UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA DE TELECOMUNICACIÓN

TRABAJO DE FIN DE MÁSTER

Development of a fast image alignment algorithm for Cryo-Electron Microscopy

OIER LAUZIRIKA ZARRABEITIA

2023

MÁSTER UNIVERSITARIO EN INGENIERÍA DE TELECOMUNICACIÓN

TRABAJO DE FIN DE MÁSTER

Título:	Development of a fast image alignment algorithm for Cryo-Electron Microscopy
Autor:	Oier Lauzirika Zarrabeitia
Tutor:	Carlos Óscar Sorzano Sánchez Narciso García Santos
Departamento:	Departamento de Señales, Sistemas y Radiocomunicaciones

MIEMBROS DEL TRIBUNAL

Presidente:	D.
Vocal	D.
Secretario:	D.
Suplente:	D.

Los miembros del tribunal acuerdan otorgar una calificación de:

Madrid, a de de 2023

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA DE TELECOMUNICACIÓN

TRABAJO DE FIN DE MÁSTER

Development of a fast image alignment algorithm for Cryo-Electron Microscopy

OIER LAUZIRIKA ZARRABEITIA

2023

Resumen

La Microscopía Electrónica Criogénica (CryoEM) ha revolucionado el campo de la biología estructural al permitir la visualización de estructuras macromoleculares en resoluciones sin precedentes. Sin embargo, el alineamiento de imágenes sigue siendo un desafío computacional en el procesamiento de imágenes de CryoEM, ya que se utiliza en numerosos pasos. El problema de alineamiento implica encontrar la traslación y rotación óptima de una imagen para que sea lo más similar posible a otra imagen dentro de un conjunto de referencias. Por consecuencia, la gran cantidad de comparaciones requeridas para resolver el problema lo convierte en un computacionalmente costoso, haciendo que se invierta una cantidad significativa del tiempo en este proceso. Asimismo, la calidad de los resultados finales está muy influenciado por la precisión de los alineamientos.

Este proyecto introduce un nuevo método de alineamiento con el objetivo de acelerar este proceso. Para ello, se utilizarán novedosas técnicas de compresión de vectores a la hora de almacenar y comparar imágenes. Estas técnicas permiten comparar imágenes de forma eficiente, facilitando la obtención de los parámetros óptimos de alineamiento.

Otro enfoque novedoso de este trabajo es el uso del consenso de alineamiento, donde múltiples ejecuciones no deterministas se combinan para mejorar la precisión del resultado. Esto mejora la fiabilidad de los alineamientos locales posteriores, ya que estas están muy sesgadas por la solución inicial.

Se han llevado a cabo pruebas con una amplia variedad de proteínas y parámetros, obteniendo así resultados empíricamente sólidos. Estos resultados demuestran que el algoritmo tiene capacidad de mejorar el rendimiento en el procesamiento de CryoEM mientras que se mantienen niveles de precisión aceptables, particularmente para alineamientos de baja resolución. Sin embargo, se han identificado limitaciones al aplicar el algoritmo en condiciones de alta resolución.

No obstante, el algoritmo proporciona una forma rápida y precisa de resolver las primeras iteraciones de un proceso de refinamiento, que normalmente implican costosos alineamientos globales. Por lo tanto, la eficacia del algoritmo puede contribuir significativamente a aumentar el rendimiento de estas primeras iteraciones, permitiendo a los investigadores obtener resultados mucho más rápidamente. Además, gracias al consenso, las iteraciones locales posteriores serán provistas con parámetros iniciales de alta calidad, disminuyendo la posibilidad de que estas caigan en mínimos locales y, por lo tanto, mejorando los resultados finales.

Palabras clave: Microscopía Electrónica Criogénica, Análisis de Partículas Aisladas, Alineamiento de imágenes, Búsqueda rápida de imágenes, Compresión de vectores

Abstract

Cryogenic Electron Microscopy (CryoEM) has revolutionised the field of structural biology by enabling the visualisation of macromolecular structures such as proteins at unprecedented resolutions. However, efficient and accurate image alignment remains a computational challenge in CryoEM image processing, as it is used in numerous steps. The alignment problem involves finding the optimal translational and rotational transformations that align an image to a set of reference images. Thus, the vast amount of image comparisons required to solve the problem renders it a computationally expensive process. For this reason, a significant amount of time spent in CryoEM image processing is dedicated to image alignment. Similarly, the quality of the final results is heavily influenced by the accuracy of the alignments.

This project introduces a new alignment method aiming perform alignments much faster than state-of-the-art techniques at little to no accuracy degradation. To do so, novel vector compression techniques will be used when storing and comparing images. These techniques can be used to efficiently compute similarity measures between the images, facilitating the identification of the optimal alignment parameters. Moreover they require less memory footprint, allowing to store more images in memory.

Another novel approach of this work is the usage of alignment consensus, where multiple non-deterministic alignments are combined to enhance the accuracy of the result. This enhances the reliability of subsequent local alignments, as these are heavily biased by the initial solution.

The algorithm has been extensively tested with a wide variety of proteins and parameters, leading to empirically solid results. These results demonstrate the algorithm's ability to enhance throughput in CryoEM image processing while maintaining acceptable accuracy levels, particularly for low resolution alignments. However, limitations in accuracy were observed when applying the algorithm to higher resolution targets.

Nevertheless, it provides a fast and accurate way to solve the first iterations of a refinement process, which typically involve expensive global alignments. Therefore, the algorithm's performance can significantly contribute to increase the throughput in these first few iterations of a refinement, allowing researchers to obtain preliminary results much faster. Additionally, the consensus provides subsequent local iterations with high quality data, diminishing the chances of falling into local minimas and thus improving the final results.

Keywords: Cryo Electron Microscopy, Single Particle Analysis, Image alignment, Fast image search, Vector compression

Contents

List of Figures

List of Tables

Glossary

AGNW Addititive Gaussian White Noise. 52

ART Algebraic Reconstruction Technique. 21, 22

BCU Biocomputing Unit. 2, 16

BLAS Basic Linear Algebra Subprograms. 28

CLI Command Line Inteface. 28

CNB Centro Nacional de Biotecnología. 2, 16, 93

CPU Central Processing Unit. 33, 35

CryoEM Cryogenic Electron Microscopy. vii, 1, 2, 5, 15, 17, 23, 27–29, 35, 44, 51–54, 56, 68, 79–83, 91, 92

CryoET Cryogenic Electron Tomography. 16, 83

CSIC Consejo Superior de Investigaciones Científicas. 2, 16, 93

CTF Contrast Transfer Function. 8, 9, 16, 17, 20, 29–31, 52, 53, 57, 58, 77, 81, 82

CUDA Compute Unified Device Architecture. 16

DCT Discrete Cosine Transform. 32

DFT Discrete Fourier Transform. 32

EBI European Bioinformatics Institute. 52

EM Electron Microscope. 5

EMBL European Molecular Biology Laboratory. 51

EMPIAR Electron Microscopy Public Image Archive. 51, 52, 54

FFT Fast Fourier Transform. 13, 19, 28

FOSS Free and Open Source Software. 28, 92

FSC Fourier Shell Correlation. 24, 25, 49

FT Fourier Transform. 12, 13, 19, 22

GPU Graphics Processing Unit. 15, 16, 33, 35, 49

GUI Graphical User Interface. 27, 28

I/O Input/Output. 28

IFT Inverse Fourier Transform. 13, 19, 22

IVF Inverted File. 35, 40, 41, 62, 63, 67

kNN k Nearest Neighbours. 20, 33, 37, 39

LS Least Squares. 22

LTI Linear Time Invariant. 30, 33

ML Machine Learining. 8, 9, 11, 33

MLE Maximum Likelihood Estimation. 48, 81

MRC Medical Research Council. 28

MSE Mean Square Error. 30

PCA Principal Component Analysis. 35, 37, 40, 62, 63, 67

PDB Protein Data Bank. 52

PQ Product Quantisation. 35, 37–41, 62, 63, 67

PSD Power Spectral Density. 8, 18, 21, 31

ROI Region of Interest. 46, 77

SGD Stochastic Gradient Descent. 10, 11

SNR Signal to Noise Ratio. 5, 8, 9, 18, 21, 25, 52, 54, 57, 81, 83

SPA Single Particle Analysis. 1–3, 5, 8, 13, 15, 16, 29, 80, 82

SSNR Spectral Signal to Noise Ratio. 25, 30, 31, 81

STA Sub-Tomogram Averaging. 83

SVD Singular Value Decomposition. 37

TEM Transmission Electron Microscope. 1, 5, 8, 18, 52

VAT Value Added Tax. 93

1. Introduction and objectives

CryoEM is an image acquisition technique that uses Transmission Electron Microscopes (TEMs) to examine a frozen sample. Unlike traditional optical microscopes, TEMs use an electron beam instead of light, which allows them to capture images at much higher resolution. As a consequence, it has become a very popular technique for collecting images of biological molecules, such as proteins[1]. These images can be used to elucidate the 3D structure of the molecule under study.

However, TEMs require very specific conditions in order to work, such as near perfect vacuum and high-energy electrons. Therefore, they are unsuitable for biological samples, as these are too fragile to endure in such conditions. Here is where the cryogenic part comes into place. In order to retain the sample intact and in place, a thin film of ice is used. The sample is cooled down very rapidly, so that the water has no time to form an ice lattice, avoiding the diffraction of the electron beam. This technique was awarded with the 2017 Nobel Prize in Chemistry[1][2]. Closely related to this, in 1982 Aaron Klug was also awarded with the Nobel Prize in Chemistry for his development of crystallographic methods to elucidate the 3D structure of proteins[3].

Usually, the sample is prepared on a copper or gold grid, which may hold thousands of specimens under study, each of them with a random orientation. Each of these specimens is known as "particle". Assuming that all particles belong to the same structure, their 2D projections can be used to mathematically infer the 3D structure of the specimen[4]. Single Particle Analysis (SPA) is a family of image acquisition and processing techniques that enables such a task.

