a resolution of 8.1 Å (the resolution of the StA map), meaning that adjusted volume has been low pass filtered at mentioned resolution previously to the subtraction. As can be seen in Fig. 4, there are differences all around the capsid, due to the difference of resolution in input maps, corroborating a lack of detail in StA map in comparison with SPA map. However, the StA reconstruction could show interesting differences in conformation due to the nature of the technique.

Thus, using this method it is possible to compare two maps of the same macromolecule obtained by different techniques and get reliable results in terms of densities as they have been numerically adjusted, which it is not the case if the subtraction is performed directly between the original SPA and StA maps as the pixel values of the maps probably will not be comparable. In Fig. 5 most notably differences between SPA and StA maps of the Hepatitis-B viral capsid are pointed by red arrows. This differences correspond to high frequency details and thus can be caused by the lack of the side chain densities in StA map due to the low resolution of the map.

3.1.3. *Model – Map*

In this section, we carried out the subtraction of the atomic model (which has been previously converted into a density map using a method based on Electron Atomic Scattering Factors, Sorzano et al. (2015)) of the envelope trimer protein of HIV BG505 in complex with the rabbit antibody E70 Fab (PDB ID6P62) from the experimental cryo-EM map (EMD-20259) from which the atomic structure was modeled.

A and B of Fig. 6 shows the central slices of the atomic model converted into density map and the experimental cryo-em map. From C to F of the same figure the results for ChimeraX, the proposed algorithm, TemPy:DiffMap method (Joseph et al., 2020) using local and global modes are shown respectively. As can be seen, the global approach of Tempy:Diff as well as ChimeraX have very noisy structures with many negative values (black pixels) and a very noisy background. For this reason, we focus the comparison of the results to the local approach of TemPy:DiffMap. The local DiffMap is noisier than the result of proposed algorithm, hindering what could be considered as real differences between the inputs and noise, also due to the presence of negative pixels

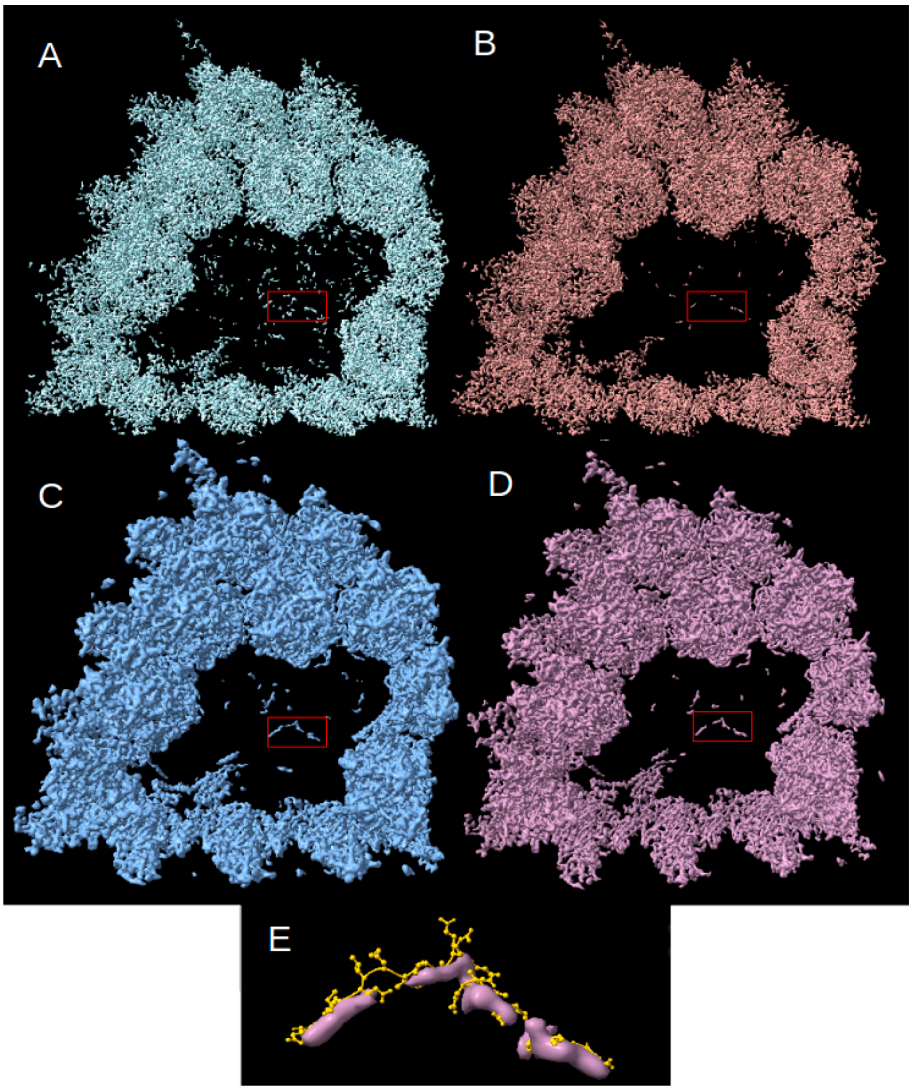


Fig. 2. Subtraction result of reference map minus map derived from ASU model performed by: A) ChimeraX subtraction B) proposed algorithm C) low pass filtered version of map in A D) low pass filtered version of map in B. Red rectangle enclosed core protein V, which is in the reference map but it is not traced in the model. E) Atomic structure of the adenovirus core V protein fitted into the density remarked in D.

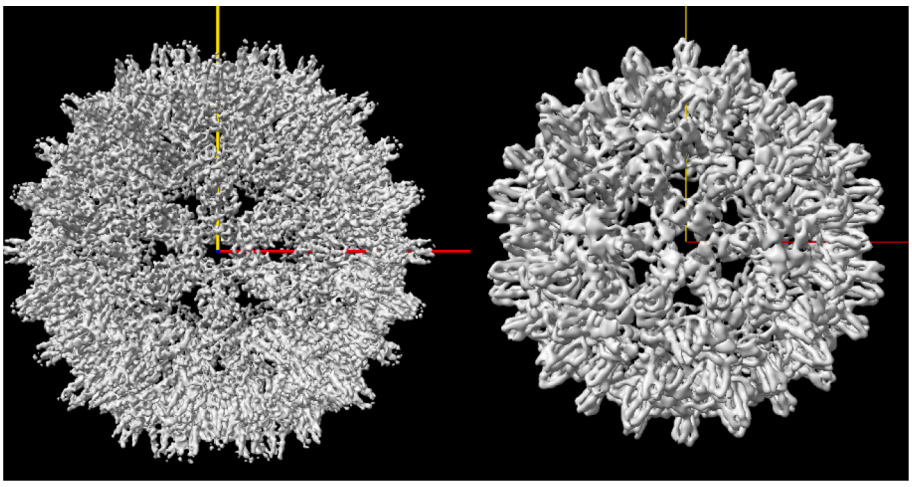


Fig. 3. Reconstruction of the Hepatitis B virus capsid using SPA (EMD-21653) (left) and StA (EMD-3015) (right).

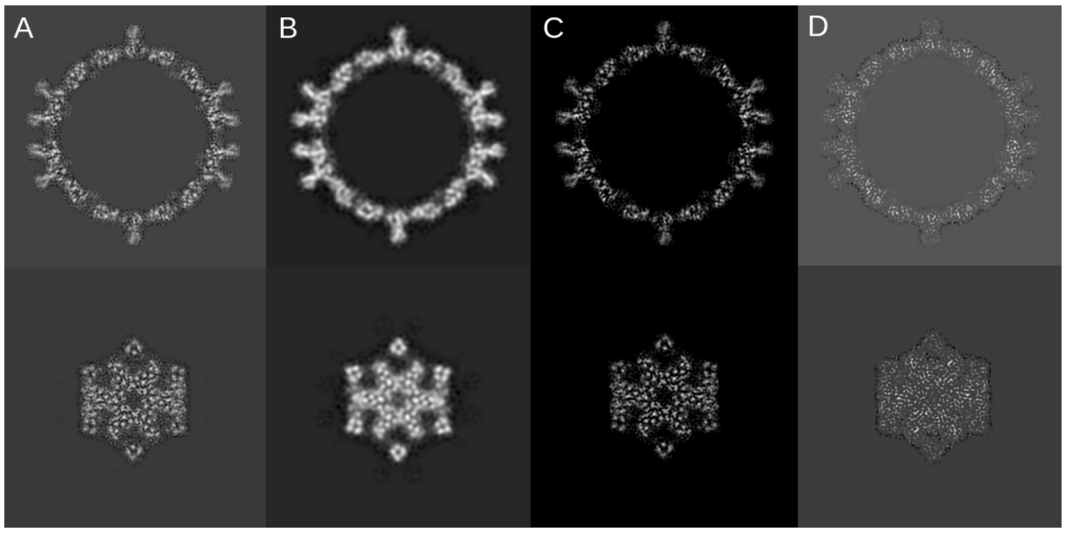


Fig. 4. Two slices from SPA Hepatitis-B viral capsid (left column), StA Hepatitis-B viral capsid (center left column), adjusted StA map to SPA map (center right column) and subtraction between both (right column).

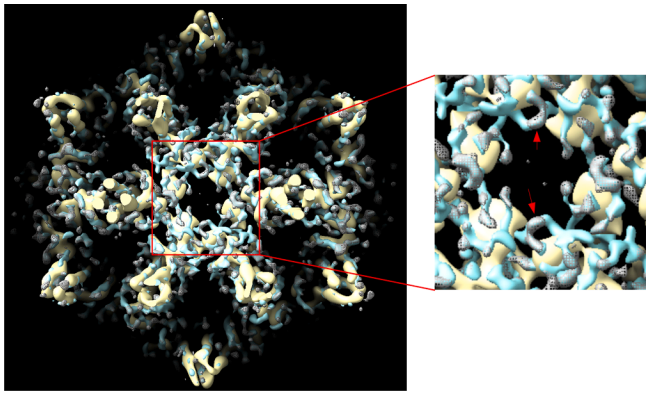


Fig. 5. Bottom view of the bottom slice in Fig. 4 of the Hepatitis-B viral capsid displayed in ChimeraX. StA volume rendered in yellow, SPA volume in light blue and difference volume in grey as a mesh. Red arrows point some remarkable differences between the subtracted volumes.

(the ones that are darker than the background) in the case of DiffMap. Nevertheless, both algorithms seem to mostly agree in the bigger differences (brighter pixels), however the brighter regions appear more blurred in the result of DiffMap.

We show in Fig. 7 the correlation of the Fourier Shell Correlation (FSC) of the difference map in Fig. 6D with the input map in Fig. 6A (FSC 1, in blue) and the input map in Fig. 6B (FSC 2). As can be seen, the FSC 1 rather correlates with the FSC of the difference map as the FSC 1 comes from a model converted into density map (Fig. 6A) and thus, in it there is high frequency information that is lacking in Fig. 6B, and that high frequency information remains in the difference map (Fig. 6D). In the case of correlation with the FSC 2, (that comes from the map in Fig. 6B), ideally the correlation would be near to 0 for all the frequencies. However, in this case the FSC 2 does not correlate that much (the correlation is smaller than 0.5, actually 0.4 is the higher correlation) so thus this indicates that there is no over-subtraction at all.

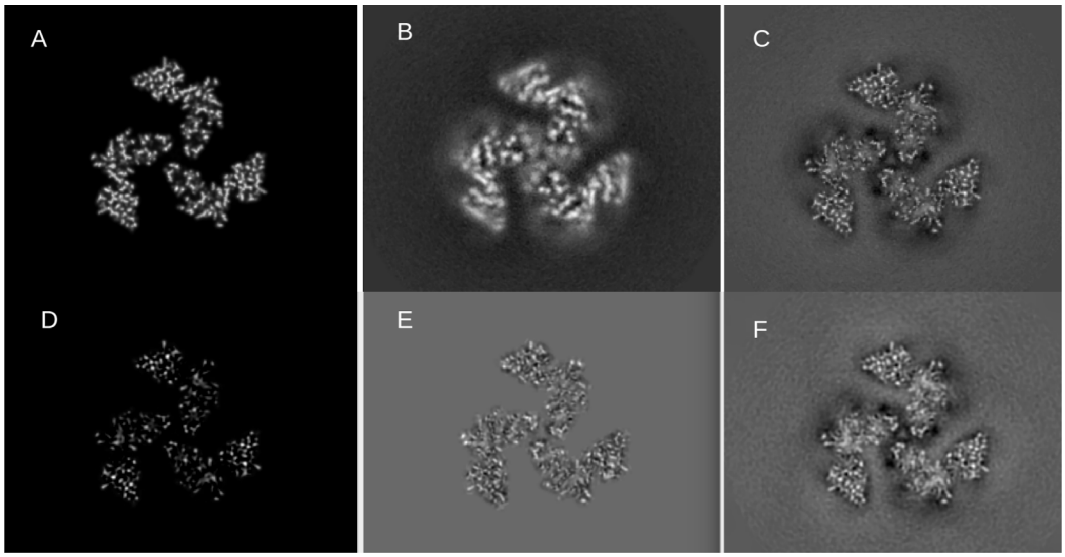


Fig. 6. Central slices of: A) PDB ID6P62 converted into density map. B) Cryo-EM density map of same structure (EMD-20259). C) Subtraction performed by ChimeraX. D) Subtraction performed by proposed algorithm. E) Local subtraction by TemPy:DiffMap. F) Global subtraction by TemPy:DiffMap.

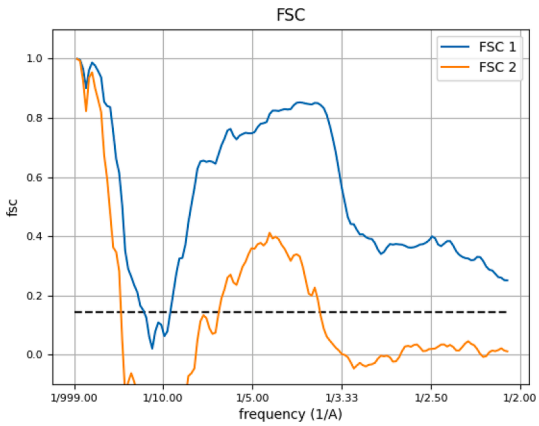


Fig. 7. Fourier Shell Correlation (FSC) of the difference map in Fig. 6D with the input map in Fig. 6A (blue) and the input map in Fig. 6B (orange).

3.2. Map sharpening

As explained in Section 2.2.2, the adjustment algorithm can be used as a post-processing step to sharpen and denoise the reconstructed map. To achieve that, a density map previously generated from an atomic model is used as reference and the experimental map is adjusted to it. To illustrate this implementation of the method, we have used an apoferritin map (EMD-11122) and its associated atomic model (PDB ID6Z9F).

The central slices of the input volume (A) and detail of the region remarked (B), reference (E) and sharpened volumes by different methods are shown in Fig. 8. We compared our result to the ones of other state-of-the-art sharpening methods: Phenix (Terwilliger et al., 2018), LocScale (Jakobi et al., 2017), DeepEMHancer (Sanchez-Garcia et al., 2020), and LocalDeBlur (Ramírez-Aportela et al., 2020).

As can be seen in Fig. 8, in this particular case, LocScale (C) has masked the structure, but the information remains very noisy and it is difficult to observe difference between the original and the sharpened map and Phenix sharpening (D) has not changed significantly the original volume. DeepEMHancer (G) denoised considerably the map, but the density remains blurred and the high frequency details are lost in some regions. This was expected as DeepEMHancer algorithm performs better in volumes whose resolution is worse than the one used here (1.56Å). In the case of LocalDeBlur (H), it has over-sharpened the result. The result of the newly proposed algorithm (F) appears denoised and sharpened and the map looks the most similar to the converted atomic model. The result is currently low pass filtered at the resolution of the original map (1.56 Å) in order to avoid over-sharpening.

In Fig. 9 we show a plot of the energy decay for the original volume, the converted atomic model and the different sharpening methods. Note that the curves are adjusted at medium frequencies (0.07 to 0.12) in order to compare the relative positions of the curves in low and high frequencies. It can be seen that LocalDeBlur (blue), LocScale (red) and Phenix (pink) have a general fall of energy similar to the one of the original map (yellow), but LocalDeBlur and LocScale move away in the high frequencies. However, the decay of energy of the sharpened results must assimilate as much as possible to the one of the converted atomic model (orange). In this case it can be seen that DeepEMHancer (green) and the proposed algorithm (purple) are the nearest. Note that the proposed algorithm actually uses the atomic model in the sharpening process, while LocalDeBlur, DeepEMHancer and Phenix do not. In the case of LocScale, it also uses the atomic model in the sharpening process but it energy decay is closer to the map in intermediate frequencies.

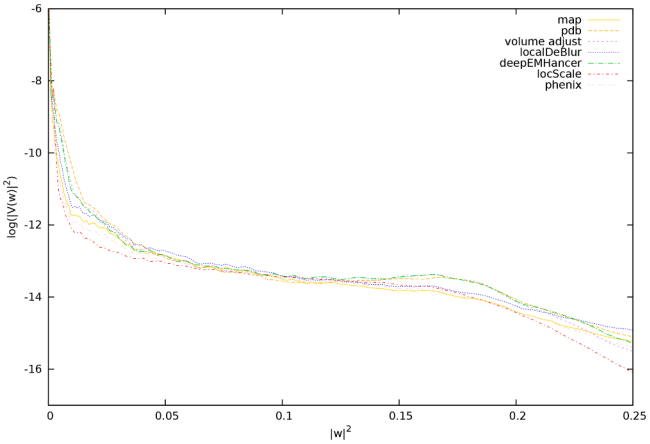


Fig. 9. Energy decay plot of original volume, converted atomic model and different sharpening algorithms. Curves are adjusted at medium frequencies.

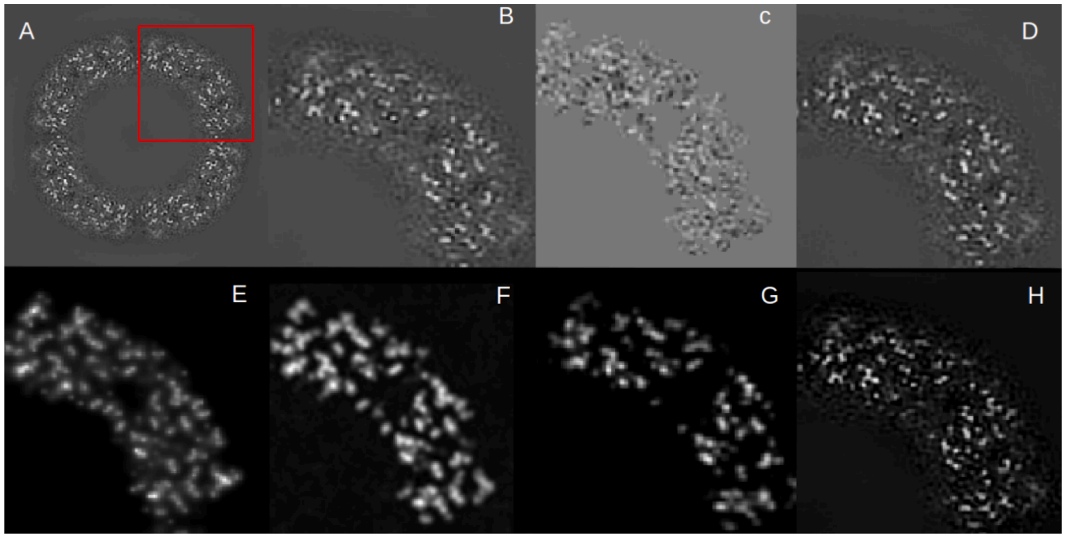


Fig. 8. Central slices of: A) Apoferritin density map EMD-11122. B) Detail of the region remarked in red in A. C) Result of LocScale. D) Result of Phenix sharpening. E) Apoferritin density map derived from atomic model PDB ID 6Z9F. F) Result of proposed algorithm as sharpening method low pass filtered at the input map resolution (1.56 Å). G) Result of DeepEMHancer. H) Result of LocalDeBlur.