At the beginning of the SPA image processing pipeline, a large quantity of noisy data is provided, from which little to no parameters are known. Therefore, all of the parameters needed for reconstruction must be estimated from the data. Many of these parameter estimations are conducted by assigning each experimental image to a reference image from which the parameters to be estimated are known. Thus, the parameters can be inherited from the assigned reference. This process is known as image alignment and it is one of the most frequent problems on a typical CryoEM image processing pipeline.

The aim of this project is to develop a fast computer program to align particles. The key innovation of this project is the usage of state-of-the-art vector search databases, which employ vector compression to store and compare vectors.

The project has been carried out at the Biocomputing Unit (BCU) research group located at Centro Nacional de Biotecnología (CNB)-Consejo Superior de Investigaciones Científicas (CSIC) facilities. This research group develops two software suites related to CryoEM, Xmipp and Scipion. The former one implements image processing algorithms, whilst the later one provides a framework to easily interoperate between state-of-the-art image processing suites. Consequently, the software developed in this project will be implemented inside Xmipp and it will be integrated into Scipion.

1.1 Objectives

The main objective of this thesis is to develop a fast and accurate image alignment algorithm for CryoEM. To be more precise, the algorithm will focus on performing 3D alignments of particles, this is, it will be used to deduce their orientation. Nevertheless, it will leave room for its application in other image alignment problems in CryoEM.

In general, the image alignment problem involves finding the best match for an image across a large set of images, also considering their rotations and translations. As a consequence, this process involves comparing many image pairs, making it computationally expensive. In addition, it is a recurrent problem in the context of SPA and other CryoEM image processing techniques.

As a result, current SPA image processing workflows allocate significant time and computational resources to execute alignment algorithms. By reducing the time required for image alignment, the overall computation time needed to solve a protein structure is greatly diminished. This enhancement in performance brings forth several advantages. Firstly, it enables more efficient utilisation of the available resources, which are often limited due to the expenses associated with high-end workstations and servers. Additionally, it allows for higher throughput, facilitating to reach further results streaming. Last but not least, biologist are able to draw conclusions and iterate must faster, accelerating the development of drugs and vaccines.

Similarly, this widespread usage of alignment processes in CryoEM also implies that the quality of the final results is heavily influenced by their accuracy. Therefore, it is important that the compromises taken to enhance the performance of the algorithm should not negatively impact its accuracy.

1.2 Structure of the document

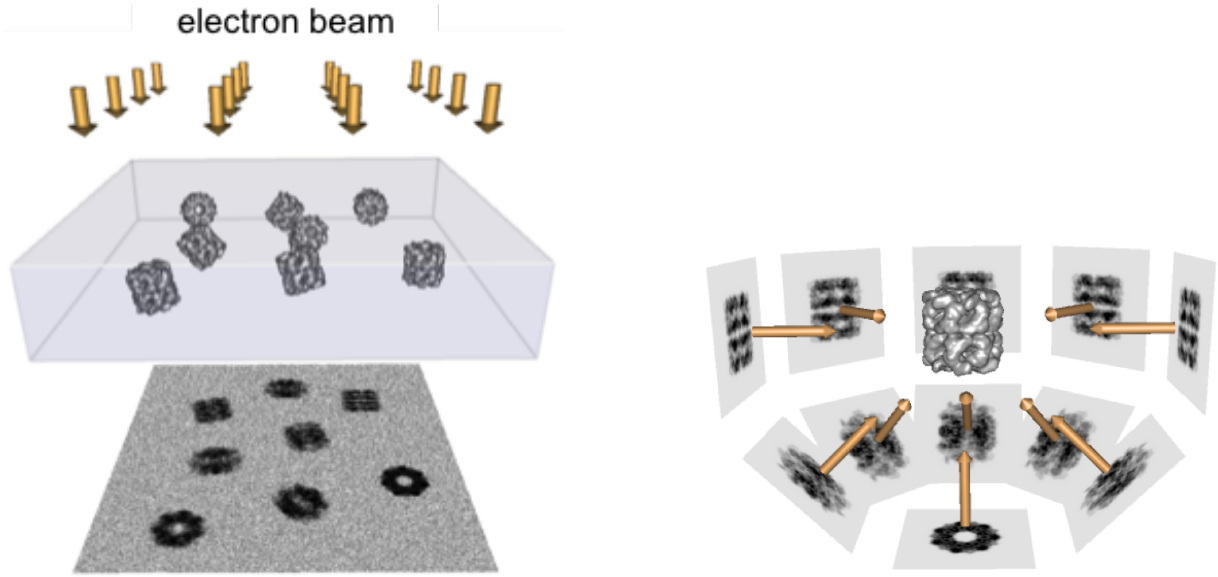
The document follows a logical and comprehensive structure to describe the development of a new image alignment algorithm. It begins in Chapter 1 with an introductory section that outlines the objectives of the study, providing a clear understanding of the project's purpose and extent. The next chapter introduces the field of SPA, which is the main use case of this algorithm. This chapter will focus on the image processing workflow that enables the discovery of the 3D structure of macromolecules. Then, the state of the art of the alignment algorithms will be described in Chapter 3, providing a thorough review of the current methodologies employed. The Implementation Chapter provides a technical description of the approaches taken to implement our own image alignment algorithm, covering the compromises and benefits associated to each design decision. This algorithm will be extensively tested, and the results will be showcased in Chapter 5. The Conclusions Chapter summarises the key outcomes and evaluates the effectiveness and limitations of the techniques employed. Finally, the document concludes with a future work chapter that identifies potential areas for further research.

2. Single Particle Analysis

SPA refers to a CryoEM technique that allows to obtain models of proteins at almost atomic resolution. Although it has been around for decades, recent technological leaps have led to an increase in interest from users and researchers. This technique involves everything from the sample preparation to the final image processing, including the image acquisition at the microscope[5]. However, this chapter will focus on explaining the image processing part of the workflow.

The essence of SPA lies on rapidly freezing thousands of specimens in a thin film of ice. In this way, each specimen will be held in place with the random orientation it had before it was frozen. At this point, the sample is scanned by a TEM, obtaining 2D projections of the specimens. These projections can be thought of as a shadow of the electron density of the sample. Using advanced image processing techniques, this collection of projections can be used to reconstruct the 3D electron density map of the specimen under study. Figure 2.1 provides an illustration of this process. Nevertheless, the reconstruction process involves several challenges, as the input images have very poor Signal to Noise Ratio (SNR) and other artefacts.

The studied sample is prepared on a copper of gold grid, which is inserted into the Electron Microscope (EM) chamber. Each of spot of the sample can only be exposed to the electron beam for a limited amount of time before degradation occurs. Recent leaps in sensor technology have sped up the required exposition time for the sensors, enabling them to capture multiple frames of the sample before degrading it. The set of frames captured from a given spot is known as movie. These movies serve as the starting point of the SPA image processing workflow. This workflow is summarised in the Figure 2.2 and it will be detailed hereafter.



(a) Image acquisition (b) Reconstruction

Images obtained from: [6]

Figure 2.1: Image acquisition and structure reconstruction

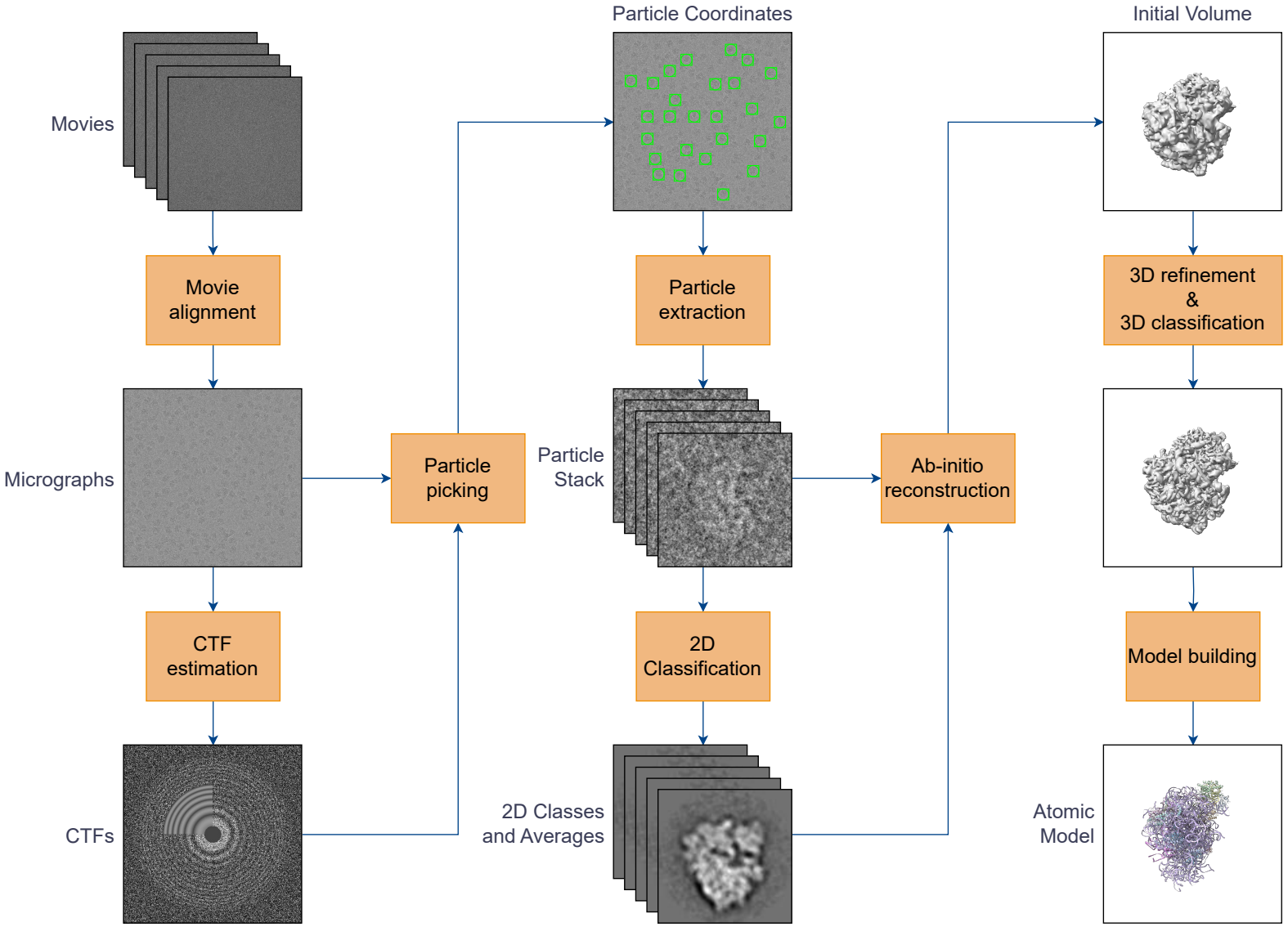


Figure 2.2: SPA workflow

2.1 SPA image processing steps

Movie alignment

In the movie alignment stage all the frames of a movie are averaged into a single image known as micrograph. This helps to increase the SNR, as the uncorrelated part of the noise tends to cancel out across images. Note that the noise induced by the vitreous ice is the same for all the frames of a given movie so it will not be removed when averaging frames.

The frames contained in a movie are in chronological order. As a result, the last images have a higher electron dose than the first ones, which translates into a more severe deterioration of their atomic structure Moreover, this deterioration has a higher influence in the higher frequencies of the image. These facts need to be taken into account when combining all the images, in such a way that the high frequencies of the last images have less weight[4].

Additionally, the electron beam positioning system drifts between frames and the sample tends to bend, which tends to produce optical flow between frames. Consequently, the frames are not aligned to one another. This needs to be fixed before attempting to average the frames, as otherwise the resulting micrograph would loose resolution.

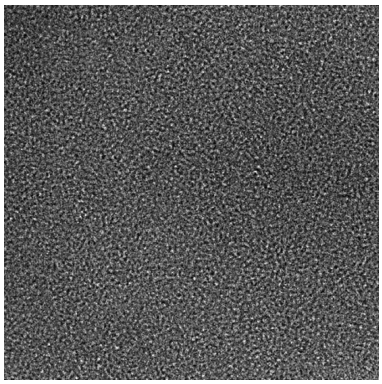
CTF estimation

TEMs do not have a planar frequency response. Instead, they "colour" the images with a characteristic Contrast Transfer Function (CTF) known as Thon rings. This transfer function has a sinusoidal appearance, with decreasing periodicity and a overall tendency to attenuate higher frequencies[4]. In addition, the rings may have elliptical shape, being wider in some axis. This is known as astigmatism. A example of a TEM CTF is shown in the Figure 2.3.

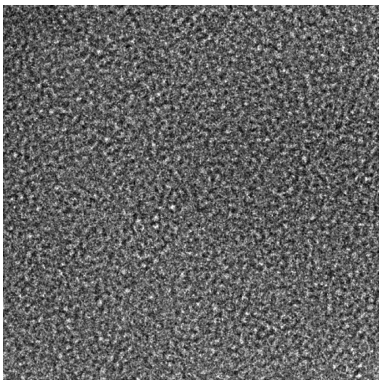
This CTF is different for each micrograph and it needs to be known by later steps. Moreover, it can be used to assess the quality of the micrographs[4]. The characterisation of the CTF is accomplished by calculating the Power Spectral Density (PSD) of the micrograph and fitting a template onto it.

Particle picking

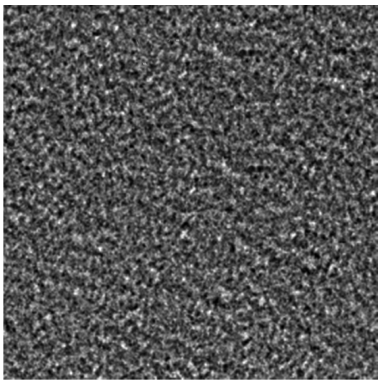
In the context of SPA, the term particle refers to the individual projections of the specimen under study. As stated earlier, a micrograph may contain many particles. Particle picking consists in pin-pointing individual particles in a micrograph. This enables extracting them to individual images in order to continue with the processing. This used to be a manual task for biologists, but recent leaps in Machine Learining (ML) have enabled the possibility of



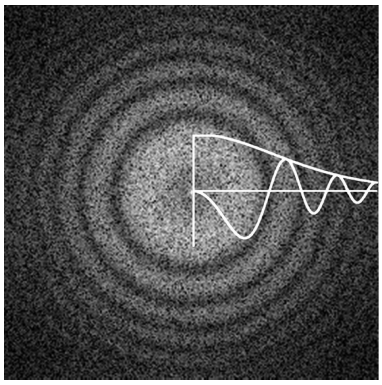
(a) Micrograph with 0.5µm defocus



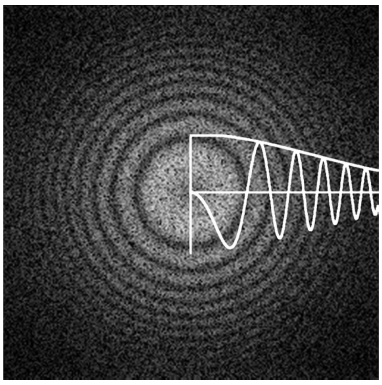
(b) Micrograph with 1.0µm defocus



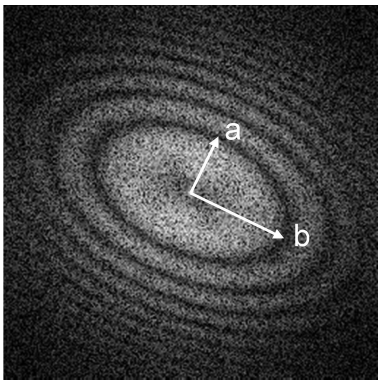
(c) Micrograph with astigmatism



(d) Micrograph PSD with 0.5µm defocus



(e) Micrograph PSD with 1.0µm defocus



(f) Micrograph PSD with astigmatism

Images obtained from: [4]

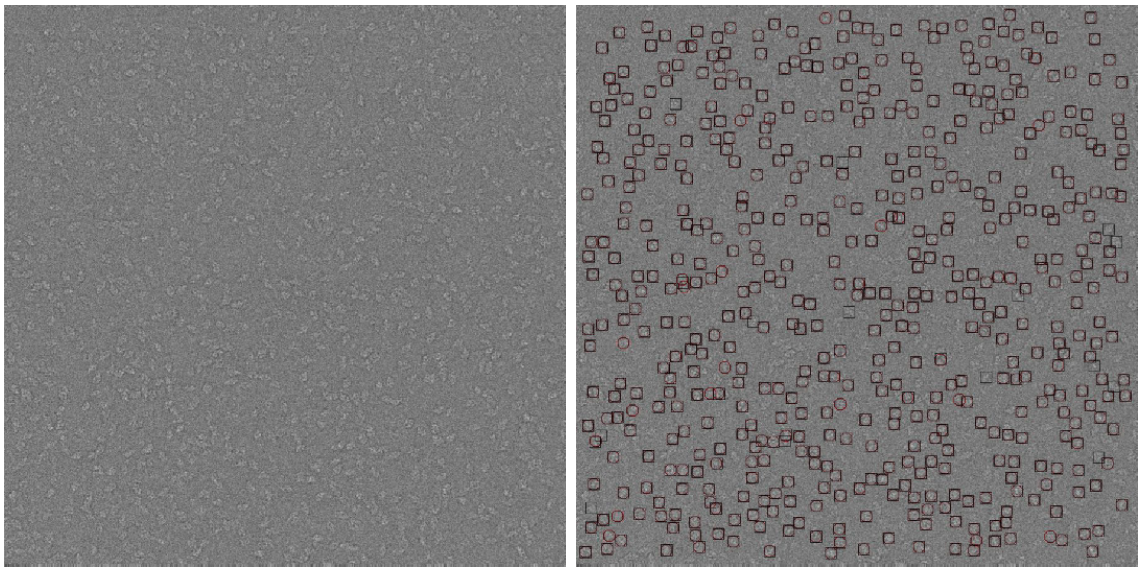
Figure 2.3: CTF examples

using supervised ML algorithms to automate this process. An example of a picking is shown in the Figure 2.4

2D Classification

2D classification consists in comparing particles to one another and clustering similar ones. These comparisons take into consideration in-plane transforms (rotations and shifts) of the particles. Therefore, clusters are invariant to translation and rotation. These clusters are averaged so that the highly correlated parts of the particles remain intact, while uncorrelated parts -noise- are attenuated, potentially increasing the SNR. Moreover, as many micrographs with unique CTFs are used, the missing information in the zeros of the CTF tends to cancel out. A example of this process is illustrated in the Figure 2.5.

These 2D classes have many applications. For instance, their averages can be used as a feedback to re-enforce the picking algorithm. Additionally, the lack of clusters can be used as an evidence of preferential orientations of the specimen. Similarly, poorly detailed clusters



(a) Original micrograph (b) Picked micrograph

Figure 2.4: Example of a picked micrograph

may indicate that particles belonging to them are invalid. Last but not least, these 2D classes may be used as input for downstream steps.