3.3. Consensus

We show here two examples of consensus maps. The first example uses several apoferritin experimental maps and it is meant to show the correctness of the method in a setup in which the solution is known. The second example shows its application to several reconstructions of the Sars-CoV-2 Spike.

3.3.1. Apoferritn

As a proof of concept, we have adjusted and fused six different reconstructions of apoferritin from EMDB with different resolutions (entries 0144 with a resolution of 1.64 Å, 6800 with a resolution of 2.90 Å, 6801 with a resolution of 3.20 Å, 3854 with a resolution of 3.15 Å, 3853 with a resolution of 2.50 Å and 4213 with a resolution of 2.14 Å, see Fig. 10). The six maps have been cropped to 200 × 200 × 200, as it was the smallest input box size and their voxel size has been set to 0.81 Å, as it was the smallest input voxel size.

Then, the six volumes were aligned and adjusted having as common reference the first volume (EMD-0144), as it was the one with best resolution (1.65 Å). The first volume was also adjusted to itself, in order to be strictly positive and low pass filtered like the rest. Each map was adjusted to the reference and low pass filtered to its own resolution. Then, we used the six adjusted volumes as input for the consensus. No parameters are required for this algorithm.

The consensus volume is shown in Fig. 11. The color scale represents the degree of difference between the inputs, being dark blue the smallest differences and red the largest ones. As expected, the consensus volume is very similar to the one with best resolution. This fact confirms the correct performance of the consensus algorithm, which is keeping in each region the information of best resolution.

In Fig. 12, the central slice of consensus volume (A and B) and the central slice for the best input volume (C) are displayed together with the central slice of PDBPDB6WXWX6 converted into density map (D) as reference. It can be seen that in general all the signal from the edges of the map is reinforced in consensus result in comparison with the input map of best resolution. Red arrows in the Fig. 12 points to regions with differences between input volume (C) and consensus (B), where the signal in the consensus is reinforced and thus, is more similar to the signal in D. These are small differences but they point out that the algorithm, as designed, is taking the most energetic coefficients of any of the input volumes, and that in these regions, not all the input volumes agree.

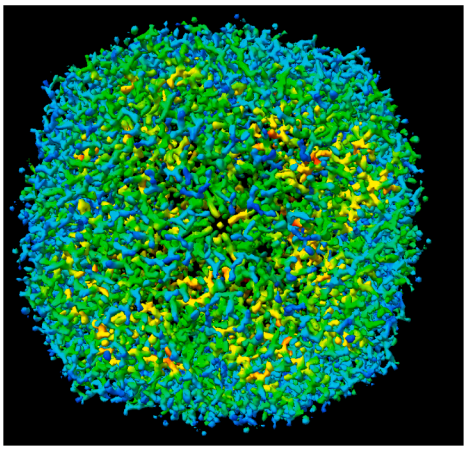


Fig. 11. Isosurface representation of the consensus of the six reconstructions of apoferritin displayed. The colors represent the degree of similarity between the signal of the input maps, being blue the smallest differences between input volumes and red the biggest differences.

3.3.2. SARS-CoV-2 spike

To better illustrate the utility of volume consensus, we have used this algorithm with different reconstructions of the spike of the SARS-CoV-2 obtained from the same dataset with different reconstruction algorithms: CryoSparc (Punjani et al., 2017), Relion (Scheres, 2012) and Xmipp HighRes (Sorzano et al., 2018). Two of the input maps correspond to the additional maps (sharpened maps) of entry EMD-11328, which processing details are described in Melero et al. (2020), while the other three are in-home reconstructions with same softwares and

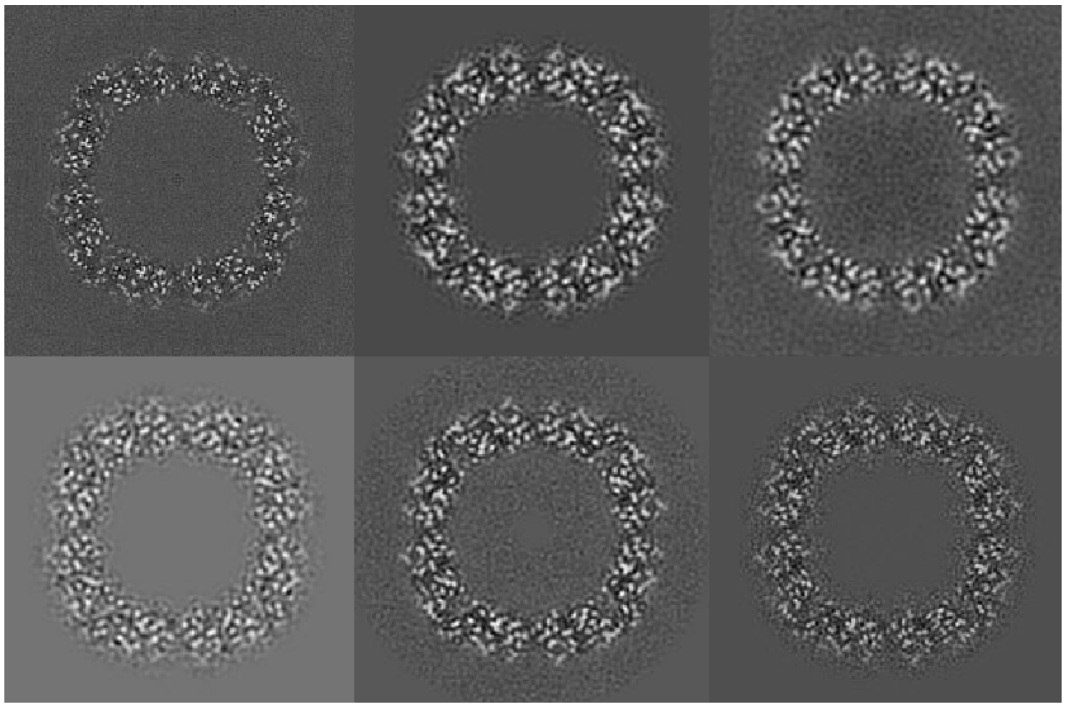


Fig. 10. Central slices of six different reconstructions of apoferritin from EMDB with different resolutions. They have been used as inputs for the consensus algorithm described in this article.

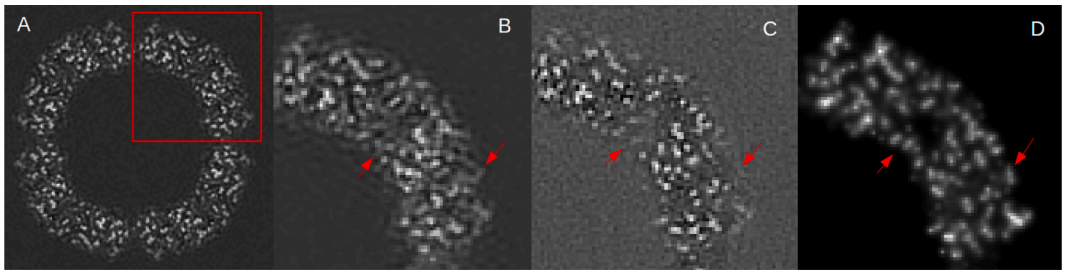


Fig. 12. A) Central slice of consensus result. B) Detail of the squared region in A. C) Same region as in B of the best resolution input apoferritin map (first map of Fig. 10) D) Same region as in B of the atomic model PDB 6WX6 converted into density map. Red arrows points to the regions where significant differences are found between the maps.

similar processing that have not been submitted to any public database. We aligned the five density maps of this structure and computed a tight mask for each one.

We remark here that these five different reconstructions of the spike are all in the same conformation, as this consensus algorithm is not intended for structures in different conformations, as the result will be a mixture of the input conformations which will make up an unreal structure.

We adjusted all of the reconstructions against the first one of the inputs, with default parameters. We arbitrarily took the first one as reference, as the five reconstructions were similar in resolution, but the result does not significantly change depending on the map selected as reference.

Then, we performed the volume consensus with the five adjusted volumes. The result is showed in Fig. 13. The colors in the figure represent the degree of similarity, being dark blue the smallest differences between input volumes and red the biggest ones. As can be appreciated, in the core of the structure there are not significant discrepancies. However, there are green to red regions at the top and bottom of the structure, indicating that these are regions with larger differences between the input reconstructions. We have used here reconstructions of the spike in the "up" conformation, however we know that the flexibility of this spike is high and the "up" conformation is not unique but there are many of them which differs in small details (Melero et al., 2020). Thus, the consensus method in this case is also useful to see in which small parts of the map the reconstructions differ due to details in similar conformations.

To explore those differences, we show the slices that correspond to that part of the structure. In Fig. 14 the same slice for the five input reconstructions and for the consensus are shown. Red arrows points to regions where there are differences between the inputs, but in the consensus appears the best input density in each region, no matter from

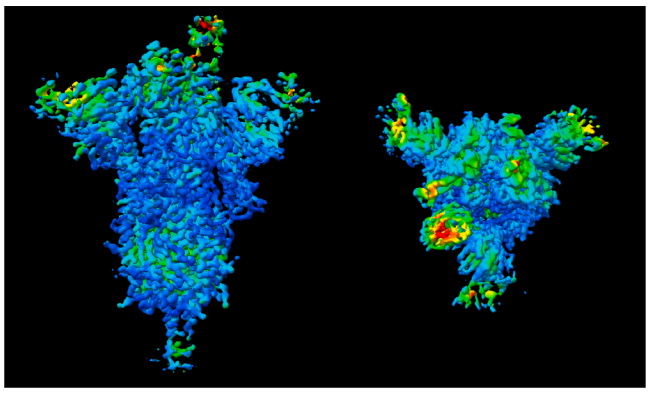


Fig. 13. Result of the consensus volume from five reconstructions of the SARS-CoV-2 Spike, side (left) and top (right) views. The colors represent the degree of similarity of the input map signal, being dark blue the smallest differences between input volumes and red the biggest differences.

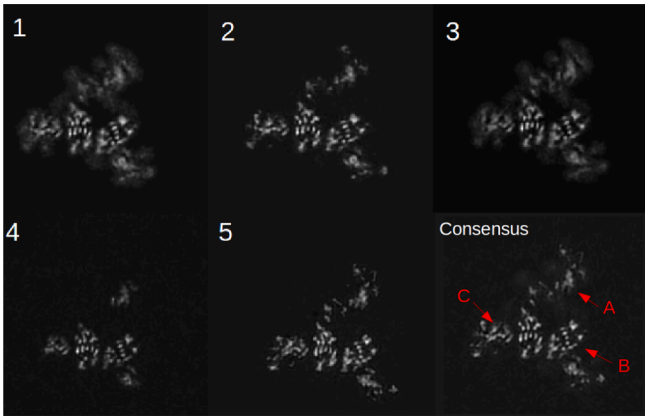


Fig. 14. Same slice of the five input reconstructions of SARS-CoV-2 Spike and the consensus result. Red arrows points to the regions where we can appreciate a fusion of the information provided by the input volumes, obtaining a better density.

what input comes from. In Fig. 15 we show the densities of each of the input volumes and the consensus volume that correspond to the regions pointed by red arrows in Fig. 14. The consensus volume is also displayed as a mesh superimposed to each of the input volumes to see the difference. As can be appreciated, in the three cases the consensus is the volume with more well-defined density in comparison with the rest, as it is build with the best parts of each of its inputs.

4. Conclusions

We have presented in this article a procedure to adjust the numerical values of two density maps. The procedure finds a trade-off between the Fourier amplitudes of one volume and the phases of another. Additionally, it imposes non-negativity, range and locality (mask) constraints to the result. We have also shown three different applications in which this operator is useful (volume subtraction, sharpening and volume consensus).

In the case of subtraction and sharpening, the results show that the proposed algorithm is at the level of the state-of-the-art methods, even improving the results in most cases. In the case of consensus algorithm, we did not find any other method among cryo-EM software packages that performs a similar task. We remark that consensus is not designed to work with structures in different conformations, but to combine different estimations of the same macromolecule on the same conformation. Moreover, in structures with high flexibility, as the case of the SARS-CoV-2 spike presented in Section 3.3.2, the consensus algorithm is not intended to generate as a result and improved map, but is a useful tool to identify the regions of the map where there are more differences among the input maps.

These methods are implemented in the Xmipp package and are user-

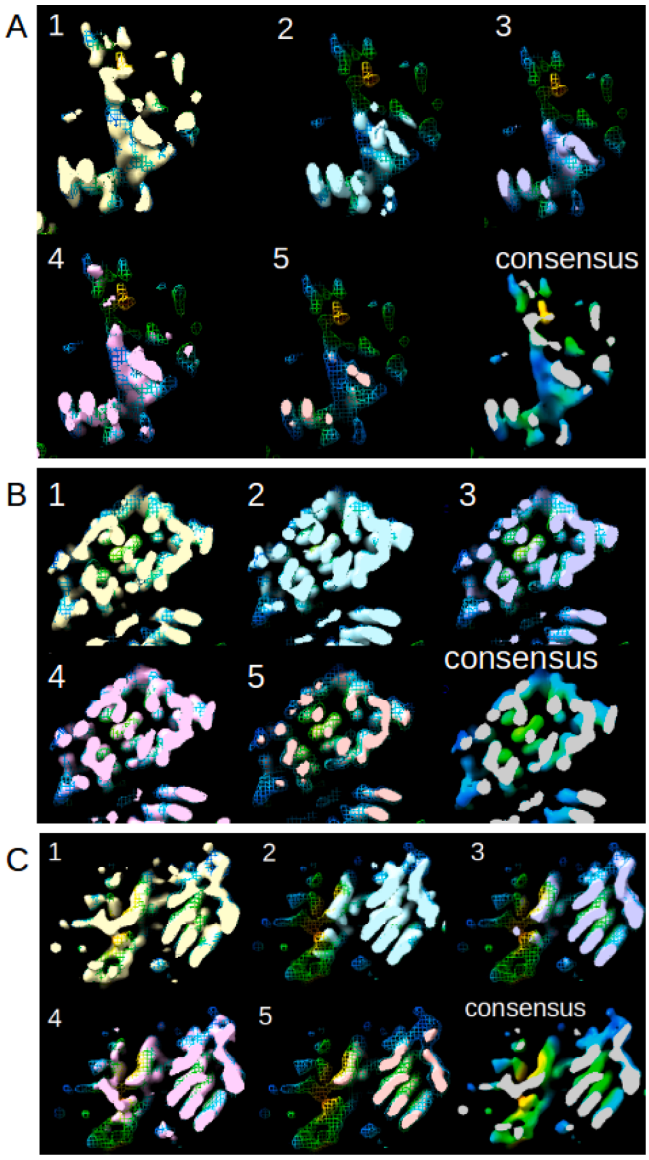


Fig. 15. Detail in 3D of the regions pointed by arrows in Fig. 14. Input volumes are displayed in different colors from 1 to 5. Consensus volume is displayed individually and superimposed to each input volume as a gradient-colored mesh.

friendly accessible through the cryo-EM workflow engine Scipion.

CRediT authorship contribution statement

E. Fernández-Giménez: Methodology, Software, Validation, Formal analysis, Writing - original draft, Visualization. **M. Martínez:** Validation, Resources, Writing - review & editing. **R. Sánchez-García:** Software, Writing - review & editing. **R. Marabini:** Resources, Writing - review & editing. **E. Ramírez-Aportela:** Resources, Writing - review & editing. **P. Conesa:** Software. **J.M. Carazo:** Writing - review & editing, Supervision. **C.O.S. Sorzano:** Conceptualization, Methodology, Formal analysis, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Spanish Ministry of Science and Innovation through Grants: SEV 2017–0712, PID2019-104757RB-I00/ AEI/ 10.13039/501100011033, the "Comunidad Autónoma de Madrid" through Grant: S2017/BMD-3817 European Union (EU) and Horizon 2020 through grants: EOSC Life (Proposal: 824087), HighResCells (ERC – 2018 – SyG, Proposal: 810057) and iNEXT-Discovery (Proposal: 871037). The authors acknowledge the support and the use of resources of Instruct, a Landmark ESFRI project.

References

Bai, X.C., Rajendra Eeson, Yang Guanghui, Shi Yigong, Scheres Sjors H., 2015. Sampling the conformational space of the catalytic subunit of human γ-secretase. Elife, 4.

de la Rosa-Trevín, J.M., Otón, J., Marabini, R., Zaldívar, A., Vargas, J., Carazo, J.M., Sorzano C.O.S., 2013. Xmipp 3.0: an improved software suite for image processing in electron microscopy. Journal of Structural Biology, 184(2), 321–328.

de la Rosa-Trevín, J.M., Quintana, A., Del Cano, L., Zaldívar, A., Foche, I., Gutiérrez, J., Gómez-Blanco, J., Burguet-Castell, J., Cuenca-Alba, J., Abrishami, V., Vargas, J., Otón, J., Sharov, G., Vilas, J.L., Navas, J., Conesa, P., Kazemi, M., Marabini, R., Sorzano, C.O.S., Carazo, J.M., 2016. Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. Journal of Structural Biology 195, 93–99.

Goddard, Thomas D., Huang, Conrad C., Meng, Elaine C., Pettersen, Eric F., Couch, Gregory S., Morris, John H., Ferrin, Thomas E., 2017. Ucsf chimerax: Meeting modern challenges in visualization and analysis. Protein Science 27 (1), 14–25.

Jakobi Arjen J., Wilmanns Matthias, Sachse Carsten, 2017. Model-based local density sharpening of cryo-EM maps. eLife, 6:e27131. ISSN 2050–084X.

Joseph Agnel Praveen, Lagerstedt Ingvar, Jakobi Arjen, Burnley Tom, Patwardhan Ardan, Topf Maya, Winn Martyn, 2020. Comparing cryo-em reconstructions and validating atomic model fit using difference maps. Journal of chemical information and modeling. ISSN 1549-960X.

Madisetti, V.J., Williams, D., 1999. Digital Signal Processing Handbook. CRC Press.

Roberto Melero, Carlos Oscar S. Sorzano, Brent Foster, José-Luis Vilas, Marta Martínez, Roberto Marabini, Erney Ramírez-Aportela, Ruben Sanchez-Garcia, David Herreros, Laura del Caño, Patricia Losana, Yunior C. Fonseca-Reyna, Pablo Conesa, Daniel Wrapp, Pablo Chacon, Jason S. McLellan, Hemant D. Tagare, Jose-Maria Carazo, 2020. Continuous flexibility analysis of SARS-CoV-2 spike prefusion structures. IUCrJ, 7(6):1059–1069. doi:10.1107/S2052252520012725.

Pajares, G., de la Cruz, J.M., 2004. A wavelet-based image fusion tutorial. Patter Recognition 37, 1855–1872.

Marta Pérez-Illana, Marta Martínez, Gabriela N. Condezo, Mercedes Hernando-Pérez, Casandra Mangroo, Martha Brown, Roberto Marabini, Carmen San Martín, Cryo-em structure of enteric adenovirus hadv-f41 highlights structural variations among human adenoviruses. Science Advances, 7(9):eabd9421.

Punjani A., Rubinstein, J.L., Fleet, D.J., Brubaker, M.A., 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nature Methods, 14:290–296. ISSN 1548-7105.

Rafie, K., Lenman, A., Fuchs, J., Rajan, A., Arnberg, N., Carlson, L.-A., 2021. The structure of enteric human adenovirus 41–a leading cause of diarrhea in children. Science Advances, 7(2):eabe0974.

Ramírez-Aportela Erney, Luis Vilas Jose, Glukhova Alisa, Melero Roberto, Conesa Pablo, Martínez Marta, Maluenda David, Mota Javier, Jiménez Amaya, Vargas Javier, Marabini Roberto, Sexton Patrick M., Carazo Jose Maria, Sorzano C.O.S., 2020. Automatic local resolution-based sharpening of cryo-EM maps. Bioinformatics, 36: 765–772.

Ramírez-Aportela Erney, Maluenda David, Fonseca Yunior C., Conesa Pablo, Marabini Roberto, Bernard Heymann J., Maria Carazo Jose, Sorzano Carlos Oscar S., 2021. Fsc-q: A cryoem map-to-atomic model quality validation based on the local fourier shell correlation. Nature Communications, 12(1), 1–7.

Sanchez-Garcia, R., Gomez-Blanco, J., Cuervo, A., Carazo, J.M., Sorzano, C.O.S., Vargas, J., 2020. Deepemhancer: a deep learning solution for cryo-em volume post-processing. bioRxiv. URL url:https://www.biorxiv.org/content/early/2020/08/17/2020.06.12.148296.

Scheres, S.H.W., 2012. Relion: implementation of a bayesian approach to cryo-EM structure determination. Journal of Structural Biology 180, 519–530.

Sorzano, C.O.S., Vargas, J., Otón, J., Abrishami, V., de la Rosa Trevín, J.M., del Riego, S., Fernández-Alderete, A., Martínez-Rey, C., Marabini, R., Carazo, J.M., 2015. Fast and accurate conversion of atomic models into electron density maps. AIMS. Biophysics 2, 8–20.

Sorzano, C.O.S., Vargas, J., de la Rosa-Trevin, J.M., Jimenez, A., Maluenda, D., Melero, R., Martinez, M., Ramirez-Aportela, E., Conesa, P., Vilas, J.L., Marabini, R., Carazo, J.M., 2018. A new algorithm for high-resolution reconstruction of single particles by electron microscopy. Journal of Structural Biology, 204:329–337. ISSN 1095-8657.

Terwilliger Thomas C., Sobolev Oleg V., Afonine Pavel V., Adams Paul D., 2018. Automated map sharpening by maximization of detail and connectivity. Acta crystallographica. Section D, Structural biology, 74:545–559. ISSN 2059-7983.

Terwilliger Thomas C., Sobolev Oleg V., Afonine Pavel V., Adams Paul D., Read Randy J., 2020. Density modification of cryo-em maps. Acta Crystallographica. Section D, Structural Biology, 76:912–925. ISSN 2059-7983.

Vilas, J.L., Gómez-Blanco, J., Conesa, P., Melero, R., de la Rosa Trevín, J.M., Otón, J., Cuenca, J., Marabini, R., Carazo, J.M., Vargas, J., Sorzano, C.O.S., 2018. MonoRes: automatic and unbiased estimation of local resolution for electron microscopy maps. Structure 26, 337–344.

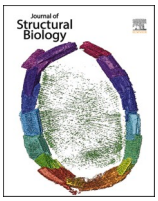
B. A new algorithm for particle weighted subtraction to eliminate signals from unwanted components in Single Particle Analysis.

Journal of Structural Biology 215 (2023) 108024

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi



Research Article

A new algorithm for particle weighted subtraction to decrease signals from unwanted components in single particle analysis

E. Fernández-Giménez [a,b], M.M. Martínez [a], R. Marabini [a,b], D. Strelak [c], R. Sánchez-García [d,e], J.M. Carazo [a], C.O.S. Sorzano [a,*]

[a] Centro Nac. Biotecnología (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain
[b] Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[c] Institute of Computer Science, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic
[d] Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, United Kingdom
[e] Astex Pharmaceuticals, 436 Cambridge Science Park, Cambridge CB4 0QA, UK

ARTICLE INFO

Keywords:
Projection subtraction
Nanodisc
Ligand
SPA
Cryo-EM

ABSTRACT

Single particle analysis (SPA) in cryo-electron microscopy (cryo-EM) is highly used to obtain the near-atomic structure of biological macromolecules. The current methods allow users to produce high-resolution maps from many samples. However, there are still challenging cases that require extra processing to obtain high resolution. This is the case when the macromolecule of the sample is composed of different components and we want to focus just on one of them. For example, if the macromolecule is composed of several flexible subunits and we are interested in a specific one, if it is embedded in a viral capsid environment, or if it has additional components to stabilize it, such as nanodiscs. The signal from these components, which in principle we are not interested in, can be removed from the particles using a projection subtraction method. Currently, there are two projection subtraction methods used in practice and both have some limitations. In fact, after evaluating their results, we consider that the problem is still open to new solutions, as they do not fully remove the signal of the components that are not of interest. Our aim is to develop a new and more precise projection subtraction method, improving the performance of state-of-the-art methods. We tested our algorithm with data from public databases and an in–house data set. In this work, we show that the performance of our algorithm improves the results obtained by others, including the localization of small ligands, such as drugs, whose binding location is unknown *a priori*.

1. Introduction

Single particle analysis (SPA) in cryo-electron microscopy (cryo-EM) has emerged as a reliable method for elucidating the atomic structure of macromolecules and biological complexes, owing to its remarkable ability to produce high-resolution electronic density maps (below 3 Å) (Neumann et al., 2018).

Nevertheless, certain challenging samples require additional processing to achieve high resolution. For example, macromolecules that consist of flexible multiple subunits often face alignment difficulties, with the larger subunit dominating the alignment and leading to poorer resolution for smaller subunits. In other cases, macromolecules may be surrounded by additional proteins (such as a specific protein within a virus capsid) or other molecules (such as a nanodisc). The presence of these surrounding molecules or domains complicates image processing, as they typically impact classifications and alignments, resulting in the molecule of interest being under-resolved.

To tackle this problem, the logical approximation is to remove the signal of the components we are not interested in from the processing. We can perform it by subtracting directly in the volume with the methods developed for this purpose in ChimeraX (Pettersen et al., 2021) or Xmipp (Fernandez-Gimenez et al., 2021). However, while volume subtraction is computationally more efficient, it only cleans the view of the final structure of interest, but it has no impact on its resolution.

Alternatively, it is possible to mask or remove the unwanted signal at the particle level. The goal is to obtain a new set of 2D projections containing solely the signal from the 3D components of interest, which can then be used for iteratively classifying, reconstructing, and refining

* Corresponding author.
E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

https://doi.org/10.1016/j.jsb.2023.108024
Received 4 July 2023; Received in revised form 22 August 2023; Accepted 4 September 2023
Available online 11 September 2023

without the unwanted signal, which may lead to an improvement of the final resolution. Once again, there are several methods available to achieve this task. For instance, Relion (Kimanius et al., 2021) offers focus refinement capabilities through masking and projection subtraction, and CryoSPARC (Punjani et al., 2017) provides a program for projection subtraction.

However, this task is not entirely solved in practice, as many undesired signals remain in the subtracted particles. In this article, we present a new method for projection subtraction, developed within the Xmipp (Strelak et al., 2021; de la Rosa-Trevin et al., 2013; Sorzano et al., 2004) software package, that improves the results of the state-of-the-art methods, as we show in the Results section. This new algorithm is also available in the Scipion (Jimenez-Moreno et al., 2021; de la Rosa-Trevin et al., 2016) framework for cryo-EM image processing.

2. Methods

The basic idea behind a projection subtraction algorithm is to take every input particle, compute the projection of the reference volume corresponding to the pose of the input particle, and subtract it from the original image, normally operating in real space. Usually, we do not want to subtract the projection of the whole particle, but only of a specific region. This region will be determined by a volume mask m, either defining the region to keep (as required by Relion) or the region to subtract (as required by CryoSPARC). See an example of the masks in Fig. 1. In the case of our algorithm, the user can choose if the input mask defines the region to keep or subtract. The complete projection subtraction workflow of our algorithm is reflected in Fig. 2:

However, the real complexity of the algorithm resides in preparing the projection before subtracting, as a projection of a reference volume and a real particle have different characteristics, and they usually have different ranges of values. Thus, instead of subtracting the projection, we will modify it previously to have an adjusted version suitable for subtraction:

$$i_s(\mathbf{s}) = i(\mathbf{s}) - p'(\mathbf{s}) \tag{1}$$

being $(\mathbf{s})$ the spatial 2D coordinate, i_s the subtracted particle, i the input particle, and p' the adjusted projection. Let $p(\mathbf{s})$ be the projection of the map to be subtracted along the same direction and the same in-plane shift as the image i.

In our algorithm, we perform this adjustment in Fourier space, previous to the subtraction in real space. Moreover, we will adjust each projection to its particle individually. For calculating the adjusted projection, let us introduce some notation. Let $I(\boldsymbol{\omega})$ and $P(\boldsymbol{\omega})$ be the Fourier transform of the i and p images, respectively. Note that these Fourier transforms are complex-valued vectors. Let us consider a set of frequencies on which we will perform the subtraction, Ω. We define the $\mathrm{vec}_\Omega\{\cdot\}$ operator that acts on the Fourier transform of an image as

$$\mathrm{vec}_\Omega\left\{ I \right\} = \begin{pmatrix} \mathrm{Re}\{I(\boldsymbol{\omega}_0)\} \\ \mathrm{Im}\{I(\boldsymbol{\omega}_0)\} \\ \mathrm{Re}\{I(\boldsymbol{\omega}_1)\} \\ \mathrm{Im}\{I(\boldsymbol{\omega}_1)\} \\ \cdots \end{pmatrix} \tag{2}$$

where Re and Im extract the real and imaginary parts of a complex number, and $\boldsymbol{\omega}_i$ goes over all frequencies in the set Ω. This vector is called $\mathbf{I}$, and its k-th component is I_k. We also define the corresponding frequency vector as

$$\mathbf{w} = \begin{pmatrix} |\boldsymbol{\omega}_0| \\ |\boldsymbol{\omega}_0| \\ |\boldsymbol{\omega}_1| \\ |\boldsymbol{\omega}_1| \\ \cdots \end{pmatrix} \tag{3}$$

and we refer to its k-th component as w_k. We also define a diagonal matrix W whose main diagonal is this vector.

Let us define H as a diagonal matrix representing the microscope Contrast Transfer Function (CTF). We now consider two scaling transformations, T_0 and T_1 that will minimize the Euclidean distance between $\mathbf{I}$ and $H\mathbf{P}$

$$\epsilon^2 = \left\| \mathbf{I} - T(H\mathbf{P}, \mathbf{w}) \right\|^2 \tag{4}$$

where T is any of the transformations (T_0 or T_1). The two transformations are given by a linear model of order zero and order one respectively

$$\begin{aligned} T_0(H\mathbf{P}, \mathbf{w}) &= \beta_{00} H\mathbf{P} \\ T_1(H\mathbf{P}, \mathbf{w}) &= (\beta_{01} + \beta_{11} W) H\mathbf{P} \end{aligned} \tag{5}$$

where β_{ij} are constants we must find to minimize the above Euclidean distance. Note that the first transformation, T_0, is a grayscale adjustment, and T_1 is a grayscale adjustment and a projection sharpening or dampening to adjust the map projection to subtract from the experimental image. We compute both models even though T_0 is contained in T_1, because if T_0 fits well enough it is not worth adding a new variable in order to avoid overfitting, as happens generally in linear regression.

The β coefficients are found by standard linear regression:

$$\begin{aligned} \beta_{00} &= \left(\left\| H\mathbf{P} \right\|^2 \right)^{-1} \langle H\mathbf{P}, \mathbf{I} \rangle \\ \begin{pmatrix} \beta_{01} \\ \beta_{11} \end{pmatrix} &= \begin{pmatrix} \|H\mathbf{P}\|^2 & \|H\mathbf{P}\|_W^2 \\ \|H\mathbf{P}\|_W^2 & \|H\mathbf{P}\|_{W^2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \langle H\mathbf{P}, \mathbf{I} \rangle \\ \langle H\mathbf{P}, \mathbf{I} \rangle_W \end{pmatrix} \end{aligned} \tag{6}$$

In the expressions above we have made use of the weighted inner product with definition $\langle \mathbf{x}, \mathbf{y} \rangle_A = \mathbf{x}^T A \mathbf{y}$, and its associated norm $\|\mathbf{x}\|_A^2 = \langle \mathbf{x}, \mathbf{x} \rangle_A$.

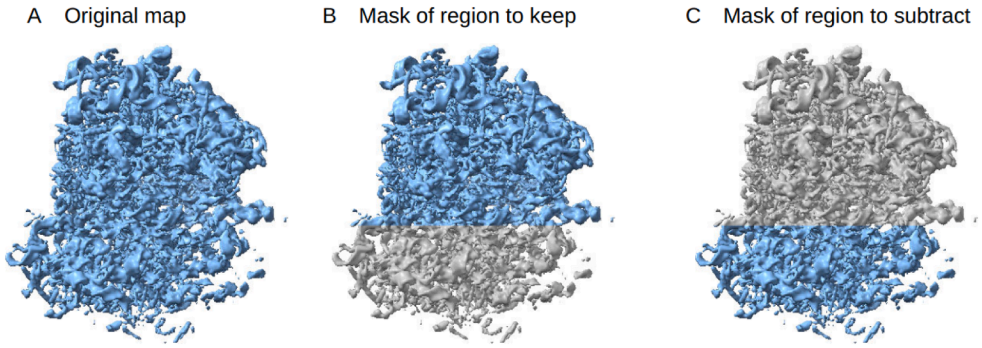


Fig. 1. Example of masks defining the region to keep or the region to subtract. The original map (A) shows a complete ribosome, from which we want to subtract the large subunit in order to keep the small one, thus, the mask defining the part to keep (required by Relion) is colored in grey in (B), and the mask defining the part to subtract (required by CryoSPARC) is colored in grey in (C). Note that the mask to keep is complementary to the mask to subtract.

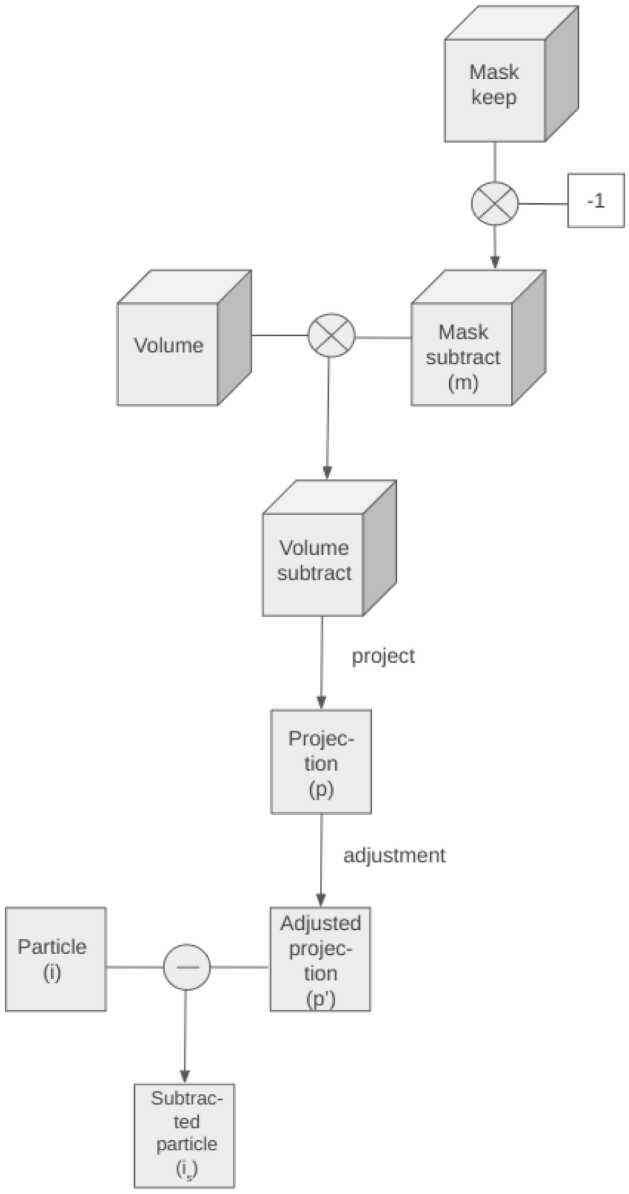


Fig. 2. Subtraction process schema. The input volume is masked with a mask that defines the region to subtract (m). However, the user can input a mask of the region to keep and its inverse will be automatically computed in order to obtain m. The subtraction volume is then projected generating p, which will be adjusted (p') in order to be subtracted to the input particle i, obtaining as a result the subtracted particle i_s.

For any experimental image $\mathbf{I}$, we choose the transformation that maximizes the determination coefficient

$$T\left(\mathbf{P},\mathbf{w}\right) = \begin{cases} T_0(\mathbf{P},\mathbf{w}) & R^2_{0,adj} > R^2_{1,adj} \\ T_1(\mathbf{P},\mathbf{w}) & R^2_{0,adj} < R^2_{1,adj} \end{cases} \tag{7}$$

where the adjusted coefficient of determination (R^2) of the i-th model is calculated as

$$\begin{aligned} R^2_{i,adj} &= 1 - \left(1 - R^2_i\right)\frac{N-1}{N-k-1} \\ R^2_i &= 1 - \frac{\|\mathbf{I} - T_i(\mathbf{P},\mathbf{w})\|^2}{\|\mathbf{I} - \bar{I}\|^2} \end{aligned} \tag{8}$$

where N is the dimension of $\mathbf{I}$, k is the degree of the polynomial in the T_i transformation, and $\bar{I}$ the average value of the vector $\mathbf{I}$.

The adjusted determination coefficient ($R^2_{i,adj}$) is reported for each particle and it is also used to rank them according to their quality, so the user can discard the worst particles.

The subtracted image is finally

$$p' = FT^{-1}\left\{T\left(HP\right)\right\}m_c \tag{9}$$

being FT^{-1} the inverse Fourier Transform and m_c a circular mask in real space whose diameter equals the particle box size. This mask is applied in order to avoid edge and corner artifacts due to the adjustment process.

Note that T_0 is similar to Relion adjustment, as Relion also multiples each particle by a constant, which is referred to as a"scale factor" in their metadata. However, Relion estimates that constant during the refinement step, which then has the same value for all the particles in the same micrograph. In our case, we estimate a different constant for each particle, which may increase our precision. We cannot evaluate the similarities or differences of our algorithm in comparison to the one of CryoSPARC as their subtraction algorithm is not published and its code is not open source. Another important difference is that we can manage masks that define the part to keep or the part to subtract, which is a useful feature depending on the application of subtraction and/or the difficulty of creating the mask. However, Relion only manages input masks of the region to keep, and CryoSPARC only manages masks of the region to subtract, which is a drawback depending on the application of the subtraction and the availability of masks, as will show in the Results section.