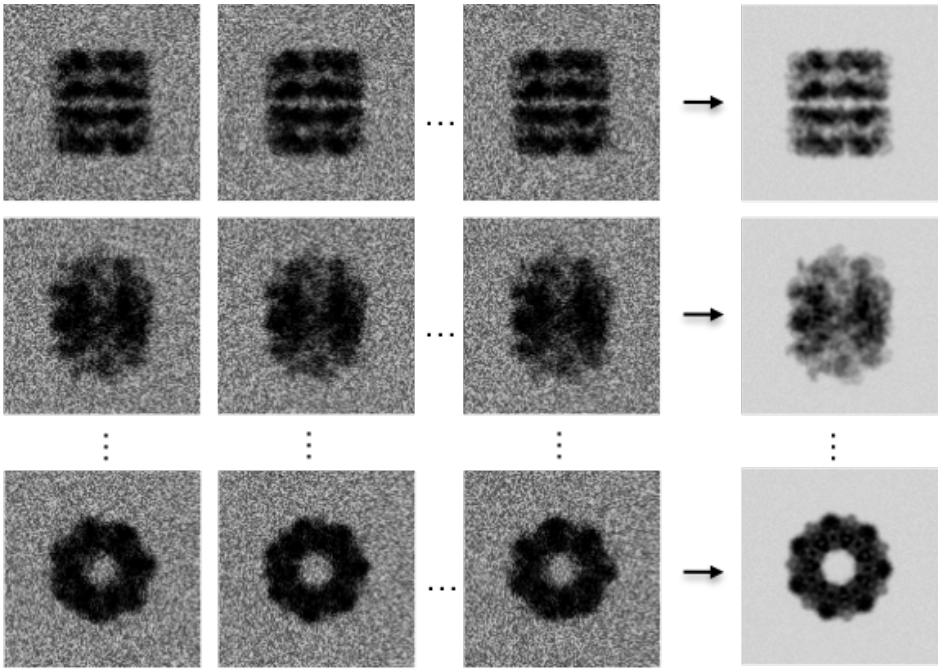


Image obtained from: [6]

Figure 2.5: Example of 2D classification

Ab-initio map reconstruction

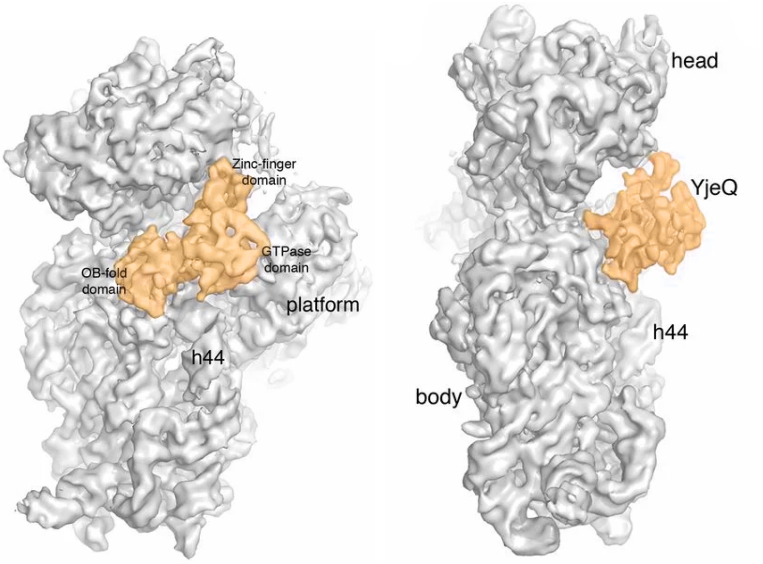
3D reconstruction is usually a Stochastic Gradient Descent (SGD) algorithm which iteratively improves a 3D electron density map of the protein under study. Therefore, choosing

a good starting point is important to improve the performance of the SGD algorithm and avoid local minimums as much as possible. This starting point is known as the initial model. As the gradient descent starts at this volume, the final result will be heavily biased by it[7].

The problem of obtaining a initial volume lies in deducing a 3D volume from a set of 2D projections that were done across unknown directions. There is a large set of approaches to address this problem. Some approaches perform a random angular assignments and then start the gradient descent from it. Some other algorithms rely on correlating a vast amount of random reconstructions[8]. Finally, there are some novel approaches that make use of unsupervised ML methods to learn a map from the particles[9].

3D Classification

Until this point we have assumed that all particles belong to the same structure. However, this is not true in many cases, as proteins may be flexible or they might have a ligand attached to them. Figure 2.6 exhibits a protein with conformational heterogeneity due to a drug binding. If this specimen was to be captured, some particles would contain the part highlighted in orange and some others would not.



(a) Front view (b) Side view

Images obtained from: [10]

Figure 2.6: 30S ribosome with a binding

3D classification consists in clustering particles based on the structure they belong to. Obviously, when the input data has no conformational heterogeneity, this step is skipped.

Usually, the differences between the considered variations of the structure are very subtle, so this is not an algorithmically easy task. Some software packages perform this task in the refinement step, in a process known as multi-reference refinement.

Refinement

The refinement step is used to obtain a high resolution 3D electron density map of the protein under study. As stated earlier, sometimes more than one map may be desired.

Most of the state-of-the art packages perform the following refinement cycle repeatedly. In essence, the algorithm tries to maximise the compatibility between the reconstructed volume and the experimental data. For that, it attempts to reproduce the experimental data from the reconstructed volume. This cycle is displayed in the Figure 2.7.

1. Project the current volume(s) from different angles to obtain a projection gallery.
2. For each experimental image find the most similar image in the gallery and assign its projection angle. Note that in-plane transformations (rotations and translations) need to be taken into account. Most of the existing solutions differ in this step, as many similarity metrics and exploration patterns can be used.
3. Reconstruct the volume(s) with the angular assigned experimental images.
4. Repeat steps 1 to 3 using the newly obtained volume. The algorithm should converge to a local minima[7]. When the loop stops producing significant changes or a desired resolution is achieved, the cycle should be stopped.

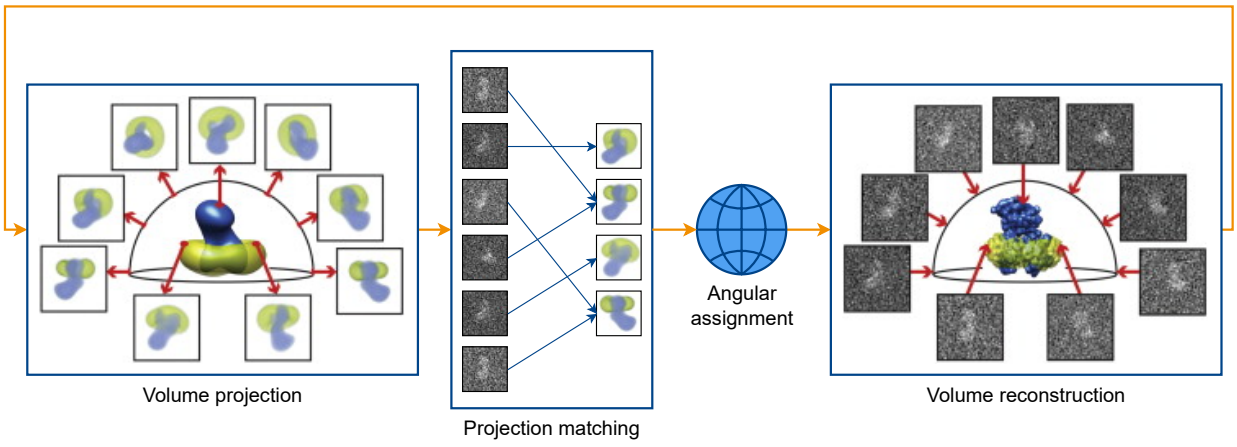


Diagram figures from: [11]

Figure 2.7: Typical refinement cycle

Nowadays, most implementations make use of the Fourier Central Slice theorem to perform steps 1 and 3. This theorem states that projecting a N-dimensional function to $N-1$ dimensions and then taking its Fourier Transform (FT) is equivalent to computing the N-dimensional FT and then extracting the central hyperplane normal to the projection direction. This equivalence is shown in the Figure 2.8. Most reconstruction algorithms leverage

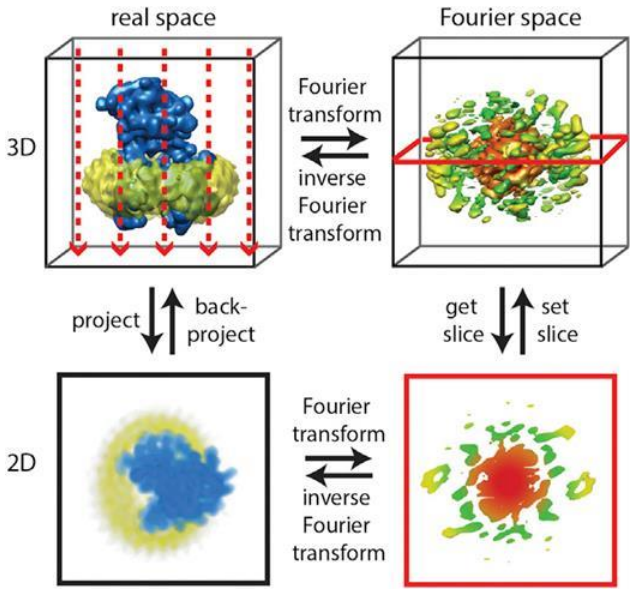


Image obtained from: [11]

Figure 2.8: Fourier Slice Theorem illustration for 3D

this fact by filling 3D Fourier space with appropriately oriented 2D FTs of the particles and then taking its Inverse Fourier Transform (IFT).

In essence, using the Figure 2.8 as an example, our goal is to obtain the 3D volume in real space (top left image), but the microscope provides a collection of 2D projections of it (lower left image). Although the direct approach would be the back-projection, following the Fourier path leads to faster results. This speed improvement is largely due to the Fast Fourier Transform (FFT) algorithm.