3. Results and discussion

To validate our algorithm, we have compared its performance with the ones obtained with the state-of-the-art projection subtraction methods: Relion and CryoSPARC. We have chosen three different scenarios: 1) focused refinement, 2) subtraction of unwanted signals, and 3) ligand discovery. As a general result of all the experiments, it has turned out that the majority of the projections have been modified just by a grayscale adjustment (T_0).

3.1. Focused refinement

3.1.1. Ribosomal small subunit

We have chosen the data set from the EMPIAR entry 10028, the *Plasmodium falciparum* 80S ribosome, with the aim of improving the resolution of the small subunit (whose molecular mass is about 1.2MDa).

As the data set already contains the extracted particles, a first reconstruction and refinement of the ribosome has been performed using CryoSPARC non-uniform refinement (Punjani et al., 2017), Fig. 3 (A). As can be seen, the quality of the map in the region of the small subunit (bottom) shows more unconnected densities than in the region of the big subunit (top). This was expected due to the fact that alignment was driven by the big subunit, as it has a larger signal than the small subunit.

In Fig. 3, the results of focal refinement by Relion (B) and subtract projections by CryoSPARC (C), Xmipp (E) and Relion (F) are shown. Fig. 3 (D) shows the mask for the region that we want to keep in subtractions and focus on focal refinement (the small subunit). As can be appreciated qualitatively in regions highlighted by red squares, Relion focal refinement and CryoSPARC are able to slightly improve the original reconstruction. However, Xmipp and Relion subtraction are able to improve it more. Additionally, CryoSPARC does not remove completely the signal from the large subunit. In this case, the results of Relion and Xmipp are comparable.

3.1.2. Crowded/Viral environment: Monomer B of hexon 1 of human Adenovirus

To compare Xmipp and Relion projection subtraction results in more

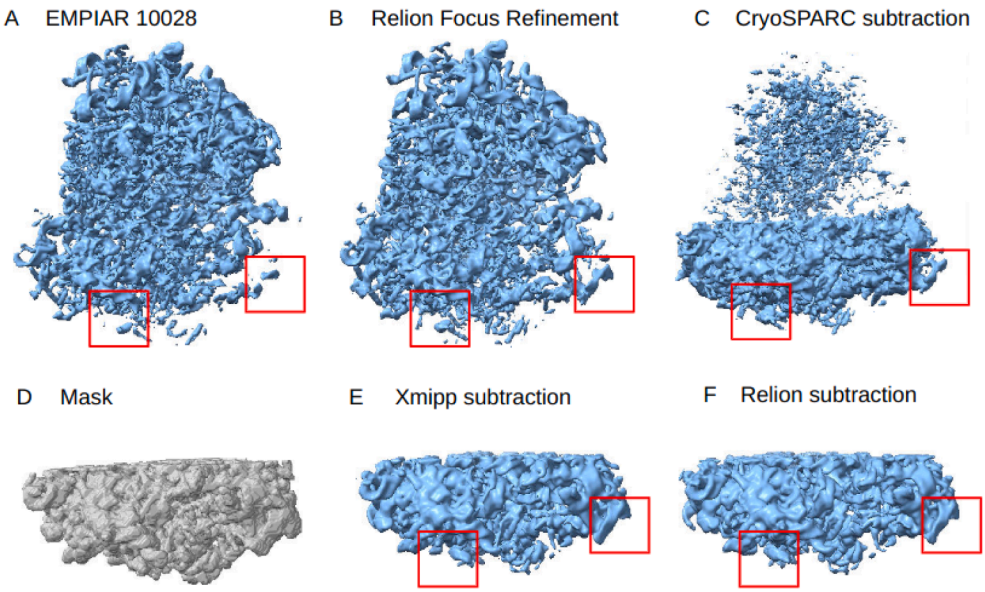


Fig. 3. (A) Refined map from particles in EMPIAR 10028. (B) Focal refinement of small subunit by using Relion focus refinement program. (C) Refined map from subtracted particles by CryoSPARC. (D) Mask of small subunit used in subtractions (region to keep) and focal refinement. (E) Refined map from subtracted particles by Xmipp. (F) Refined map from subtracted particles by Relion. Parts with remarkable differences among the maps are squared in red.

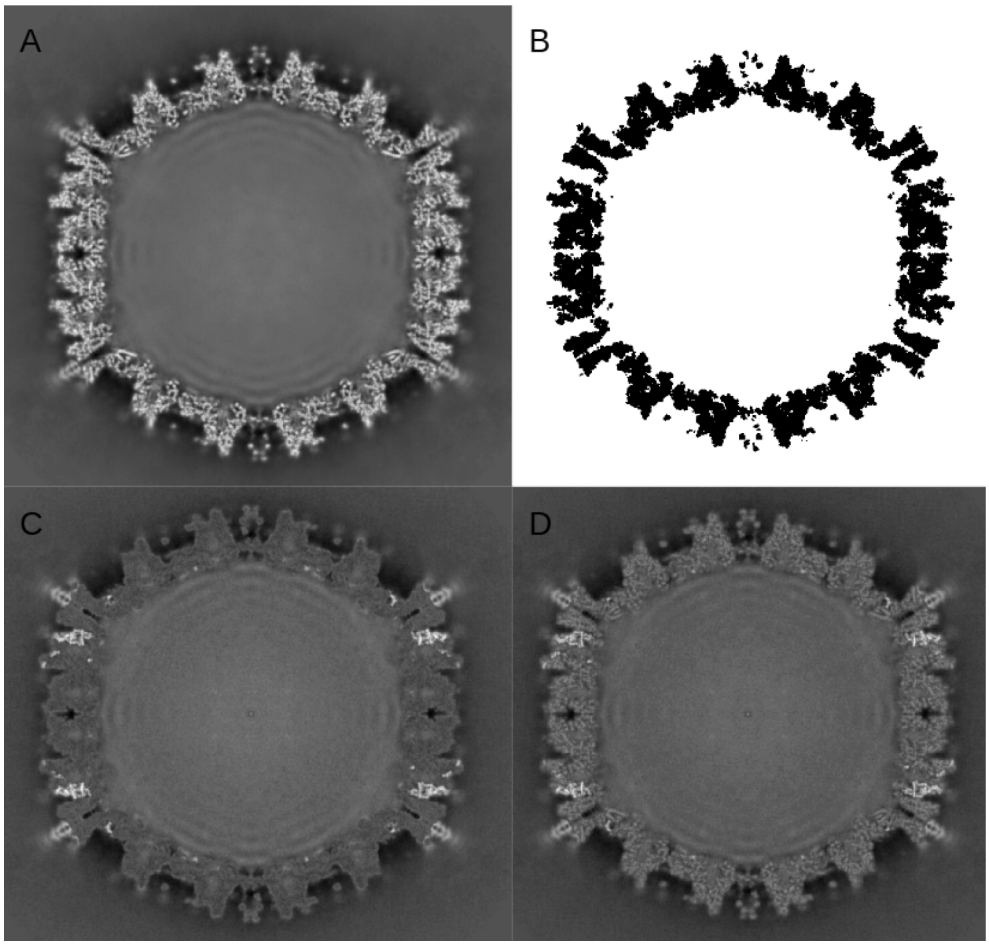


Fig. 4. Human Adenovirus (central slice) (A) Reconstruction without subtraction (B) Mask of the capsid (in black is the region to subtract) (C) Map reconstruction of subtracted particles by Xmipp. (D) Map reconstruction of subtracted particles by Relion.

detail, we have chosen a smaller target in a crowded environment: the monomer B of the hexon 1 of the human Adenovirus (molecular mass 108KDa) from an in–house data set of the Adenovirus capsid. Central slice of the complete Adenovirus previous to subtraction is shown in Fig. 4 (A). In this case, we wanted to subtract the whole capsid except for the monomer B of hexon 1, therefore a mask of the region that we wanted to keep (thus, the capsid is in black) is built and its central slice is shown in Fig. 4 (B). Central slices of the volumes reconstructed after the subtractions with Xmipp and Relion are shown in Figs. (C) and (D) respectively.

A map of monomer B is shown in Fig. 5 (A). It has been obtained by converting PDB entry 6b1t (Dai et al., 2017) into a density map at 3Åresolution. A reconstruction of monomer B with Xmipp subtracted particles is shown in Fig. 5 (B), and the result of refining it is shown in (D). Analogously, the reconstruction with Relion subtracted particles is shown in Fig. 5 (C), and the result of its refinement is in (E). Both refinements have been done with Relion auto-refine (Kimanius et al., 2021) with the same parameters. As can be seen, on the right side of (C) and (E), there are considerable amounts of signals that do not correspond to monomer B, but they have not been removed by Relion subtraction. We can appreciate the undesired remaining signal also in the central slices, in the second row of Fig. 5. The region pointed out with arrows from (B) to (E) correspond to the adjacent monomer B, as monomer B is arranged in a pentameric form around the penton. Thus, it remains in all the cases because it was included in the mask used to define the region to keep (monomer B) in the projections subtraction. The result produced by CryoSPARC subtraction is shown in the Supplementary material Fig. S-1. We consider that it is substantially worse than the previously presented results, as it leaves more signal on the right side and loses part of the signal in the monomer region.

3.2. Subtraction of unwanted signals

As stated in the introduction, having nanodiscs in the sample is very useful for sample preparation of membrane proteins. Still, it is inconvenient when doing image processing as they may drive image alignment. Thus, subtraction of the signal generated by the nanodisc in the particles can improve the final result. To check our ability to handle this case we have used the particles from EMPIAR entry 10005, which sample is a capsaicin receptor that has been embedded in nanodiscs during the sample preparation process. We have generated a mask of the region to keep from related EMDB entry 5778, which is the capsaicin receptor without the nanodisc.

In Fig. 6 (top), we compare two different slices of the map after particle refinement as it is in the entry of EMPIAR 10005 with equivalent two slices of the map obtained after refining the same data set (with the same refinement parameters) once Xmipp has subtracted them.

We can appreciate that the nanodisc is present in the first two cases, while in the subtraction it has been removed (see red arrows in the figure). Even though the reported resolution by FSC does not improve, the local resolution improves as shown in Fig. 6 (bottom), especially in the part where the nanodisc was (bottom of the structure), as the signal from the structure is now stronger in that part. There are also important improvements at the top of the structure in the Figure, as it seems the alignment without subtraction was perturbed by the nanodisc than by the structure itself.

The result obtained by Relion subtraction (see Fig. S-2 in Supplementary materials) with the same particles and the same refinement parameters produces a structure highly degraded. For CryoSPARC subtraction, a mask of the region to subtract is needed, which in this case is the nanodisc. It is difficult to obtain a thigh mask of the nanodisc as there is no direct way to get it, however, we compute an approximate mask of the nanodisc by subtracting the volume in EMD-5778 from the reconstructed volume from the particles in EMPIAR 10005. Even though the mask represents approximately the nanodisc, the result of CryoSPARC subtraction also produces a structure highly degraded (it is shown in Fig. S-3 of Supplementary materials).

3.3. Ligand discovery

3.3.1. Simulated data

To further evaluate the precision of our subtraction method, we decided to test its performance in detecting a small molecule in the context of a much larger one. In this case, we are interested in removing the signal of the small subunit of the ribosome, keeping just a ligand that is bound to it.

In this experiment, we have used the same data set as in Section 3.1.1, as the small subunit of the 80S ribosome of the data set has a drug bounded (emetine). We have used the related PDB entry 3j7a to generate a map at a 2Åresolution without the emetine ligand by removing it from the PDB in ChimeraX and converting it to a density map. To check the performance of the subtraction, we have to avoid other sources of errors, such as alignment errors. Thus, we have not used the original particles of the data set, but we have created a set of projections from the volume in

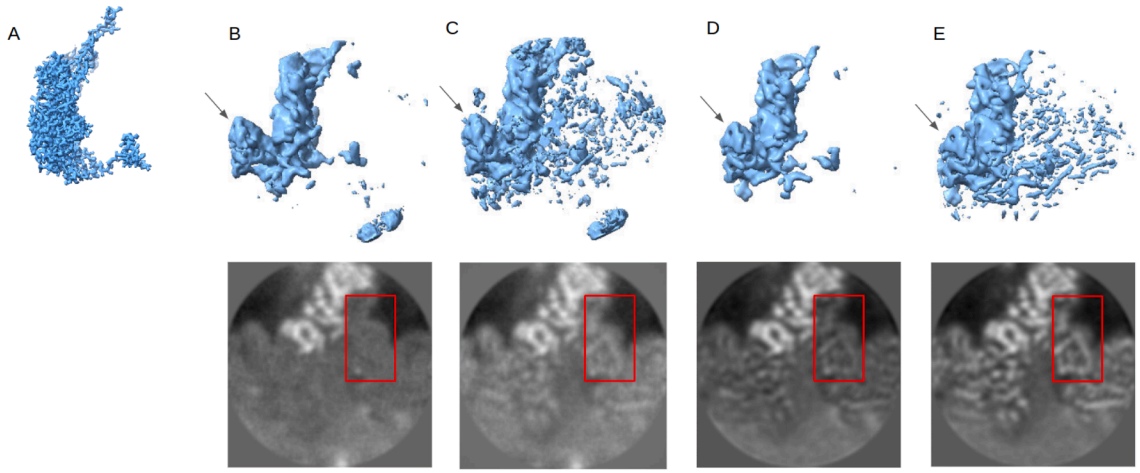


Fig. 5. Top: Monomer B of hexon 1 of the human Adenovirus. Below: a central slice of each map. (A) Converted 3Åmap from PDB entry 6b1t (Dai et al., 2017). (B) Map reconstruction of subtracted particles by Xmipp. (C) Map reconstruction of subtracted particles by Relion. (D) Map refinement of subtracted particles by Xmipp, the reported resolution by FSC is 4.0Å. (E) Map refinement of subtracted particles by Relion, the reported resolution by FSC is 4.1Å. An area showing different signal levels is marked by a red rectangle (ideally, they should have been subtracted). Arrows point to a zone that does not belong to monomer B but was included in the mask used for the subtractio.n in both cases.

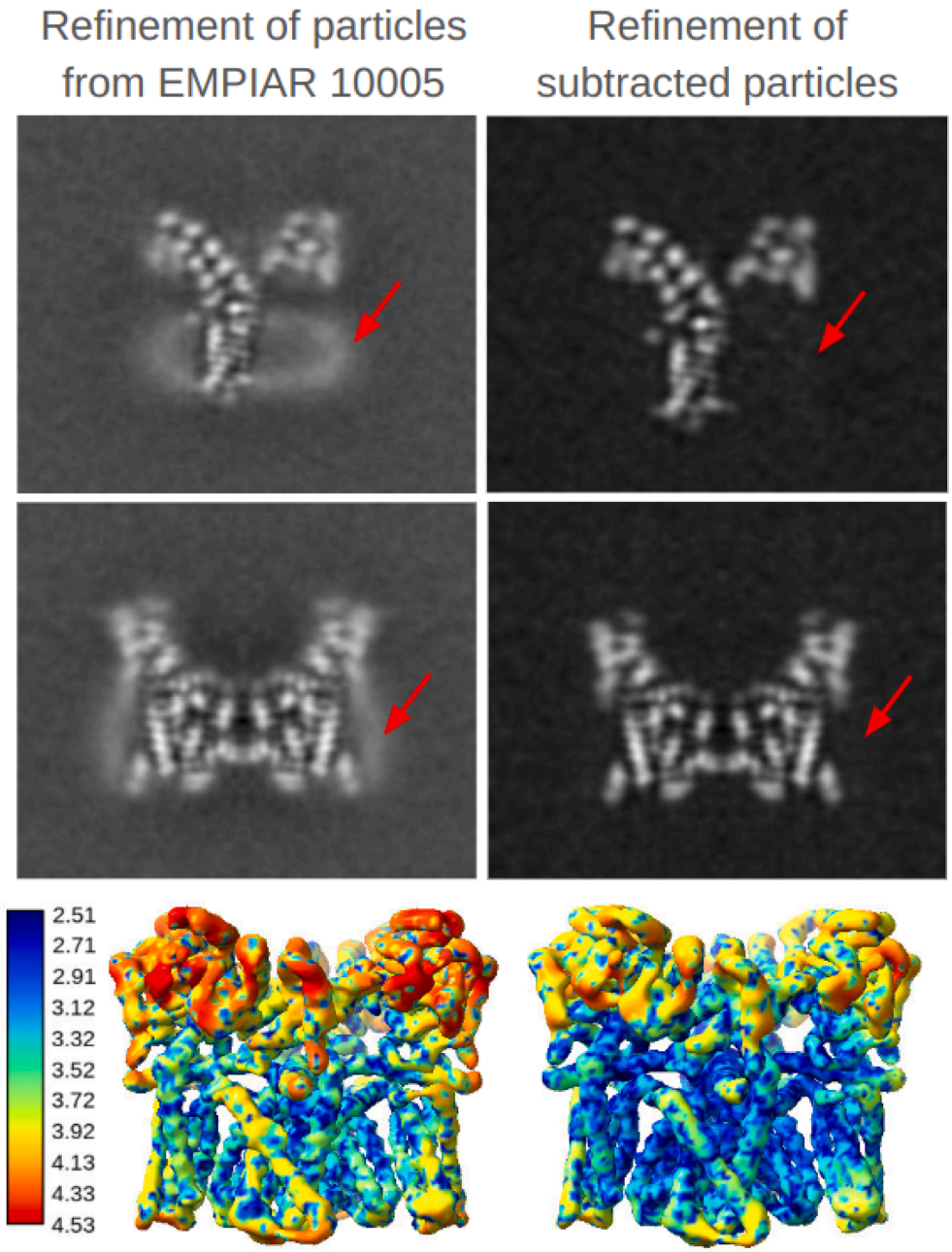


Fig. 6. (Top) The image displays two different slices from the volumes obtained through the refinement of particles in EMPIAR 10005. The left side represents the particles before undergoing subtraction by Xmipp, while the right side shows the particles after subtraction. The same refinement parameters have been applied. The red arrows highlight the signal produced by the nanodisc, which has been eliminated in the subtracted case. (Bottom) Local resolution measure with MonoRes (Vilas et al., 2018) of the refined volumes for (left) particles in EMPIAR 10005 and (right) the same particles once Xmipp has .subtracted them.

EMDB-related entry 2660 to use them as input particles. We have added to these projections a simulated CTF and Gaussian noise with a standard deviation of 50 (see Fig. S-4 in Supplementary material).

The result of reconstructing the subtracted particles with Xmipp is shown in ChimeraX in Fig. 7. The map has been filtered with a Gaussian filter with a standard deviation of one to see the density corresponding to emetine better. We have performed the same experiment with CryoSPARC subtraction, and it has not been able to remove completely the density of the small subunit of the ribosome (see the result in Supplementary material Fig. S-5). We cannot perform this experiment with Relion as it needs as input a mask of the region to keep instead of the region to subtract, requiring knowing ahead the location of the ligand.

3.3.2. *Experimental data*

In this section, we have used signal subtraction for ligand discovery, as in Section 3.3.1, but in this case, with experimental data. The data consists of the human enzyme pyruvate kinase M2 with two pairs of ligands, a sugar (1,6-di-O-phosphono-beta-D-fructofuranose, which is the one we are interested in) and an amino-acid (threonine). We have used our subtraction approach, starting from the movies available in EMPIAR 10647. As a reference volume, we have used the related PDB entry 6th, removing the ligands in ChimeraX and converting the resulting structure (only the enzyme) into a density map at 2Åresolution by using Xmipp (Sorzano et al., 2015). Then, we created a binary tight mask for this map to be used as the mask of the region to subtract because the aim is to get just the signal from the ligands. The result is shown in Fig. 8 (volume displayed in ChimeraX with PDB as reference) and a slice of the volume in Fig. 9 (A). We have called this section "ligand discovery" as this method can be used to find a ligand that is bound to our sample, but still, we do not know where it is. Thus, as in Section 3.3.1, we cannot perform this experiment with Relion because it needs as input the mask of the region to keep. CryoSPARC subtraction cannot recover the signal from the ligand in this case (see Fig. S-6 in Supplementary Materials).

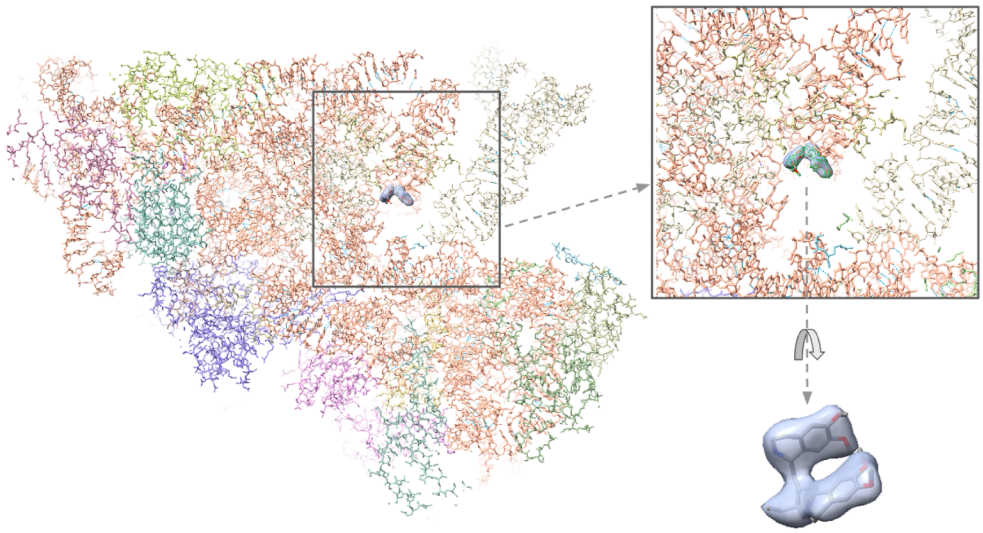


Fig. 7. Reconstructed map of subtracted particles with Xmipp (gray) over PDB PDB3j7a (small subunit of 80S ribosome bound to drug emetine). A zoomed region and detail of the region with density, where emetine fits (rotated to see the fitting better). A Gaussian filter with a standard deviation of one has been used to improve the visualization of the density of the emetine.

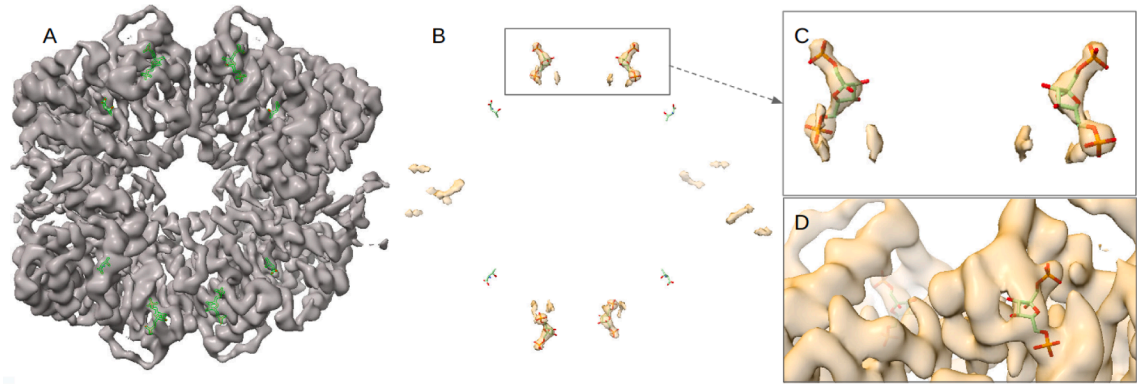


Fig. 8. (A) reconstruction of particles obtained by processing EMPIAR 10647 and (B) reconstruction of same particles subtracted by Xmipp to keep the ligand (1,6-di-O-phosphono-beta-D-fructofuranose), which PDB (6tth) is fitted and (C) zoom of the region of interest. (D) the same region from the original map (before subtraction). The subtraction map does not have any filter or further post-processing.

While volume subtraction is the preferred option because is more computationally efficient, particle subtraction is required if further image processing (such as 3D classification or refinement), is going to be needed. In Fig. 9 we compare the result of Xmipp projection subtraction (A) which implies an adjustment previous to subtraction, Xmipp volume subtraction (Fernandez-Gimenez et al., 2021)(C) which uses another kind of volume adjustment and remove negative values, (B) volume subtraction without any adjustment and, (D), volume subtraction adjusting just mean and standard deviation (D). In (E) we show the result of multiplying (A) and (C) in order to make a consensus of both results and in (F) we show the differences between (A) and (C). As can be seen in (E), the signal coming from the ligand (pointed by red arrows) gets reinforced and in (F) it disappears, showing the agreement between the results.

4. Conclusions

Complementing the now classical pipelines for cryo-EM SPA, special image processing methods are needed to achieve high-resolution information from challenging samples. This is the case of macro-molecules with flexible subunits, proteins embedded in other complex macro-molecules, such as viral capsids, and samples with external components needed for stabilization (nanodiscs, antibodies, etc.). In all these cases, performing a subtraction in the particles of the signal that is not of interest can improve the resolution of the region of interest during the reconstruction and refining process.

Currently, there are methods to perform projection subtraction in Relion and CryoSPARC. However, they do not fully solve the subtraction problem (see Supplementary materials). In this work, we have developed a new method for projection subtraction in Xmipp that improves the performance of subtraction as applied to data from several examples of challenging specimens taken from public databases (EMPIAR, EMDB, and PDB). These improvements come from the fact our approach is able to compute an adjustment for each projection to each input particle individually. Even though we propose two transformations (just grayscale adjustment or grayscale adjustment plus sharpening or dampening of frequencies) to adjust each projection, it has turned out that in most cases, a sharpening or dampening of frequencies is not justified according to our algorithm criteria. Moreover, unlike others, our method can work with the input mask of the region to keep or subtract, providing greater flexibility.

Besides, we have shown that our subtraction approach is very appropriate for the task of finding a small ligand bounded to a large macro-molecule without any assumption on the location of the binding

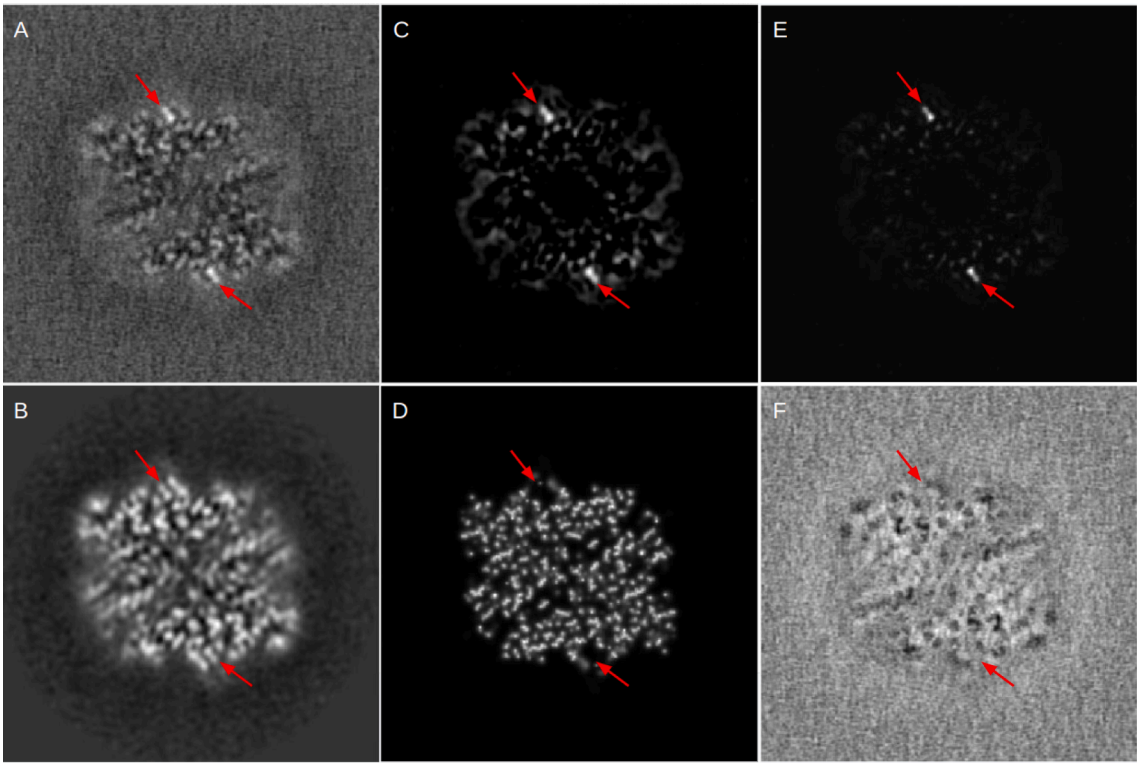


Fig. 9. Equivalent slice of the map coming from particles obtained by processing EMPIAR 10647 so as to keep the ligand (1,6-di-O-phosphono-beta-D-fructofuranose) by (A) Xmipp projection subtraction, (B) volume subtraction without adjustment, (C) Xmipp volume subtraction (Fernandez-Gimenez et al., 2021), (D) volume subtraction with just mean and standard deviation adjustment, (E) multiplication of (A) and (C) as a consensus solution which reinforce the signal from the ligand, (F) difference between (A) and (C). Red arrows point to the regio.n of the ligand.

site.

The algorithm is publicly available at https://github.com/I2PC/xmipp/blob/devel/src/xmipp/libraries/reconstruction/subtract_projection.cpp and can be used through Scipion Framework under the protocol 'subtract projection'.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The authors acknowledge the economic support from MICIN to the Instruct Image Processing Center (I2PC) as part of the Spanish participation in Instruct-ERIC, the European Strategic Infrastructure Project (ESFRI) in Structural Biology. Grant PID2019-104757RB-I00 is funded by MCIN/AEI/ 10.13039/ 501100011033 and "ERDF A way of making Europe", by the European Union. The "Comunidad Autónoma de Madrid" through Grant S2022/BMD-7232, the European Union (EU) and Horizon 2020 through grant HighResCells (ERC-2018-SyG, Proposal: 810057). The authors also acknowledge grant PID2019-104098 GB-I00/AEI/10.13039/501100011033, cofunded by the Spanish State Research Agency and the European Regional Development and grant 2023AEP082 by Agencia Estatal CSIC. Ruben Sanchez-Garcia is funded by an Astex Pharmaceuticals Sustaining Innovation Post-Doctoral Award.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jsb.2023.108024.

References

Dai, X., et al., 2017. Atomic structures of minor proteins VI and VII in human adenovirus. J. Virol. 91 (24), e00850–17.
de la Rosa-Trevin, J.M., et al., 2016. Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. J. Struct. Biol. 195, 93–99.
de la Rosa-Trevin, J.M. 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. doi: 10.1016/j.jsb.2013.09.015.
Fernandez-Gimenez, E., et al., 2021. Cryo-EM density maps adjustment for subtraction, consensus and sharpening. J. Struct. Biol. 213 (4), 107780.
Jimenez-Moreno, A., et al., 2021. Cryo-EM and Single-Particle Analysis with Scipion. J. Visual. Exp.: JoVE.
Kimanius, D., et al., 2021. New tools for automated cryo-EM single-particle analysis in RELION-4.0. Biochem. J. 478 (24), 4169–4185.
Neumann, P., Dickmanns, A., Ficner, R., 2018. Validating resolution revolution. Structure 785–795.
Pettersen, E.f. et al., 2021 UCSF ChimeraX: Structure visualization for researchers, educators, and developers. In: Protein Sci. 30 (2021), pp. 70–82.
Punjani, A., et al., 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods 14, 290–296.
Sorzano, C.O.S., et al., 2015. Fast and accurate conversion of atomic models into electron density maps. AIMS Biophys. 2, 8–20.
Sorzano, C.O.S., et al., 2004. XMIPP: A new generation of an open-source image processing package for Electron Microscopy. J. Struct. Biol. 148, 194–204.
Strelak, D., et al., 2021. Advances in Xmipp for Cryo-Electron Microscopy: From Xmipp to Scipion. Molecules 26, 20.
Vilas, J.L., et al., 2018. MonoRes: automatic and unbiased estimation of Local Resolution for electron microscopy Maps. Structure 26, 337–344.

Supplementary material of "A new algorithm for particle weighted subtraction to eliminate signals from unwanted components in Single Particle Analysis"

E. Fernández-Giménez[1,2], M.M. Martínez[1], R. Marabini[1,2], D. Strelak[3], R. Sánchez-García[4], J.M. Carazo[1], and C.O.S. Sorzano[1,*]

[1]Centro Nac. Biotecnología (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain
[2]Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[3]Institute of Computer Science, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic
[4]Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, United Kingdom
[*]Corresponding author at: Centro Nac. Biotecnología (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain. E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

April 2023

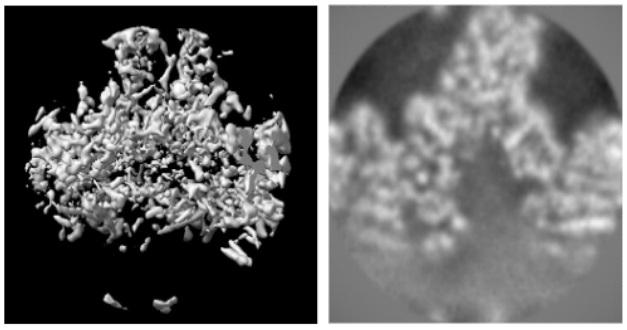


Figure S-1: Monomer B of hexon 1 of the human Adenovirus: Map reconstruction of subtracted particles by CryoSparc (left) and central slice of it (right).

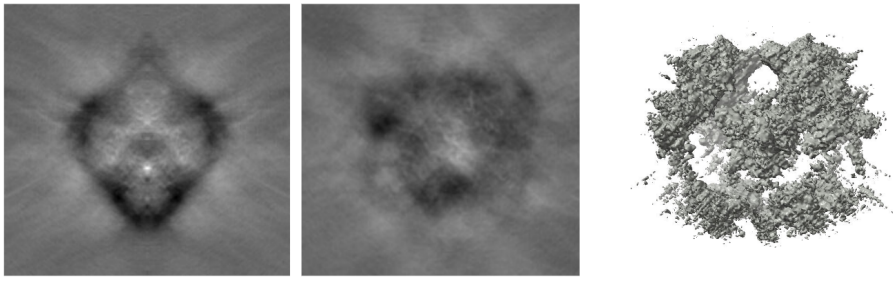


Figure S-2: Two slices (left and center) of the volume obtained by the refinement of the particles in EMPIAR 10005 after being subtracted by Relion and the volume displayed in ChimeraX (right).

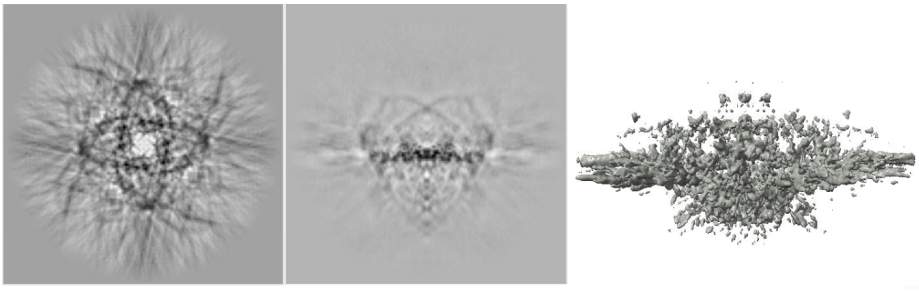


Figure S-3: Two slices (left and center) of the volume obtained by the refinement of the particles in EMPIAR 10005 after being subtracted by CryoSPARC and the volume displayed in ChimeraX (right).

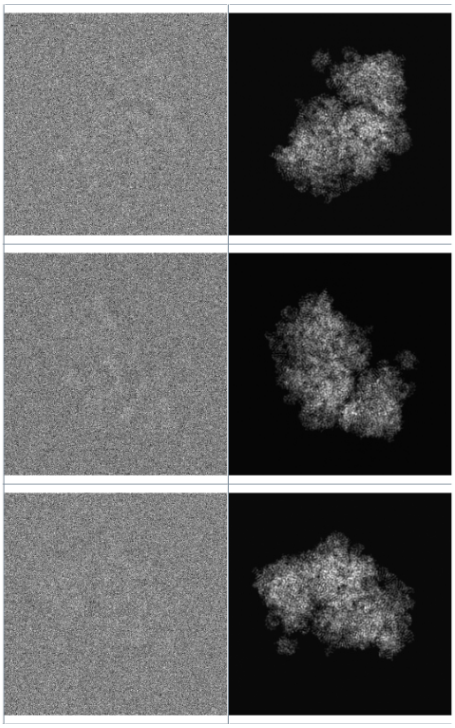


Figure S-4: Three examples of simulated particles obtained by projecting EMD 2660, add simulated CTF and Gaussian noise with a standard deviation of 50 (left) and its corresponding clean projections without CTF nor noise (right).

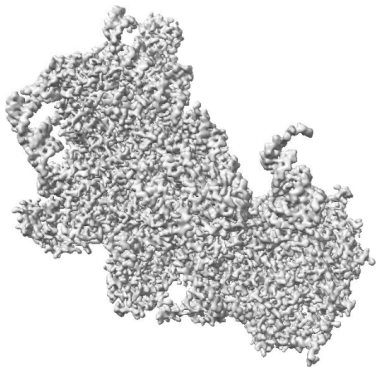


Figure S-5: Reconstructed map of subtracted particles with CryoSparc (subtraction of small subunit of the ribosome in order to keep just the ligand, emetine).

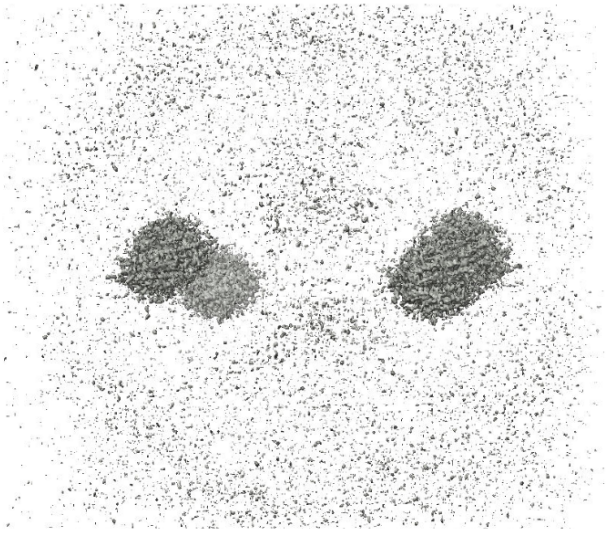


Figure S-6: Reconstruction of particles obtained in the processing of EMPIAR 10647 and subtracted by CryoSparc.