Model building

The final step in SPA consists in deducing the atomic structure of the protein under study. This is a labour intensive task where a biologist needs to fit an amino acid sequence into the newly reconstructed 3D electron density map. A example of this process is displayed in the Figure 2.9

2.2 Conclusions

The complexity of the SPA image processing can not be overstated. The starting point is a vast amount of data representing thousands of random projections of the specimen under study. This data is heavily contaminated with various sources of noise and other artefacts. Moreover, most of the parameters, including the projection directions, are unknown. Many times we can not even affirm that all projections belong to the same structure. All these

(a) 3D electron density map (b) Solved structure

Images obtained from: [12]

Figure 2.9: Example of model building

unknowns need to be estimated from the data before attempting to perform a reconstruction. At the end, the atomic model of the protein can be deduced from this reconstruction.

However, the effort required to obtain these atomic models is highly justified. These models give researchers a lot of knowledge and power to develop new drugs and vaccines.

3. State of the art

3.1 SPA image processing software packages

SPA has significantly increased its popularity in the last decades. As a consequence, several image processing packages have arisen. All of them chase similar ambitions: Obtain accurate high resolution maps in the least amount of time possible. Most of the state-of-the-art CryoEM image processing packages have converged into the same image processing pipeline. This pipeline follows a conventional structure, although it is somewhat malleable. The difference between packages lies on the algorithmic approach they use to accomplish individual tasks of the pipeline. Usually, each package is only proficient in a handful of steps. In fact, some packages do not implement the whole pipeline and rely on others to be able to process from beginning to end.

Traditional software packages in the context of SPA are Spider[13], Imagic, Eman[14], Cistem[15], Relion[16] and Xmipp[17]. In 2016 the introduction of Cryosparc[18] was disruptive due to its significant performance improvements. Closely related to this, Scipion[19] is a platform that enables end users to easily interoperate between different image processing packages.

One of the recent leaps in the context of CryoEM has been the usage of hardware accelerators such as Graphics Processing Units (GPUs) to significantly reduce processing times. Although GPUs are only well suited for highly parallelizable operations, in those cases, the computation time is reduced by several orders of magnitude. Indeed, this has been one of the main factors leading to the recent growth of CryoEM.

Currently, all of the image processing suites are steering towards streaming image processing. This means that data is processed at the same time that it is acquired in the microscope. This allows to adjust acquisition parameters in real-time, optimising resource utilisation and improving the quality of the final results.

Xmipp

Xmipp is a image processing package aimed at obtaining 3D electron density maps of biological samples. It is developed at the BCU group at the CNB-CSIC research centre. It was introduced at 1996, although it has suffered many major overhauls since then. Even though its primary focus is on SPA, it has diversified to many other microscopy techniques such as Cryogenic Electron Tomography (CryoET)[17].

Currently, it is on its third major version, which gets a minor version bump-up every 4 months. It has been mostly implemented in the C++ programming language, but it includes parts written in Python and Java. Xmipp offers methods for all steps in the SPA image processing pipeline, being proficient at movie alignment (Flexalign)[20], CTF estimation, particle picking, 2D classification and 3D refinement.

Xmipp developers have ported many crucial programs to run on GPU accelerators, significantly decreasing overall computation times. This has been achieved using Compute Unified Device Architecture (CUDA), a GPU computing platform commercialised by NVIDIA Corporation.

Scipion

As mentioned earlier, Scipion does not implement any image processing algorithms. Instead, it provides a common scaffolding to integrate image processing packages though plugins. This enables end users to easily build SPA image processing workflows using the strengths of each processing package. Moreover, it provides methods to consensuate the outputs of multiple programs, further increasing the quality of the results. In fact, the benefits of Scipion have been extended to other domains such as Virtual Drug Screening[21] or CryoET[22].

In the context of SPA, all widespread image processing tools have been integrated into Scipion. As shown in the Figure 3.1, usage statistics prove that users do have different preferences for each step of the processing workflow. For instance, 3D classification is almost always done with Relion, whilst particle picking is primarily done though Xmipp. This manifests the need for such a software, as manually inter-operating between packages is a very time consuming and error prone process. At the same time, being locked-in with a particular package leads to suboptimal results, as that particular package may waver in some steps.

Scipion is highly modular, as it can be extended with plugins. These plugins are usually related to the integration of a image processing suite, such as Relion or Cryosparc. A plugin provides a set of protocols, which can can be seen as a “steps” in the image processing workflow. Then, the user can easily build its own workflow, freely choosing the procedure used for each stage. What is more, the user may repeat the same step using different protocols

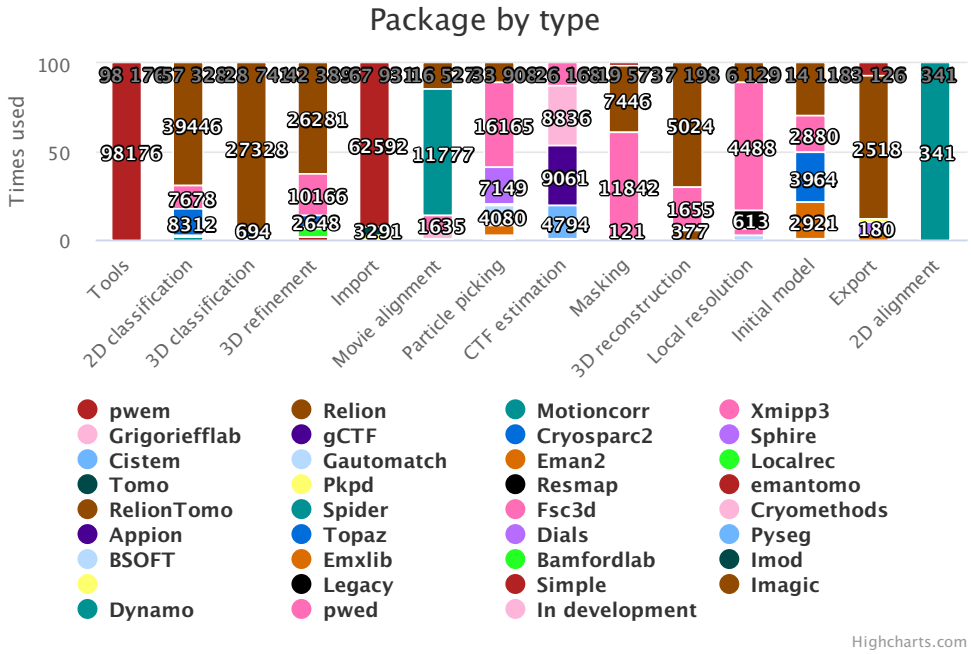


Figure 3.1: Scipion package usage statistics by type

and consensuate their outputs. Therefore, Scipion not only integrates alien algorithms, but it also provides some added value to the results.

Many of the current Scipion developments focus on implementing streaming workflows, where all the other benefits stated earlier still apply. Moreover, there is some innovation related to the automated control of the microscope from the image processing software. This control feedback loop enables microscope operation with little human intervention, significantly reducing costs.

3.2 Refinement algorithms

Most CryoEM image processing suites such as Relion[16], Cryosparc[18], Cistem[15] and Xmipp[23] implement a refinement step in which a 3D model is iteratively improved. This process should converge to a high resolution solution which is compatible with the provided images.

The input for this step is a large set of noisy images presumably containing the projection of a particle. A low-resolution estimation of the volume is also provided (initial volume)[24].

Input images are not clean, in fact, they have many artefacts. Firstly, the spatial frequency response of the microscope is not planar. This means that the acquired images have been filtered in frequency space with a transfer function known as CTF. This transfer function has been estimated in previous steps, so it does not need to be deduced (although it can be fine-tuned). Moreover, the particle is not perfectly centred in the image box. Last but not

least, the images contain a vast amount of noise. Indeed, the SNR is in the order of -10dB to -20dB[25]. This means that the noise has a greater contribution to the image than the specimen itself. The PSD of the noise roughly resembles to pink noise, this is, its PSD is inversely proportional to the frequency. However, the exact PSD of the noise is unknown and should also be estimated from data[26]. The main source for this noise is the amorphous ice crystal structure that holds the biological sample in place.

Before attempting to reconstruct the 3D structure of the specimen, the projection directions of each of the images need to be deduced in a process known as 3D particle alignment[24]. This is a computationally expensive task and much effort has been put into it to reduce the amount of time expended on this process.

The alignment process relies on projecting the current volume from multiple perspectives, somewhat mimicking the microscope's behaviour. Then, each of the experimental images is searched across all simulated projections (references), considering in-plane transformations (rotations and translations). The actual similarity metric used for matching varies across the existing solutions[24].

Once a best match has been found, the projection parameters of the selected reference image and its best transform can be assigned to the experimental one. This enables using the experimental images to reconstruct a new volume, potentially with a higher resolution. This last volume can be used as the initial volume for the next iteration. This cycle is illustrated in the Figure 2.7.

Projection gallery generation

The projection gallery represents a set of projections of the current volume from relevant directions. This collection of images is generated by projecting the initial volume in the same way that a TEM microscope would do. This enables performing comparisons between experimental and generated images.

In rough terms, TEMs fire a electron beam through the sample and capture the "shadow" of its Coulomb potential density[25]. In other words, areas where the beam encounters electrons will appear dim. This process is illustrated in the Figure 2.1a. Nevertheless images are usually complemented so that bright areas relate to the presence of matter.