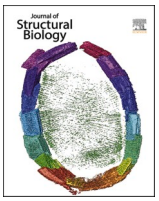
C. Local defocus estimation in Single Particle Analysis in Cryo-Electron Microscopy

Journal of Structural Biology 215 (2023) 108030

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi



Research Article

Local defocus estimation in single particle analysis in cryo-electron microscopy

E. Fernandez-Gimenez [a,b], J.M. Carazo [a], C.O.S. Sorzano [a,*]

[a] Centro Nac. Biotecnologia (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain
[b] Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain

ARTICLE INFO

Keywords:
Local defocus
Defocus
CTF
High-resolution
SPA
Cryo-EM

ABSTRACT

Single Particle analysis (SPA) aims to determine the three-dimensional structure of proteins and macromolecular complexes. The current state of the art has allowed us to achieve near-atomic and even atomic resolutions. To obtain high-resolution structures, a set of well-defined image processing steps is required. A critical one is the estimation of the Contrast Transfer Function (CTF), which considers the sample defocus and aberrations of the microscope. Defocus is usually globally estimated; in this case, it is the same for all the particles in each micrograph. But proteins are ice-embedded at different heights, suggesting that defocus should be measured in a local (per particle) manner. There are four state-of-the-art programs to estimate local defocus (Gctf, Relion, CryoSPARC, and Xmipp). In this work, we have compared the results of these software packages to check whether the resolution improves. We have used the Scipion framework and developed a specific program to analyze local defocus. The results produced by different programs do not show a clear consensus using the current test datasets in this study.

1. Introduction

Single Particle Analysis (SPA) by cryo-electron microscopy (cryo-EM) has become an established field to elucidate the atomic model of macro-molecules and biological complexes, as it can obtain high-resolution (below 3 Å) Coulomb potential maps. Achieving good resolution maps is becoming easier using state-of-the-art methods(Neumann et al., 2018; Vilas et al., 2022). However, to obtain high resolution, details matter. One of the best-known sources of issues in cryo-EM is the proper correction of the contrast transfer function (CTF) (Sorzano et al., 2021b).

Out-of-focus image acquisition, i.e., defocus, is the main phase contrast mechanism in cryo-EM (Danev et al., 2020). A good CTF estimation allows for correcting aberrations and defoci, as CTF is affected by these acquiring conditions. Thus, this function must be determined for each micrograph during its processing. CTF estimation and correction is one of the first steps in the general workflow used in SPA. However, this is usually done in a global approach, as the CTF is estimated and corrected for the whole micrograph, i.e., each particle of the same micrograph has the same CTF correction.

Nevertheless, it is known that a more precise CTF correction could be done by refining the global estimation for each particle, called local CTF estimation or refinement. In this way, based on the global CTF determination for each micrograph, some methods estimate the local CTF correction for each particle that appears in that micrograph. In this study, we are specifically interested in the local defocus estimation.

The local defocus refinement step is important when one wants to improve resolution once the map is already at high resolution (or close to it), as the sample has a certain thickness and the macro-molecules in it have been frozen at different heights inside the ice layer(Noble et al., 2018) (see Fig. 1). This causes small differences in each particle's defocus value, even though they are in the same micrograph. Still, these differences are big enough to cause blurring if the same defocus is used for all particles in the same micrograph. There are several state-of-the-art methods that perform local defocus estimation, such as Gctf (Zhang, 2016), Relion(Zivanov et al., 2020), CryoSPARC(Punjani and Fleet, 2020), and Xmipp(Strelak et al., 2021; de la Rosa-Trevin et al., 2013; Sorzano et al., 2004).

All these packages estimate the local defocus, and in this work, we try to estimate how precise those estimations are. As we cannot compare the different estimations versus the (unknown) ground truth, we have performed a comparative study between the results of the four state-of-

* Corresponding author.
E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

https://doi.org/10.1016/j.jsb.2023.108030
Received 1 July 2023; Received in revised form 30 August 2023; Accepted 21 September 2023
Available online 25 September 2023

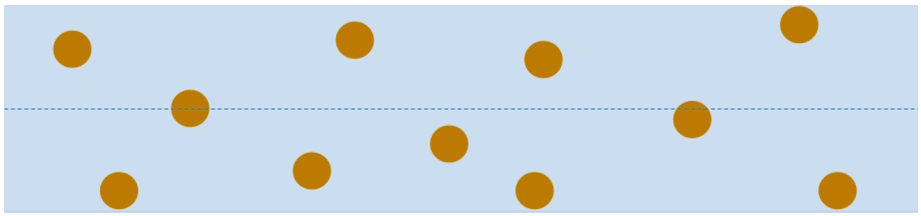


Fig. 1. Distribution of proteins along the ice thickness of the sample(Noble et al., 2018). The dotted line represents the global defocus estimation for the micrograph. However, the height of each particle (schematically represented by yellow dots) in the sample does not agree in many cases with the height corresponding to the global defocus.

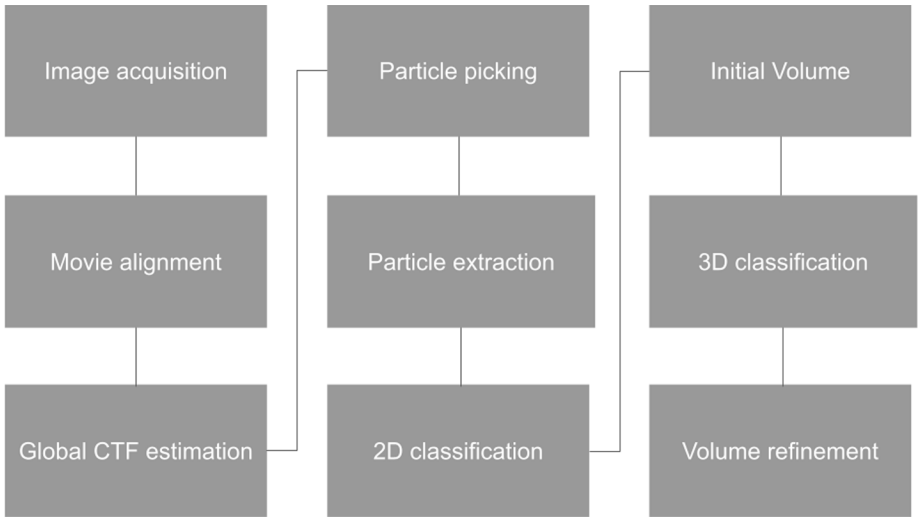


Fig. 2. Standard single particle analysis (SPA) workflow.

the-art methods mentioned before. We aim to know how similar these estimations are between them and corroborate if it is true that local estimations are precise enough to improve resolution, as this is a final step that is performed only in some specific workflows where the user wants to push forward the resolution after refinement and it is difficult to validate.

2. Methods

To perform the comparative study on local defocus estimation, we use Scipion (Jimenez-Moreno et al., 2021; de la Rosa-Trevin et al., 2016) since it is a general framework for cryo-EM image processing that allows the combination and compatibility of different cryo-EM software packages. This simplifies the comparison of results and their interpretation. Thus, we have performed a conventional SPA workflow from movies to final refined volumes. Then, we have included and focused on the different state-of-the-art methods to estimate local defocus (Gctf, Relion, CryoSPARC, and Xmipp).

2.1. Data

The data for this study is an apoferritin sample acquired with a 300 kV CryoArm microscope and a K3 camera at the Spanish National Center for Biotechnology (CNB) Cryo-EM Facility, initially used to check the microscope setting and performance. It consists of 8,721 movies with a pixel size of 0.23Å/pixel, leading to 7,815 micrographs (leaving some of the movies out due to excessive motion blur). Apoferritin has a diameter of approximately 12 nm, and the thickness of the ice in this sample is

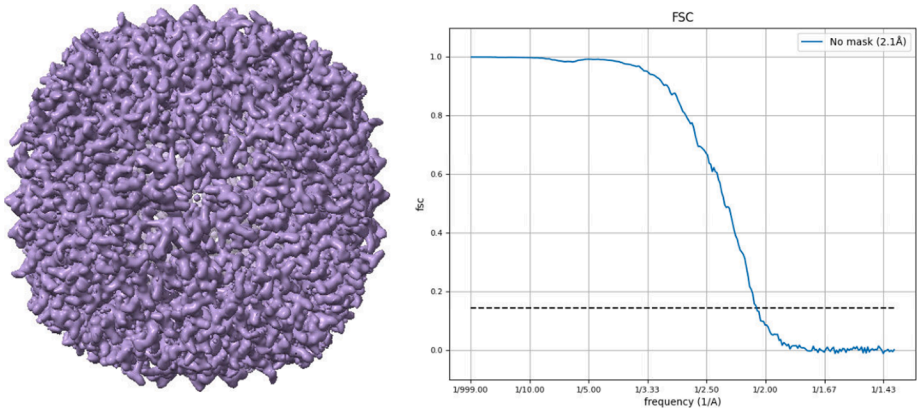


Fig. 3. Final density map for the apoferritin refined with CryoSPARC (considering global defocus estimation) and reported Fourier Shell Correlation (FSC).

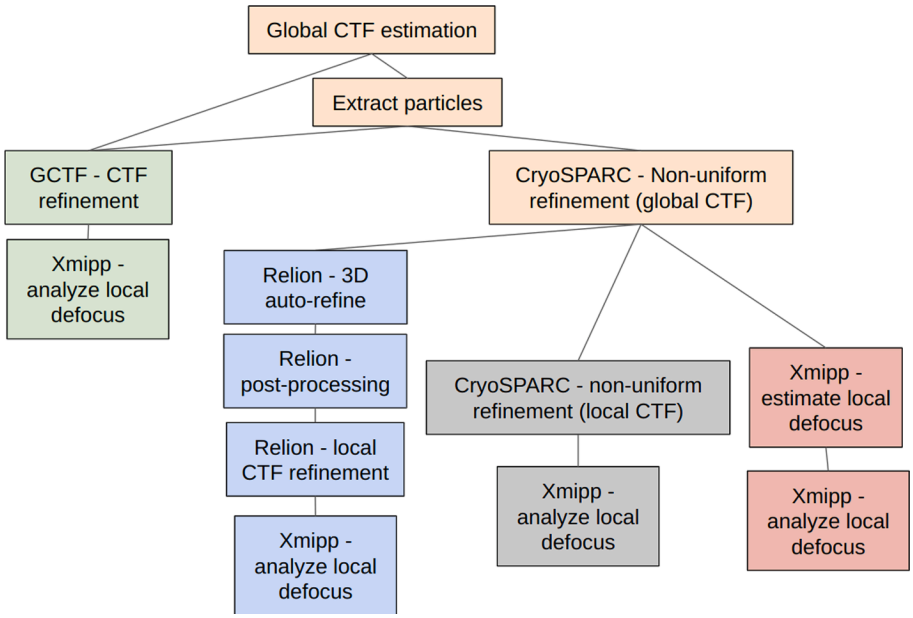


Fig. 4. Local defocus workflow in Scipion developed for this study, including the proposed analysis protocol.

approximately 40 nm, which leaves enough space for the sample to be placed at different heights (defocus) in the thickness of the sample. Apoferritin has been deeply studied in cryoEM (Yip et al., 2020), and it is known to be a very stable protein with high-order symmetry that allows high resolution with standard SPA processing. An in–house acquisition has been chosen instead of a data set coming from a public database to have all the acquisition information and intermediate steps along the processing. This procedure gives us more flexibility to execute the different defocus estimation algorithms at different workflow points.

2.2. Workflow

A common SPA workflow was carried out at the Instruct Image Processing Center (I2PC), combining different software packages in Scipion. The main steps followed in the workflow are listed below, and a standard SPA workflow is summarized in Fig. 2:

1. Movie alignment using MotionCor2 (Zheng et al., 2017).
2. CTF estimation using Gctf (Zhang, 2016).
3. Particle picking with Cryolo(Wagner et al., 2019) and Gautomatch.
4. Extraction of downsampled particles with Relion (Zivanov et al., 2020).
5. Building of initial model with CryoSPARC(Punjani et al., 2017).
6. Several iterative steps of 2-dimensional classification with CryoSPARC to discard bad particles.
7. Re-extraction of original size particles with Xmipp (Strelak et al., 2021; de la Rosa-Trevin et al., 2013; Sorzano et al., 2004).
8. Several iterative steps of CryoSPARC non-uniform refinement(Punjani et al., 2017) with initial model and re-extracted original size particles as input, with global CTF refinement and global beam tilt refinement but no local defocus refinement, achieving a map resolution of 2.1Å as shown in Fig. 3.

Once a refined reconstruction is obtained, we tried to push forward the reconstruction quality by performing a local defocus estimation and correction. To do that, different algorithms were used.

Firstly, we include “Gctf - ctf refinement” after the re-extraction of particles at their original size. This algorithm needs as input the set of particles without alignment, the set of micrographs, and the corresponding set of global CTF estimations. Note that to use the “Gctf - ctf refinement” program, we had to use an older version of Gctf (v. 1.06), as the current one (v. 1.18) does not support local CTF refinement anymore.

As for the defocus refinement algorithms of Relion (v.3.0.0), CryoSPARC (v.4.0.7), and Xmipp (v.22.4.0), they all make use of the same kind of input: the aligned set of particles and a refined reference volume. Note that to perform CryoSPARC defocus refinement, we can carry out the whole non-uniform refinement program because it is one of its options or we can run it as a separate step after the refinement (obtaining practically identical results in both cases). In the case of Relion, there is a specific program to run CTF refinement. However, it needs as input the output of the Relion post-processing program, which in turn needs as input the output of Relion auto-refine (which needs as input the set of particles and a reference volume). In both cases, these complex specificities occur because we are using these different programs outside of their standard workflows, and it makes clear the point that mixing different software suites is not easy; however, Scipion can bridge these issues easily. Finally, Xmipp has a dedicated program to compute CTF refinement, taking as input directly the particles and volume out from any refinement program. Fig. 4 shows the workflow for all the local defocus estimation software tested in this work.

2.3. Xmipp algorithm for Local Defocus Estimation

Each software suite uses different procedures to obtain the local defocus estimation. In this section, we describe how it is done by the algorithm inside the software suite that we develop in our laboratory, that is, Xmipp. Xmipp estimation of local defocus consists of several steps that begin with the computation of the projection of the reference volume corresponding to the alignment of each input particle. Then, the global defocus estimated previously is applied to the corresponding projection of every particle. After that, the correlation between the input particle and the corresponding projection is computed. Finally, the global defocus estimation is refined using the Powell optimization method to look for a local minimum that better matches the particle and the corresponding projection as measured by the correlation. These steps are executed for each input particle. The method is stand-alone to refine local defocus for a set of aligned particles and their refined volume (as used in this study). Still, it is also used in the refinement algorithm HighRes(Sorzano et al., 2018) when the “optimize defocus” option is selected.

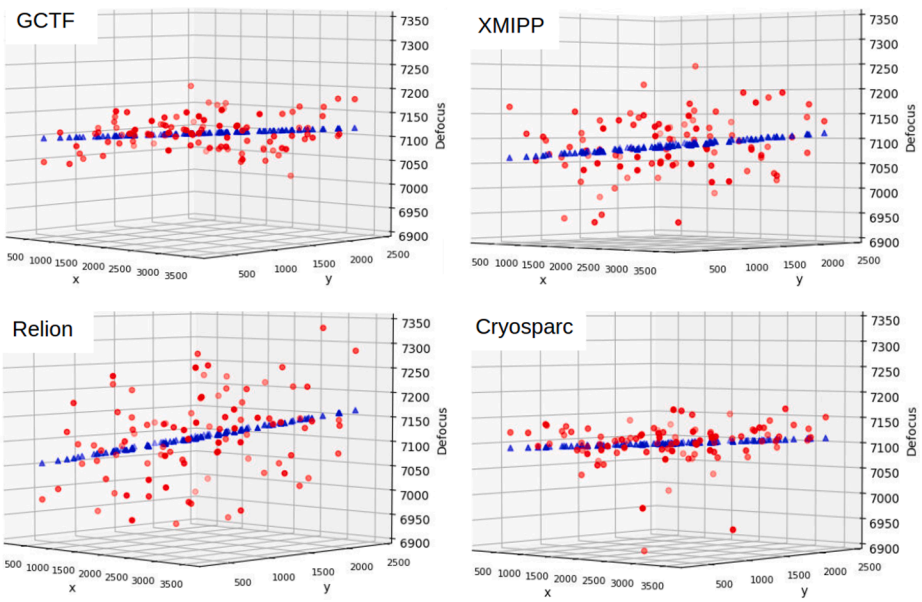


Fig. 5. Plots computed by our analysis protocol of the distribution of all the particles in red (coordinates X and Y are the positions of the particle in the micrograph, while Z coordinate corresponds to the local defocus estimated for the particle in Å) along the thickness of the sample for a particular micrograph of the data set according to the different local defocus estimation software. Note that small differences in the rotation of the axes in each subplot have been made for convenience to better visualize the adjustment plane (blue) in each case.

2.4. Local defocus estimation analysis

To compare different local defocus estimations, we have developed a Scipion protocol that analyses the output of any local defocus estimation software. This protocol uses a least-squares approach to compute a linear fitting (a plane) according to Eq. 1 from the set of local defocus estimations of the particles in each micrograph.

$$\Delta f_i \approx \widehat{\Delta f}_i = \Delta f_0 + a y_i + b x_i, \tag{1}$$

where Δf_0 is the global defocus, Δf_i the local defocus of the i-th particle, and (x_i, y_i) its position in the micrograph. The result is a plot per micrograph of the three-dimensional distribution of the particles in each micrograph, that is, the $(x_i, y_i, \Delta f_i)$ points. The three-dimensional position of each particle in the micrograph and the computed fitting plane are plotted. Thus, this plot approximates the distribution of heights of particles inside the ice and informs about the local defocus estimated variations of the particles in the micrograph.

3. Results and discussion

Section 3.1 shows quantitative differences between the local defocus estimations computed by the different state-of-the-art methods (Gctf, Relion, CryoSPARC, and Xmipp). As the reader will notice, there are perceptible differences, thus, in Section 3.2, we try to elucidate which estimations are reliable.

3.1. Differences between estimations by different software

To better illustrate that the different software programs estimate different local defocus for each particle, we have chosen one representative micrograph of the data set. We show in Fig. 5 the plot produced by our analysis protocol for the same micrograph with the different local

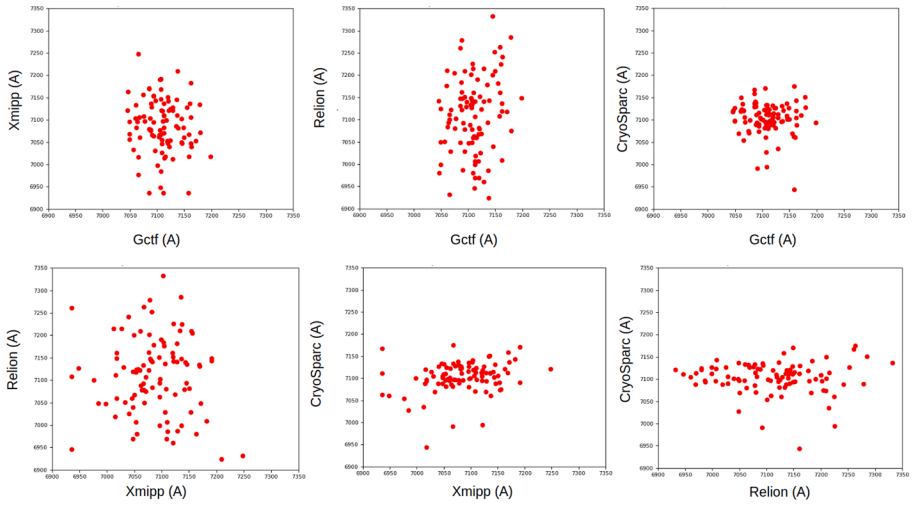


Fig. 6. Scatter plots comparing the local defocus estimations (in Å) of each method versus the others for particles in a selected micrograph of the data set to see how they correlate (perfect correlation would be a diagonal line).

Table 1
Average correlation matrix between the local estimations by the different software and global estimation. The major correlation between methods is in bold.

	Gctf	Relion	Xmipp	CryoSPARC	Global
Gctf	1				
Relion	0.12	1			
Xmipp	0.17	0.12	1		
CryoSPARC	0.26	0.15	**0.40**	1	
Global	0.04	0.25	0.04	0.04	1

defocus estimations. Remarkably, the shape of the defocus distribution of particles in the same micrograph differs according to each method (both the adjusted plane and the distribution around it). It is true that if we consider the dispersion range, the differences are small, meaning that the estimations of the different programs are not very far away. However, as this is considered a refinement step to push forward resolution once the density map is already in high resolution, this result may question the accuracy of the different methods.

To show and compare the differences between the results of the different methods with this chosen example micrograph, we have computed a pair-wise scatter plot of the estimations of the different methods to see how they correlate. Scatter plots are shown in Fig. 6. Note that for total correlation, a diagonal line should appear in the plot; thus, from Fig. 6, we can state that the different methods do not agree. Moreover, their different estimates are completely independent of each other, that is, what one would expect from a random estimation.

To quantify these differences, we have computed the correlation matrix between all the methods for each micrograph. Then, we averaged all matrices, obtaining the average correlation matrix shown in Table 1. Note that we cannot directly compute the correlation matrix of the particles in all micrographs, as the global defocus for each micrograph is different (as they have been acquired at different defocus on purpose). As can be seen from the average correlation matrix, in this particular data set, the major agreement between methods is between CryoSPARC and Xmipp (but note that it is not even 0.5), and the worst is between Gctf and Relion. However, this fact may not be extrapolated to any other data set.

3.2. Which local defocus estimation should we trust?

From the previous section, we have shown that local defoci estimations by different software have considerable differences. Unfortunately, the ground truth about local defoci is unknown, so we cannot be sure if any or none of the different estimates are correct. Thus, we have tried to evaluate the quality of each estimation separately.

Firstly, we have computed the scatter plots of each estimation versus the global for the same subset of random particles, shown in the Supplementary material Fig. S-1. Note that in this plot, the range is much bigger as they consider the defocus of every micrograph (and usually, in image acquisition, micrographs are acquired at different defocus on purpose). Thus, we can see that all the estimations correlate fairly well with the global estimation, as expected, meaning that none of them is completely erroneous, as a local defocus estimation may vary from the global, but less than half of the ice thickness (if the variation is bigger, it would mean that the particle is out of the sample, which obviously would be an erroneous estimation).

Afterward, we checked the stability of the methods, as differences in local defoci are small and can be highly affected by noise in the computations. Thus, we have executed two times each local defocus estimation program for particles in the example micrograph used before, and we have computed the scatter plot of each pair of executions to see the correlation. The plots are shown in Fig. 7.

To avoid identical inputs, we have performed two different CryoSPARC non-uniform refinements with the same input particles and reference volume and the same parameters: symmetry, filter type, global CTF refinement parameters (as minimum resolution, global beam tilt refinement, and spherical aberration). Note that refinement is not 100% stable and thus, angles and shifts may slightly vary from one execution to another even with the same parameters(Sorzano et al., 2021a). The outputs of these two refinements have been used separately with each of the local defocus estimation methods (except for Gctf, due to it requiring other inputs). In summary, we have repeated the workflow in Fig. 4 from "CryoSPARC - Non-uniform refinement (global CTF)" to the end.

In the case of Gctf, which uses as input a set of non-aligned particles (i.e., directly from the extraction step), to avoid identical input, we have

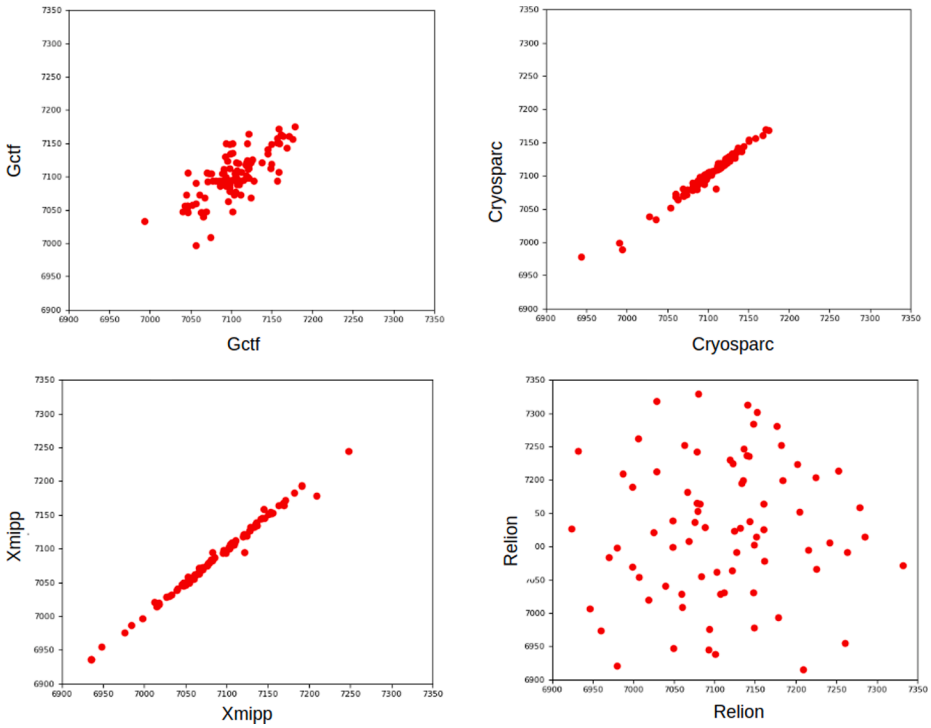


Fig. 7. Scatter plots comparing two executions of each local defocus estimation method for a specific micrograph of the data set to evaluate their stability.

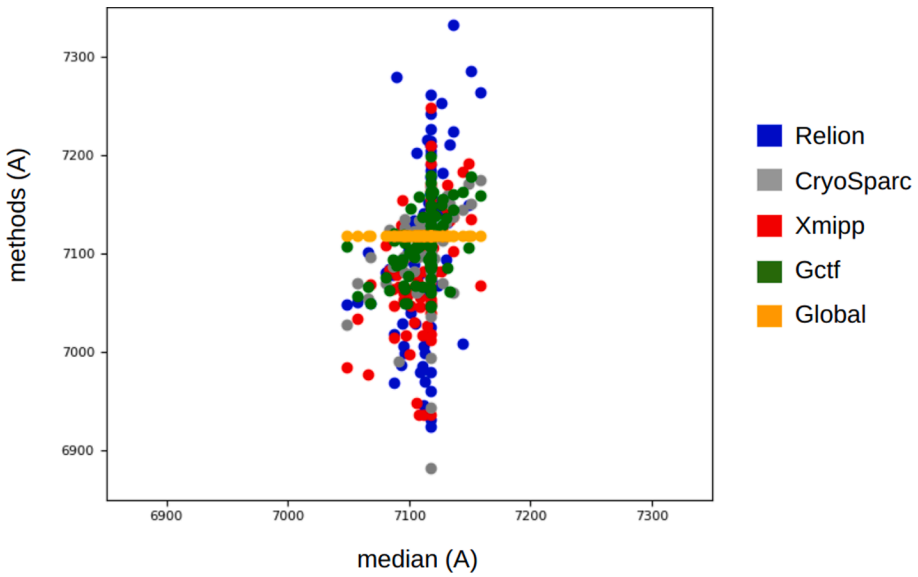


Fig. 8. Scatter plot comparing the median local defocus estimation (in Å) with each local defocus estimation method (in different colors) for all the particles in the chosen example micrograph to see how they correlate and how the different estimations spread around the unknown ground truth.

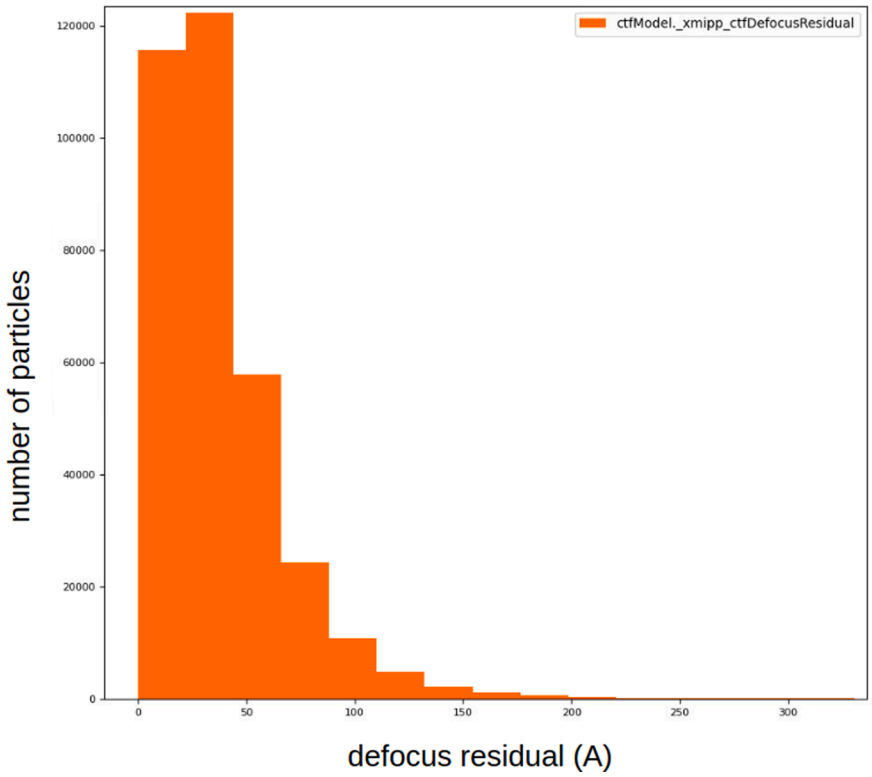


Fig. 9. Histogram of the residuals (computed as the median absolute deviation, MAD) between the median of all local defocus estimations and the original global estimation for each particle.

slightly modified the X and Y coordinates of the input particles by moving them the equivalent of half the size of the protein (i.e., as if particles have been picked off-center). As can be appreciated in Fig. 7, in this case, CryoSPARC and Xmipp are more stable than Gctf and Relion.

As discussed above, we cannot know if a local defocus estimation is correct. We can check if some estimation is too far from the original global one, which would be a clue for an incorrect estimation. We can also check its stability as a measure of reliability. Although some methods are (at least in this case) more stable than others, the differences in the results of different executions are indeed too small to be a reason to fully discard an estimation method (but one must be aware of this instability).

We can compare their estimations to see if they agree, which is not proof of correctness. But if two results computed by two different methods agree makes one think that it is more likely that the solution is correct, as suggested in Sorzano et al. (2021a). Thus, by comparing estimations, we can check that they are close enough not to be able to discard any, and we can also check that the result seems correct for the

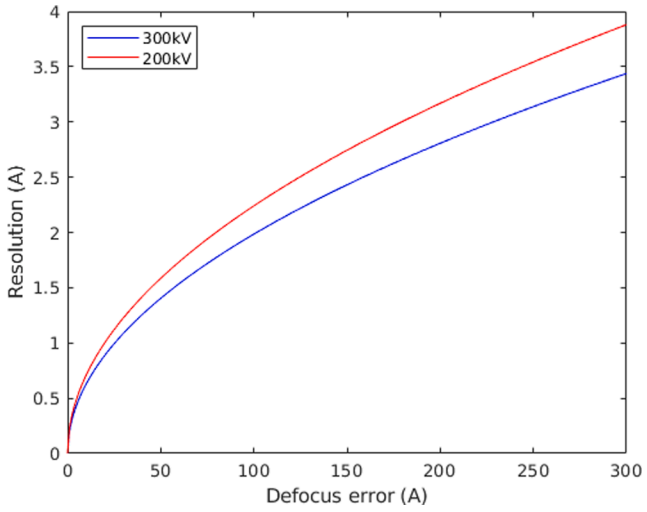


Fig. 10. Resolution limit, $|\mathbf{R}|^{-1}$, with different defocus error for a 300 kV microscope (blue) and for a 200 kV microscope (red) when $\Delta\chi(\mathbf{R}) = \pi/2$.

reason stated above. Small fluctuations around the ground truth are expected for unbiased estimates of the true local defocus (Sorzano et al., 2021a; Sorzano et al., 2021b), and this is certainly the case in this experiment (see Fig. 8). Suppose we want to gain accuracy in this step. In that case, we may take an average (or median) of the different estimates, and the average should have less noise than any of the individual estimations. Fig. 9 shows the Median Absolute Dispersion (MAD) between the estimations of the local defocus for each particle. It can be observed that most of the local defocus estimates are within 50Å. A defocus error of 50Å means a shift of $\pi/2$ in the CTF at a resolution of 1.4Å, which will cause a wrong correction of the CTF, which will worsen the final resolution (see Supplementary material Fig. S-2). Logically, the accuracy in the estimation of the local defocus is directly related to the achievable resolution in the reconstructed map (Zhang and Zhou, 2011). This is shown in Fig. 10, which has been generated from Eq. 2(Sorzano et al., 2007),

$$\begin{aligned}\chi(\mathbf{R}) &= \pi\lambda\left(|\Delta f\left(\mathbf{R}\right)||\mathbf{R}|^2 + \frac{1}{2}C_s|\mathbf{R}|^4\lambda^2\right)\\ \Delta\chi(\mathbf{R}) &= \pi\lambda\Delta|\Delta f\left(\mathbf{R}\right)||\mathbf{R}|^2\end{aligned} \tag{2}$$

where C_s represents the spherical aberration coefficient and λ is the electron wavelength computed as:

$$\lambda = \frac{1.23\times10^{-9}}{\sqrt{V + 10^{-6}V^2}}, \tag{3}$$

V is the acceleration voltage of the microscope. The MAD of the different local defocus estimates can be used to filter out those particles whose defocus is uncertain.

Finally, we have performed a reconstruction for each local defocus estimation keeping the same alignment (angles and shifts) in all the cases, which are shown in Fig. 11 showing the local resolution (local resolution histograms are in Supplementary material Fig. S-3) and the corresponding Fourier Shell Correlation (FSC, computed in Xmipp) in Fig. 12. Note that for this specific dataset the only local defocus estimation that has improved the global resolution is the one computed by CryoSPARC, which has obtained an improvement in global resolution of 0.1Å(from 2.1Åto 2.0Å). A similar result has been obtained with EMPIAR 10647 dataset (a non-globular protein), shown in Figs. S-4 and S-5 of Supplementary materials.

3.3. Local defocus refinement is not always helpful

In this section, we show the global and local defocus estimations for beta-galactosidase (EMPIAR 10061). As shown in Fig. 13 (local resolution histograms are shown in the Supplementary material Fig. S-6). In this case, local defocus estimations worsen the resolution achieved in the reconstruction, except for Gctf estimation, which remains similar to global defocus estimation. Thus, in this case, it seems that is not worth refining local defocus estimation. The cause is unknown, but for some reason, the local refinement programs cannot correctly estimate the local defocus.

4. Conclusions

To achieve the highest possible resolution in Cryo-EM SPA, it is important to be precise when estimating every parameter. Thus,

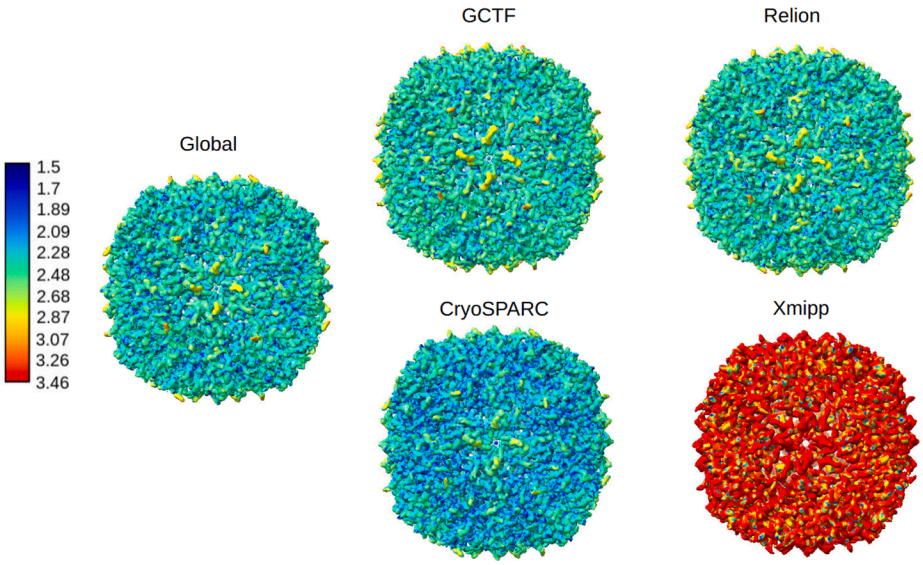


Fig. 11. Local resolution computed with Monores(Vilas et al., 2018) of reconstructed density maps with global defocus and local defocus computed by the existent state-of-the-art methods, keeping the same alignment (angles and shifts) in all the cases.

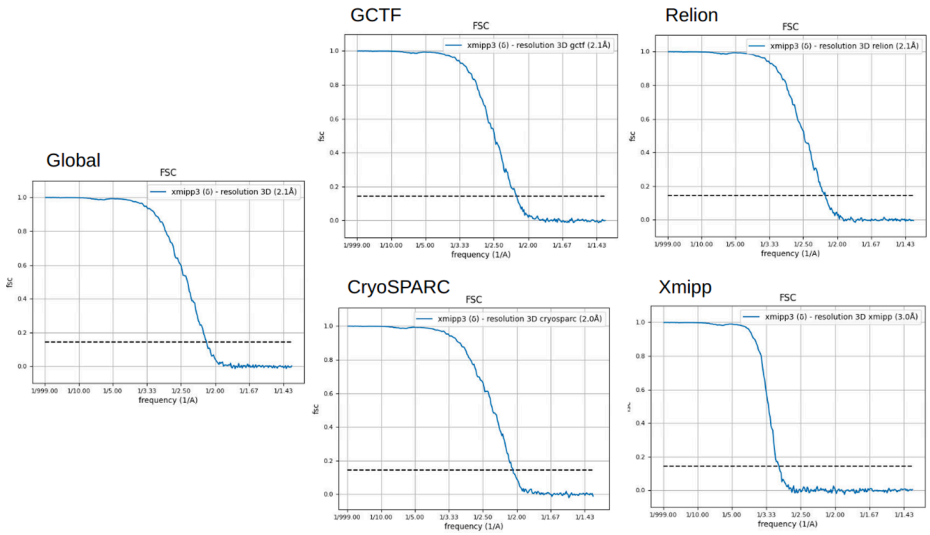


Fig. 12. Fourier Shell Correlation (FCS) to measure the global resolution of each apoferritin reconstruction in Fig. 11.