This behaviour is mathematically described with the expression (3.1), which consists in computing the integral across a set of parallel lines normal to the projection plane. The volume is represented by the function $V(\tilde{\boldsymbol{r}}) : \mathbb{R}^3 \to \mathbb{R}$ and the projected image is represented by the function $I_{\tilde{A}}(\tilde{\boldsymbol{s}}) : \mathbb{R}^2 \to \mathbb{R}$. Both $\tilde{\boldsymbol{s}} = (s_x, s_y, 1)$ and $\tilde{\boldsymbol{r}} = (r_x, r_y, r_z, 1)$ are homogeneous coordinates[24]. Homogeneous coordinates allow to introduce a shift to the projection, so that the particle can be off-centred.

$$I_{\tilde{A}}(\tilde{\boldsymbol{s}}) = \int_{-\infty}^{\infty} V(\tilde{A}^{-1}\tilde{H}^T\tilde{\boldsymbol{s}})dt \tag{3.1}$$

where $\tilde{H}^T$ is the projection matrix defined as in (3.2). It can be deduced that for homogeneous 2D coordinates, this matrix will provide the integration variable t in the Z axis.

$$\tilde{H}^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & t \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow \tilde{H}^T\tilde{\boldsymbol{s}} = (s_x, s_y, t, 1) \tag{3.2}$$

$\tilde{A}$ encodes the projection direction using the 3D rotation matrix R and in-plane shift $\boldsymbol{\delta} = (\delta_x, \delta_y, 0)$ in an affine matrix defined in (3.3). Due to the fact that the rotation matrix R is orthonormal, its inverse matrix can be easily computed by transposing ($R^T = R^{-1}$).

$$\tilde{A} = \begin{pmatrix} R & \boldsymbol{\delta} \\ \mathbf{0}^T & 1 \end{pmatrix} \Leftrightarrow \tilde{A}^{-1} = \begin{pmatrix} R^T & -R^T\boldsymbol{\delta} \\ \mathbf{0}^T & 1 \end{pmatrix} \tag{3.3}$$

This projection operation can be significantly accelerated using the Fourier Central Slice Theorem. This theorem states that projecting a N-dimensional function to $N-1$ dimensions and then taking its FT is equivalent to computing the N-dimensional FT and then extracting the central hyperplane normal to the projection direction[27]. This equivalence is shown in the Figure 2.8. Therefore, a set of projections can be generated by extracting 2D planes from the 3D FT of the input volume and then computing their IFT.

Note that when using this approach, only one 3D FFT is computed, and then for each projection direction, a 2D central slice is extracted from it, which contains the 2D FT of the projected image. More often than not, the following steps are performed in Fourier space, so computing the IFT of the slices is not needed.

Using the previous notation, the projection operation in Fourier Space is expressed in (3.4). The volume and image functions now define Fourier space, so that they are complex functions $\hat{V}(\tilde{\boldsymbol{R}}) : \mathbb{R}^3 \to \mathbb{C}$ and $\hat{I}_{\tilde{A}}(\tilde{\boldsymbol{S}}) : \mathbb{R}^2 \to \mathbb{C}$. Moreover, the $\tilde{\boldsymbol{s}}$ and $\tilde{\boldsymbol{r}}$ spatial variables have been substituted by the frequency vectors $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{R}}$, respectively[24].

$$\hat{I}_{\tilde{A}}(\tilde{\boldsymbol{S}}) = e^{-j\langle R\boldsymbol{\delta}, \tilde{\boldsymbol{S}}\rangle}\hat{V}(\tilde{A}_F^{-1}\tilde{\boldsymbol{S}}) \tag{3.4}$$

where

$$\tilde{A}_F = R \Leftrightarrow \tilde{A}_F^{-1} = R^T \tag{3.5}$$

Projection matching and angular assignment

Each of the input particles needs to be searched across the projections generated in the prior section, also considering all possible in plane transformations (rotations and shifts).

This search can be either local or global. For global searches, the projection gallery is generated with a uniform spacing between projection angles and all their in-plane transforms are also generated at a regular interval. Then, for each experimental image all combinations are tested to find a best match. However, as the target resolution increases, the sampling rate of the parameters also needs to be increased. This makes global searches unfeasible beyond low-resolution targets. Therefore, for high resolution, local searches are used. This means that images are only sampled around a narrow range centred in their previous assignment, significantly increasing throughput[18][16]. However, local searches need to be applied with precaution, as they involve the risk of falling in local minimas.

When comparing reference images and experimental images, the CTF needs to be considered. Experimental images have been "coloured" with a characteristic transfer function induced by the microscope. As opposed to this, the reference gallery was generated artificially without considering any frequency response. Therefore, this transfer function needs to be addressed before attempting to compare images to one another. Most of the current implementations choose to filter the reference images with the experimental's estimated CTF. Note that the CTF may vary from particle to particle[24]. Therefore, for each particle, the CTF needs to be re-applied to the reference gallery.

In essence, this step can be seen as a k Nearest Neighbours (kNN)[23] problem with $k = 1$. This means that we have a large set of images composed of all the projections of the current volume and its in-plane transformations. For each experimental image we want to find the most similar one, this is, the image which minimises some distance metric.

The problem can be mathematically expressed with the expression (3.6) where C is the estimated CTF of the experimental image I_{exp}. $I_{\tilde{A}}$ is the volume projection defined in (3.1). C, and I_{exp} remain constant for a given search. Moreover, the volume to be projected is also constant throughout a search. Therefore, only the projection parameters $\boldsymbol{R}$ and $\boldsymbol{\delta}$ regarding $\tilde{A}$ need to be optimised.

$$\min_{\boldsymbol{R},\boldsymbol{\delta}} \text{dist}(CI_{\tilde{A}}, I_{exp}) \tag{3.6}$$

A distance function is used to evaluate similarity between pairs of reference and experimen-

tal images. As mentioned earlier, the election of this function may vary across different implementations. More often than not, this distances are calculated in Fourier space, as high frequencies contain little information due to low levels of SNR. This allows to compute distances in a reduced set of coefficients corresponding to low frequencies, allowing more efficient computations.

- **Euclidean**: The euclidean distance is defined as the square root of the sum of squared coefficients. The square root part can be ignored, as it is not necessary for comparisons. In (3.7) it is defined for complex numbers, necessary for comparing in Fourier space.

$$dist^2_{\mathrm{L2}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^H \cdot (\boldsymbol{x} - \boldsymbol{y}) \tag{3.7}$$

- **Weighted euclidean**: The weighted euclidean distance features a weight matrix to give more importance to some coefficients. In order to remain a weight function, this matrix needs to be positive semi-definite. In fact, it is usually a diagonal matrix, providing independent weights to each coefficient. In the case of Relion[16] and Cryosparc[18] these weights are derived from the Maximum Likelihoods Estimation. In essence, the weights correspond to the inverse of the noise variance, this is, the inverse of the noise power. Therefore, PSD of the noise needs to be known.

$$dist^2_{\mathrm{L2,W}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^H \cdot \boldsymbol{W} \cdot (\boldsymbol{x} - \boldsymbol{y}) \tag{3.8}$$

- **Pearson correlation**: Pearson correlation is not a distance metric but a similarity metric. Nevertheless it can be easily converted to a distance applying a monotonically decaying function such as $1 - x$.

$$\rho_{x,y} = \frac{(\boldsymbol{x} - \bar{\boldsymbol{x}})^T (\boldsymbol{y} - \bar{\boldsymbol{y}})}{\sqrt{(\boldsymbol{x} - \bar{\boldsymbol{x}})^T (\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot (\boldsymbol{y} - \bar{\boldsymbol{y}})^T (\boldsymbol{y} - \bar{\boldsymbol{y}})}} \tag{3.9}$$

Finally, once a best reference match is found, the projection parameters and in-plane transform of the reference are assigned to the experimental image, as presumably it has been captured in such orientation.

3D reconstruction

The 3D reconstruction step consists in building a 3D electron density map using the angular-assigned experimental particles. Originally, this was performed using back-projection algorithms. Later this was replaced by the Algebraic Reconstruction Technique (ART) method, which approaches the reconstruction problem as a Least Squares problem. Current solutions rely on the Fourier Central Slice theorem, which was previously stated.

ART reconstruction

ART is a iterative reconstruction method. It formulates the reconstruction problem as a linear equation $Ax = b$ where x is the vector encoding the reconstructed volume, b is the set of experimental images and A is derived from the projection directions. Then, the problem can be seen as a Least Squares (LS) problem[27] trying to solve for x. In practice, A is too large to be solved with the conventional LS equation. Therefore, a iterative gradient descent is used. As LS is a convex problem, the iterative method is guaranteed to find the global minima[28][24].

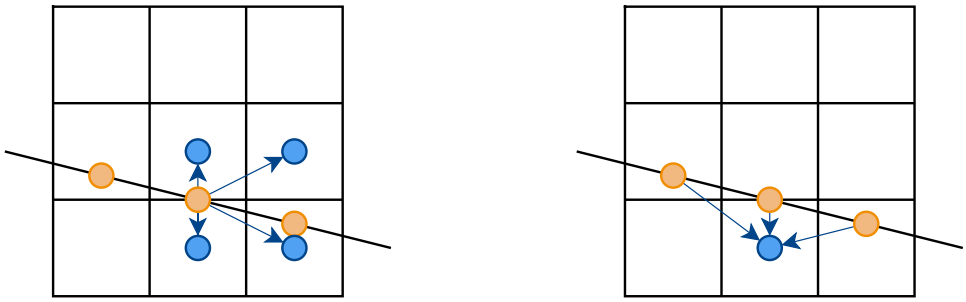
Fourier Reconstruction

Earlier, it has been stated that the Fourier transform of a 2D projection is equivalent to taking a 2D central slice from the 3D FT of the volume. Fourier reconstruction leverages this fact by filling the 3D Fourier space with the appropriately oriented 2D Fourier transforms of the experimental images. Assuming that projections from all possible directions have been provided, the whole 3D Fourier space is defined, which means that the volume can be unequivocally determined by computing the IFT. This principle is illustrated in the Figure 2.8[24].

In practice, there are some caveats related with the Fourier reconstruction regarding the discreetness of the samples. Low frequencies tend to get defined repeatedly when filling with slices, as these frequencies are close to the point where all planes intersect. In fact, the central point representing the lowest frequency, the DC component, is defined by all the slices. Contrary do this, some of the higher frequencies may not be defined because no slice goes though them. These problems are usually addressed using coefficient interpolation.

There are two approaches used when filling 3D Fourier space with 2D slices: Scatter and gather. In the first approach, each pixel of the 2D slices contributes to the voxel(s) that it “touches”. The second strategy approaches the problem inversely, for each voxel of the volume it determines from which pixels of the 2D slices it can consume from[29]. These strategies are illustrated in the Figure 3.2.

The main benefit of the scatter approach is that it can be executed in $O(n^2)$, as it loops over the Fourier coefficients of an image. Contrary to this, a naive implementation of the gather approach would loop over all Fourier coefficients of a volume, requiring $O(n^3)$ time[29]. However, there are heuristic methods that allow exploring only a subset of the 3D Fourier space, effectively achieving $O(n^2)$ time complexity for the gather approach. Moreover, this approach is more suitable for paralellisation, as it avoids race conditions. Indeed, this is the approach used in Xmipp[29].



(a) Scatter (b) Gather

Image based on: [29]

Figure 3.2: Gather and Scatter approaches

Multi-reference refinement

The refinement algorithms rely on the assumption that all input particles belong to the same structure. However, this is not always true, as the dataset might be heterogeneous. This heterogeneity can be either discrete or continuous.

The continuous heterogeneity relates to flexible macromolecules. These macromolecules have a certain amount of freedom to deform. Therefore, when projecting them, no single conformation can be attributed to the dataset. The study of these kind of proteins with CryoEM techniques is a novel field which is currently under study[30].

Contrary to this, discrete heterogeneity involves a discrete amount of conformations in the dataset. For instance, a biologist may want to test how a particular ligand binds to the protein. In that case, it is reasonable to consider that some of the input particles may come from a structure with the ligand attached, whilst some others won't have it.

The most common way to solve discrete heterogeneity is to generalise the refinement algorithm to N volumes. This is done by enabling multiple projection gallery inputs to the projection matching stage. Here, not only the projection parameters need to be considered, but also the reference volume. Then, each of the volumes can be reconstructed with the particles that were classified as such[31].

Although the working principle of multi-reference refinement is simple, there are many problems associated to it. Firstly, several input volumes need to be provided. These volumes have to be somewhat different so that the algorithm is able to converge to distinct volumes. Additionally, a problem known as attraction may appear. This problem is related to a class being able to gather increasingly more particles in each iteration, so that the other volumes are degraded and diverge. This last problem is usually solved by penalising the cost function for very populated classes[25].

3.3 Map quality metrics

Until this point, the map resolution term has been used repeatedly. However, the term has not been formally defined. The aim of this section is to provide some insights about map quality and resolution measurements.

Angular assignment error

Provided that the ground truth angular assignment of the experimental particles is known, the estimated angular assignment of can be compared against it to calculate the average deviation of the estimation. However, the ground truth angular assignment is not known, as after all, that is the purpose of angular assignment. Nevertheless, a secondary angular assignment can be provided by some other method. Therefore, this measurement is interesting for comparing and consensuating various algorithms. Additionally, this metric can be useful when assessing an algorithm with artificially generated experimental images (Phantoms), where ground truth projection parameters are known.

In the case of comparing the results of two independent refinement algorithms, it is probable that the reconstructed volumes may not have the same orientation, which will induce a systematic error on this metric. To avoid this error, the reconstructed volumes should be aligned to one another before comparing angles.

A major drawback of this method is that it may not reflect well the error produced when reconstructing the volume. For instance, when aligning a pseudo-symmetric protein, there is a set of views that are extremely similar. If an incorrect view is chosen for a particle, its angular error will be high. However it will not introduce a large error into the reconstruction.

Fourier Shell Correlation

When there is no other estimation about the angular assignment, the prior method can not be used. Moreover, the angular error does not explicitly provide any information about the quality of the reconstructed volume. The Fourier Shell Correlation (FSC) tries to solve these issues by measuring the resolution of the reconstructed volume[32].

The FSC is calculated by splitting the angular assigned particle set into two equally sized random subsets and generating a volume from each of the subsets. Then spherical surfaces (shells) are extracted from the Fourier transforms of the volumes, each one of them representing a certain frequency band. Comparing all shell pairs with the correlation metric, a correlation value can be assigned to each frequency band. This correlation in function of the frequency is known as the FSC, which is expressed in the equation (3.10) [33][32].

$$FSC(\omega) = \frac{\sum_{\omega_i \in \omega} F_1(\omega_i) \cdot F_2(\omega_i)^*}{\sqrt{\sum_{\omega_i \in \omega} |F_1(\omega_i)|^2 \cdot \sum_{\omega_i \in \omega} |F_2(\omega_i)|^2}} \tag{3.10}$$

Typically, the FSC tends to have a low-pass behaviour, as the SNR is worse at high frequency. This empirical fact is shown in the Figure 3.3. In fact, There is a direct relation between the FSC and the Spectral Signal to Noise Ratio (SSNR):

$$SSNR(\omega) = \frac{FSC(\omega)}{1 - FSC(\omega)} \Leftrightarrow FSC(\omega) = \frac{SSNR(\omega)}{1 + SSNR(\omega)} \tag{3.11}$$

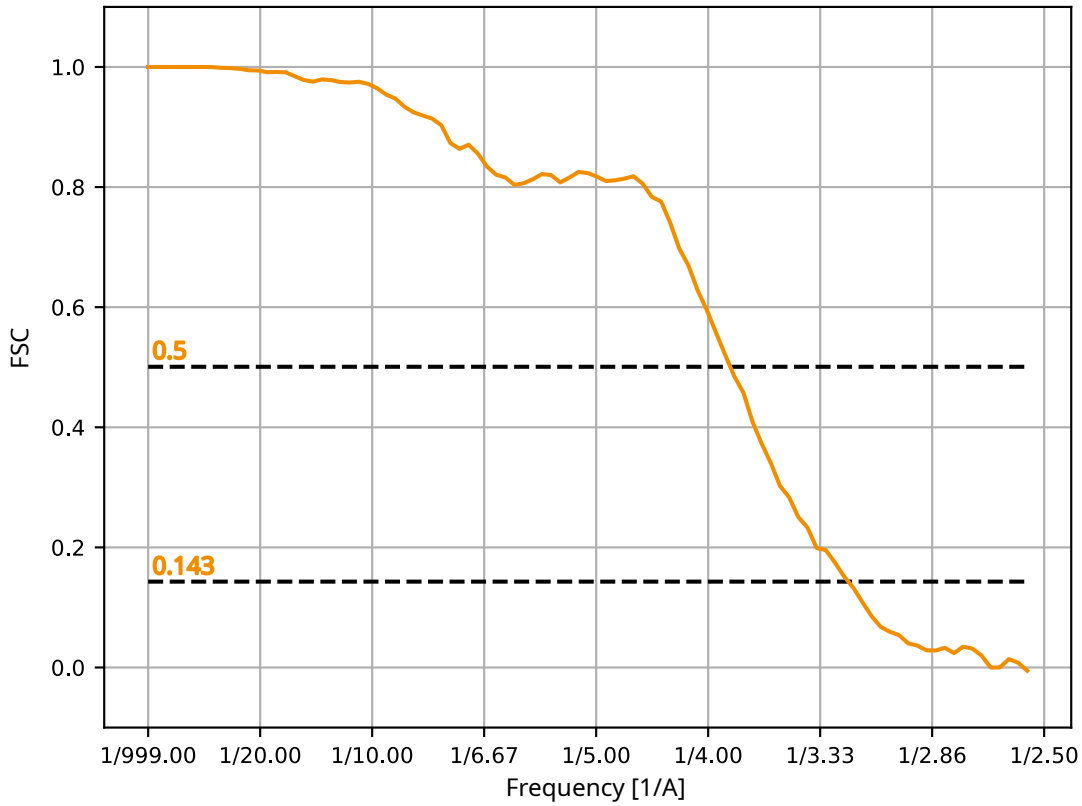


Figure 3.3: Example of a FSC function

Due to this empirical low-pass tendency, it is useful to establish a threshold. In this way, the resolution of a map can be determined as the first value where this FSC threshold is crossed. This provides a numerical value that quantitatively defines the quality of the map. The FSC threshold value election is source for a long lived discussion in academia. Some experts argue that this threshold should be 0.5, in reference to the cutoff frequency in the context of signal processing. Some other experts prefer to use 0.143. This value reflects better the resolution of the map that would be obtained when considering the whole particle

set (instead of halves)[34]. As a consequence of this debate, the threshold value used for determining the resolution of the map needs to be provided along the actual measurement.

4. Implementation

The main focus of this project is the development of a fast and reliable image alignment algorithm for CryoEM. Particle alignment is a frequent problem in the context of CryoEM, but solely addressing the alignment problem does not provide a solution to any practical objective. Consequently, we have chosen to utilise the 3D refinement problem as a platform for evaluating the effectiveness of the alignment algorithm. In this chapter we will describe both the image alignment algorithm itself and its integration in a 3D refinement cycle.

4.1 Architecture

This project has been developed in the context of Xmipp and Scipion framework. The image processing algorithms have been implemented inside Xmipp, whilst the refinement logic has been implemented in the Xmipp's Scipion plugin.

The image processing part implemented in Xmipp has been named as *swiftalign*, the union of the words *swift* and *align*. This name precisely describes the purpose of these programs: Fast image alignment. The *swiftalign* framework implements two programs: `swiftalign_train` and `swiftalign_query`.

These two programs, along other Xmipp programs, will be invoked from the Scipion protocol named as *swiftres*. This protocol is used to implement the refinement logic. This means that it will be responsible of orchestrating calls to the image processing algorithms and provide an easy to use Graphical User Interface (GUI) for the end user.

swiftalign

Swiftalign is a framework of image processing programs that is part of the Xmipp image processing suite. These programs specialise in image alignment, this is, matching images considering all their in-plane transformations. Although several utility programs have been implemented during the development of this project, the most prominent ones are `swiftalign_train` and `swiftalign_query`. All the framework has been implemented in

Python and it uses PyTorch library to accomplish computations such as common Basic Linear Algebra Subprograms (BLAS) routines and FFTs. This library is Free and Open Source Software (FOSS) and it is currently being developed by Facebook's AI Research lab.