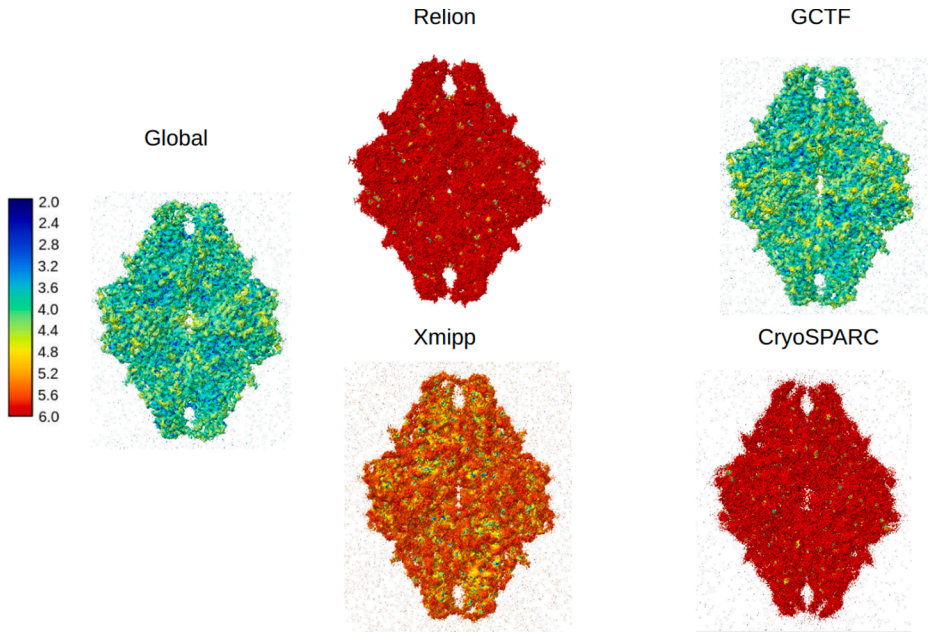


Fig. 13. Local resolution of refined density maps with global defocus and the different local defocus computed by the existent state-of-the-art methods.

estimating an accurate defocus for each particle instead of just a global value per micrograph should result in a more reliable reconstructed map. A global defocus per micrograph assumes that all particles within the micrograph are at the same height (same defocus) with respect to the imaging plane. But we know that it is not true. If these differences in defocus in the same micrograph are not properly corrected, these inaccuracies will lead to blurring and to an inaccurate CTF correction when refining the volume. These factors are limiting as a high resolution is approached (Zhang and Zhou, 2011).

In this work, we have tried four different state-of-the-art programs (Gctf, Relion, CryoSPARC, and Xmipp) to compute local defocus. Their estimations are close, although a deeper analysis reveals noticeable differences, especially because refinement is a step performed to gain accuracy. As with many other parameters in SPA, we cannot know the ground truth for the local defocus of each particle, and thus, we cannot know which estimation (if any) is correct. Then, the best we can do to evaluate our estimations before computing the reconstructions, in order to help with the interpretation of local CTF estimations (not the CTF estimation itself) potentially in those datasets that show large CTF variations, is to compare the estimations computed by the different defocus refinement methods. To do so, we have developed an analysis protocol inside Scipion that allows us to study in detail the resulting defocus estimation of the different programs and put them in the same framework to compare them directly.

Moreover, we have checked that none of the estimations produces unreasonable results that would immediately suggest discarding them. However, the results produced by different programs do not show a clear consensus using the current test datasets.

We recall that it is known that relatively small errors in defocus (50–100Å) translate into noticeable offsets ($\pi/2$) in CTF at high resolution (1.4–2Å), which are indeed resolution values reported in resolved structures nowadays. This fact indicates that the type of defoci differences between methods that we are reporting in this work may be a limiting factor when working at very high-resolution regimes.

Local CTF estimation remains a complex issue and, as a final note, we remark that local defocus refinement does not result always in an

improvement, especially when the signal-to-noise ratio (SNR) is low and the resolution achieved is not good and homogeneous, as we show with the beta-galactosidase example.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

We thank the Spanish National Center for Biotechnology Cryo-EM Facility for yielding the data for this study, especially Javier Chichón and Roberto Melero from I2PC. The authors acknowledge also the economic support from MICIN through grant PID2019 - 104757RB - I00 funded by MCIN/ AEI/ 10.13039/ 501100011033/ and "ERDF A way of making Europe", by the "European Union". To the European Union (EU) and Horizon 2020 through grant: HighResCells (ERC - 2018 - SyG, Proposal: 810057).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jsb.2023.108030.

References

Danev, R., et al., 2020. Fast and accurate defocus modulation for improved tunability of cryo-EM experiments. IUCrJ 7 (3), 566–574.
de la Rosa-Trevin, J.M., et al., 2016. Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. J. Struct. Biol. 195, 93–99.
de la Rosa-Trevin J.M., 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. doi: 10.1016/j.jsb.2013.09.015.
Jimenez-Moreno, A., et al., 2021. Cryo-EM and Single-Particle Analysis with Scipion. J. Visual. Exp.: JoVE.
Neumann, P., Dickmanns, A., Ficner, R., 2018. Validating resolution revolution. Structure 785–795.
Noble, A.J., et al., 2018. Routine single particle cryoEM sample and grid characterization by tomography. eLife.
Punjani, A., Fleet, D.J., 2020. 3D variability analysis: directly resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM images. bioRxiv. https://doi.org/10.1101/2020.04.08.032466.
Punjani, A., et al., 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods 14, 290–296.
Sorzano, C.O.S., et al., 2018. A new algorithm for high-resolution reconstruction of single particles by electron microscopy. J. Struct. Biol. 204, 329–337.
Sorzano, C.O.S., et al., 2007. Fast, robust and accurate determination of transmission electron microscopy contrast transfer function. J. Struct. Biol. 160, 249–262.
Sorzano C.O.S. et al., 2021a. Image Processing in Cryo-Electron Microscopy of Single Particles: The Power of Combining Methods. In: Structural Proteomics (2021), pp. 257–289.
Sorzano, C.O.S., et al., 2021b. On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy. Acta Crystallogr. Sect. D 78, 410–423.
Sorzano, C.O.S., et al., 2004. XMIPP: A new generation of an open-source image processing package for Electron Microscopy. J. Struct. Biol. 148, 194–204.
Strelak, D., et al., 2021. Advances in Xmipp for Cryo-Electron Microscopy: From Xmipp to Scipion. Molecules 26, 20.
Vilas, J.L., et al., 2018. MonoRes: automatic and unbiased estimation of Local Resolution for electron microscopy Maps. Structure 26, 337–344.
Vilas, J.L., Carazo, J.M., Sorzano, C.O.S., 2022. Emerging themes in CryoEM-single particle analysis image processing. Chem. Rev. 122 (17), 13915–13951.
Wagner, T., et al., 2019. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. In: Communications Biology 2, p. 218.
Yip, K.M., et al., 2020. Atomic-resolution protein structure determination by Cryo-EM. Nature 587, 157–161.
Zhang, K., 2016. Gctf: real-time CTF determination and correction. J. Struct. Biol. 193, 1–12.
Zhang, X., Zhou, Z.H., 2011. Limiting factors in atomic resolution cryo electron microscopy: no simple tricks. J. Struct. Biol. 175, 253–263.
Zheng, S.Q., et al., 2017. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods 14, 331–332.
Zivanov, J., Nakane, T., Scheres, S.H.W., 2020. Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in Relion-3.1. IUCrJ 7, 253–267.

Supplementary material

Estrella Fernandez-Gimenez[1,2], J.M. Carazo[1], and C.O.S. Sorzano[1,*]

[1]Centro Nac. Biotecnologia (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain
[2]Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[*]Corresponding author at: Centro Nac. Biotecnologia (CSIC), c/Darwin, 3, 28049 Cantoblanco, Madrid, Spain. E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

July 2023

1 Apoferritin

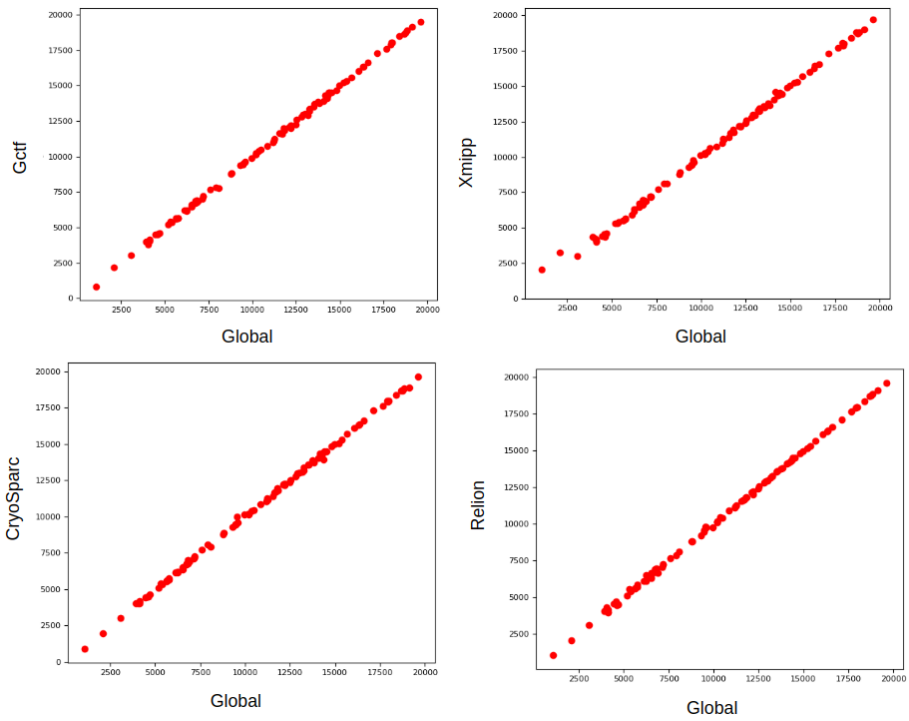


Figure S-1: Scatter plots comparing the local defoci estimations of each method versus the global defoci estimation for a set of random particles of the data set (from micrographs with different defoci) to see how they correlate.

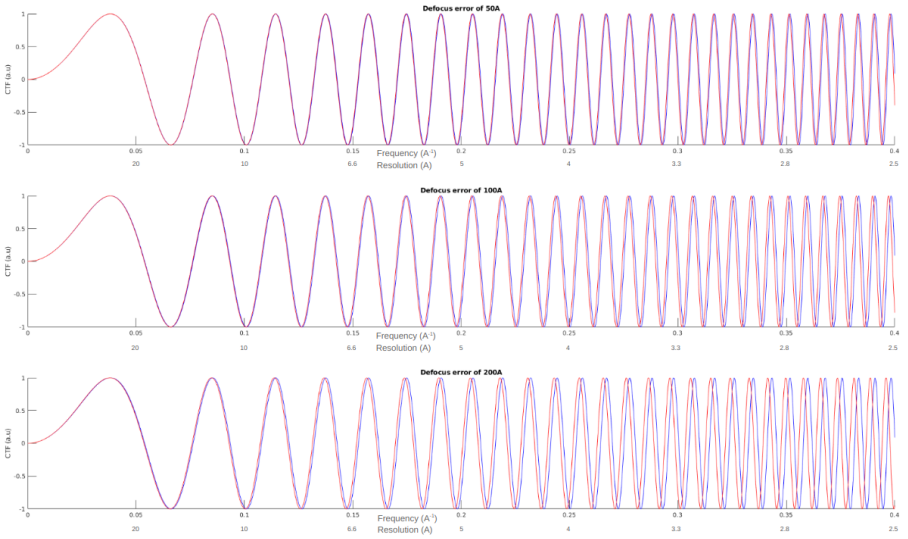


Figure S-2: Offset in CTF produced by different defocus errors. An offset of $\Delta\chi(\mathbf{R}) = \pi/2$ with a defocus error of 50Å occurs approximately at 1.4Å resolution. With a defocus error of 100Å, it occurs at 2Å and with a defocus error of 200Å occurs at 3.3Å, which are all resolutions achievable nowadays.

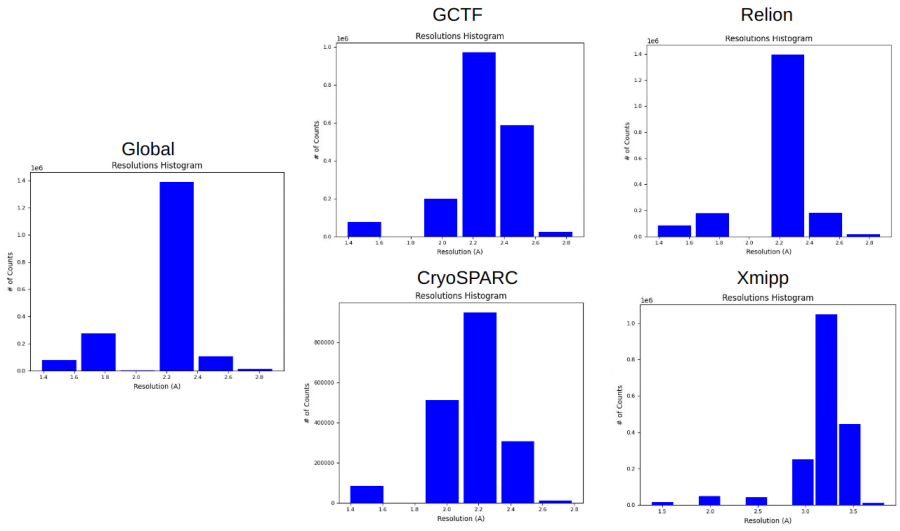


Figure S-3: Resolution histograms of apoferritin with different defocus. The histograms have been computed with MonoRes.

2 PKM2 Enzyme (EMPIAR 10647)

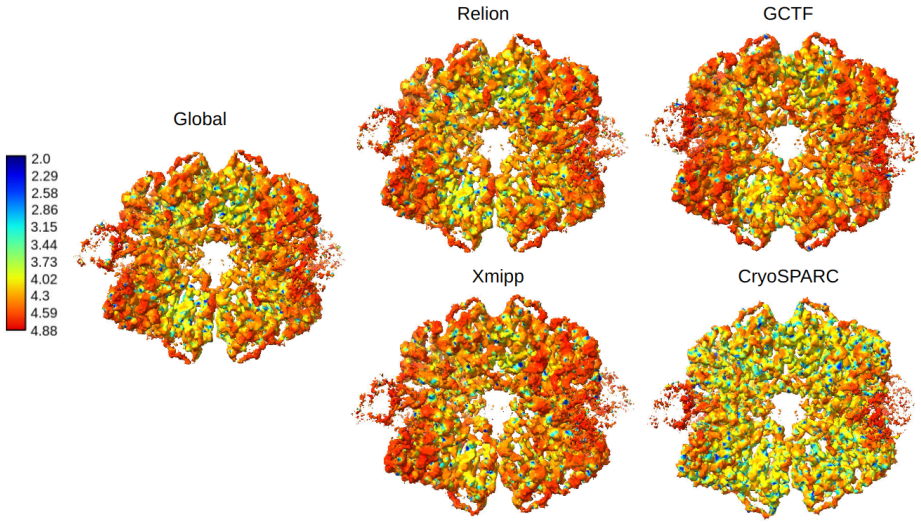


Figure S-4: Local resolution of refined density maps with global defocus and local defocus computed by the existent state-of-the-art methods.

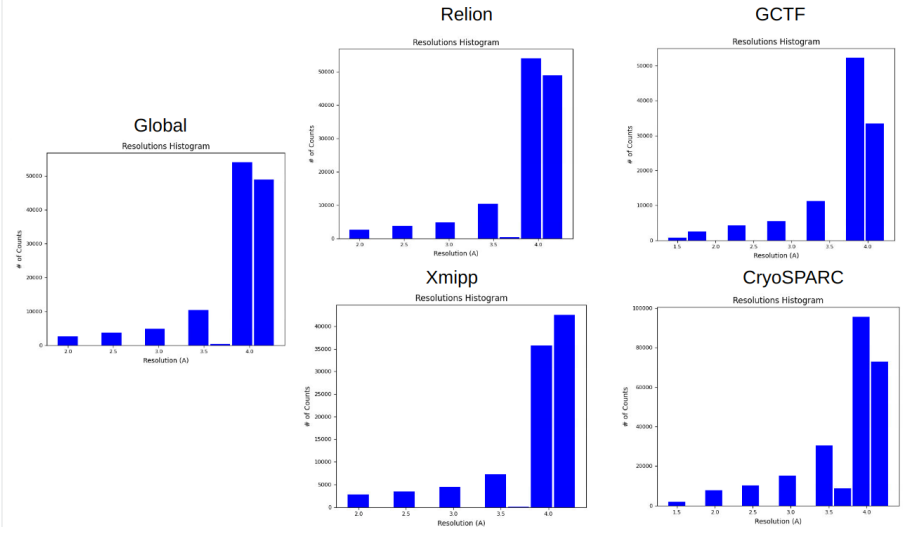


Figure S-5: Resolution histograms of PKM2 enzyme with different defocus. The histograms have been computed with MonoRes.

3 Beta-galactosidase (EMPIAR 10061)

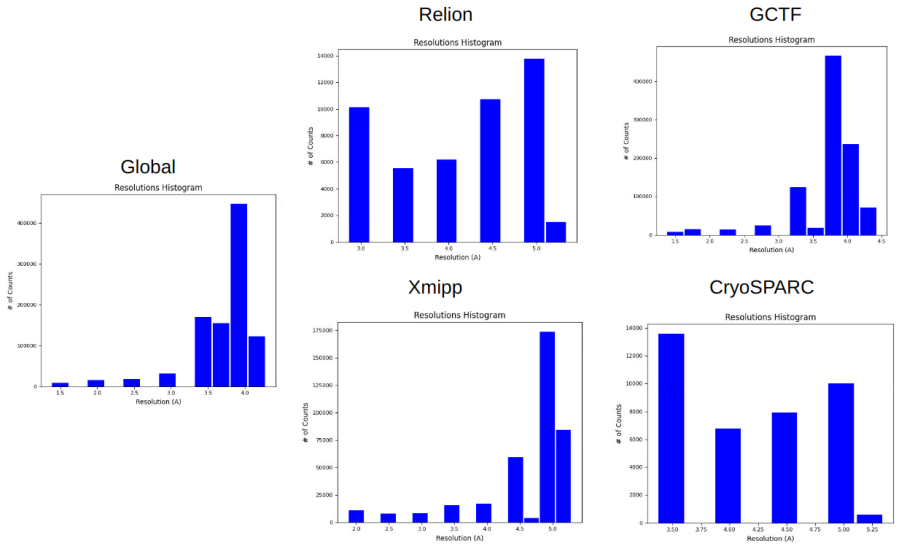


Figure S-6: Resolution histograms of beta-galactosidase with different defocus. The histograms have been computed with MonoRes.