One of the key innovations of the project is the usage of state-of-the-art vector databases. These databases are implemented inside the FAISS library, which is also developed by Facebook. Therefore, a good integration between the former libraries is expected.

Additionally, some file Input/Output (I/O) libraries are used to be able to read and write common file formats used in the context of CryoEM. This includes the `mrcfile` library used for reading images in `mrc` format and `starfile` library used for reading metadata files stored in `star` format. Both of these libraries are developed by the Medical Research Council (MRC).

Many vector databases need to be trained before being populated. This training needs to be performed with data that resembles the actual data that is going to be used, so that the database can be structured optimally. This task is responsibility of `swiftalign_train` which in rough terms performs a data augmentation of reference images to train a FAISS database.

Later this trained database is populated by `swiftalign_query` using all possible in plane transformations of the reference gallery. Once populated, it is used to query experimental images and assign the estimated alignment parameters. Hence, the actual particle alignment will be carried out in this program.

swiftres

Both of the previous programs have been implemented as a Command Line Inteface (CLI) utility. This means that they can be invoked by the user from a command line prompt. However, the usage of parameters and invocation is not trivial. Therefore, a Scipion protocol was implemented that wraps these calls to perform a 3D refinement. This protocol, not only provides logic for determining invocation parameters but it also offers an easy to use GUI. This protocol is named as *swiftres* and it will be introduced in this section.

The GUI of Scipion protocols is usually implemented as a form which lets the user choose the parameters for execution. These parameters may include the output of a previous step, a numeric value, toggle switch... When the appropriate parameters are selected, the protocol can be executed or saved for later use. In this case, the basic parameters to run *swiftres* include the input particles, the initial volume(s), the resolution of the initial volume and the symmetry associated to the protein under study. A screenshot of this is displayed in Figure 4.1.

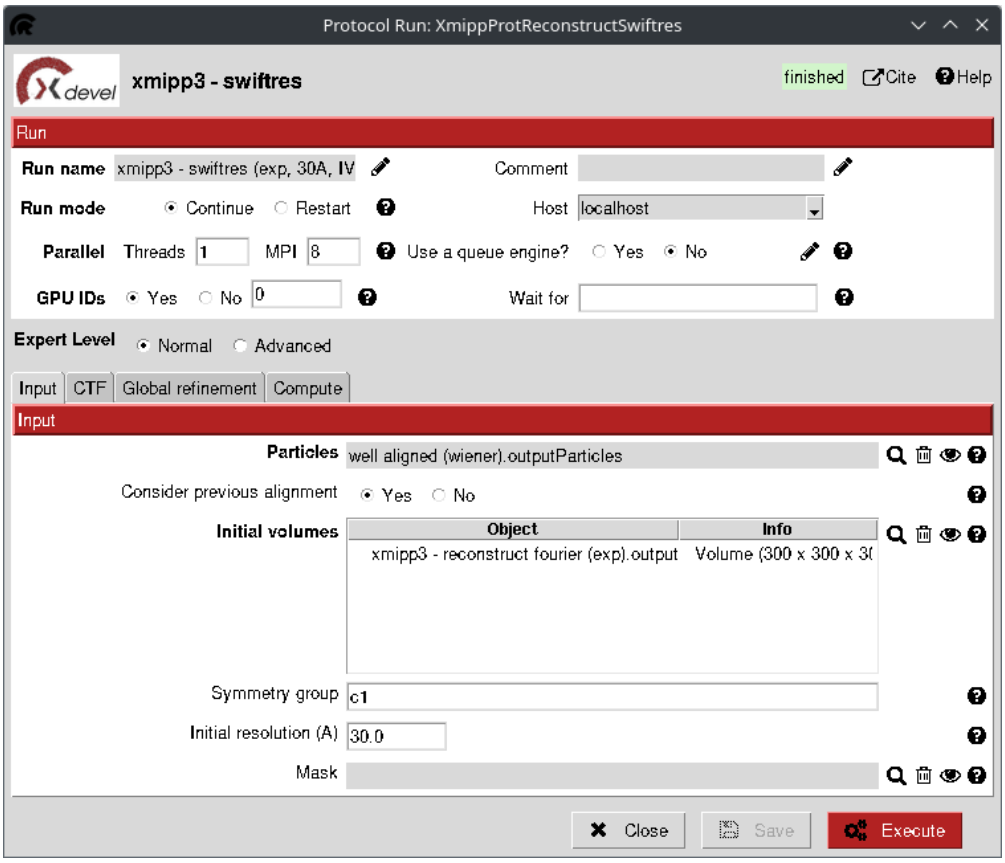


Figure 4.1: Screenshot of the user interface for running *swiftres* protocol in Scipion

4.2 Fast image alignment

Image alignment is a core problem in the context of CryoEM image processing. Several key steps involved in the SPA image processing pipeline involve performing an image alignment. Algorithms such as 2D classification, ab-initio volume reconstruction, 3D refinement and 3D classification are examples that fall in this category.

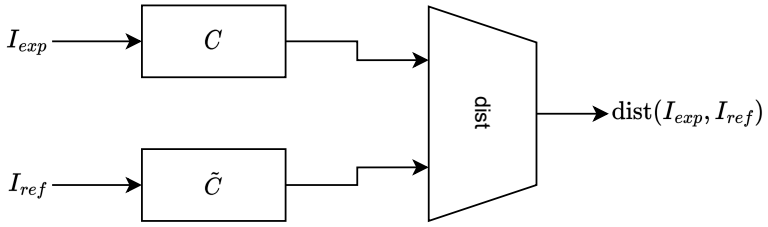
Moreover, image alignment is a very expensive algorithm, as it involves comparing millions of pairs of images. Thus, it can be stated that a significant fraction of the time spent in a typical SPA workflow can be attributed to this problem. The main focus of this project is to develop a novel image alignment method that outperforms existing solutions without sacrificing the quality of the results.

CTF Correction

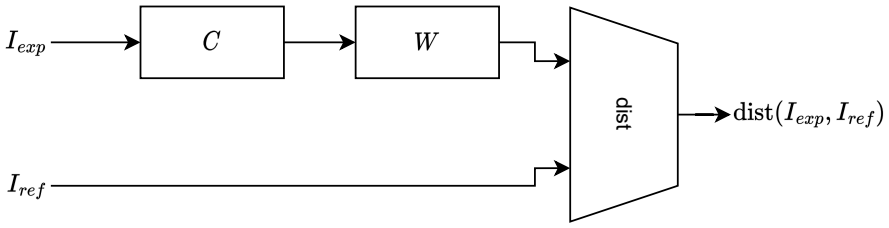
When comparing experimental images to reference ones, the CTF must be taken into consideration. Most implementations choose to apply the CTF to the reference images. The main drawback of this election is that the CTF varies across images, hence, the CTF needs to be re-applied to the reference gallery for each experimental image[31].

A possible solution to this problem consists in clustering similar CTFs into a reduced amount of discrete groups, so that for a given group, the CTF is only applied once to the reference gallery. However, this procedure adds complexity to the algorithm, specially considering that in streaming only a reduced subset of the data is available at a given time. Moreover, some datasets have a wide range of unique CTFs, which implies that a lot of small clusters would be generated, cancelling all the benefits of this approach.

This algorithm addresses this issue by correcting experimental images with a Wiener filter. As shown in the figure 4.2, the comparisons are made with the clean reference images, so the gallery does not need to be modified across comparisons.



(a) CTF to reference



(b) Wiener correction

Where C is the CTF, $\tilde{C}$ is the CTF estimation and W is the Wiener filter for correcting the estimated CTF.

Figure 4.2: CTF correction approaches

Among other applications, the Wiener filter can be used to deduce the inverse filter for a Linear Time Invariant (LTI) system, as is the case of the CTF model. This process is also known as the Wiener deconvolution, where the filter is obtained in such a way that the Mean Square Error (MSE) of the output is minimized[35]. The definition of the Wiener filter is displayed at (4.1), which also expresses it in function of the SSNR. Note that in absence of noise (infinite SSNR), the filter tends to behave as the inverse filter[36]. The term $N'(f)/S(f)$ acts to prevent overamplification when the direct filter's gain is low.

$$\hat{H}^{-1}(f) = \frac{H^*(f) \cdot S(f)}{|H(f)|^2 \cdot S(f) + N'(f)} = \frac{H^*(f)}{|H(f)|^2 + \frac{N'(f)}{S(f)}} = \frac{1}{H(f)} \cdot \frac{\frac{S(f)}{N'(f)} \cdot |H(f)|^2}{\frac{S(f)}{N'(f)} \cdot |H(f)|^2 + 1} \tag{4.1}$$

where the $S(f)$ term refers to the signal PSD before the direct filter, whilst the $N'(f)$ term refers to the noise PSD after the filter[35]. These signals are represented in the block diagram of Figure 4.3.

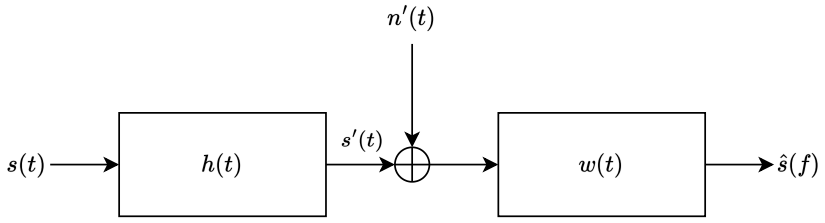


Figure 4.3: Wiener deconvolution block diagram

The Wiener filter was already implemented in Xmipp[17] as `xmipp_ctf_correct_wiener2d` program, so the existing implementation was re-used for this algorithm. This program assumes a constant SSNR profile across all frequencies. This constant SSNR value is derived from the Filter's mean energy. It has been empirically proven that using a 10% of the filter's energy leads to good results[37].

The main drawback of this approach is that the Wiener filter is not able to correct frequencies where the direct filter has a zero. As shown in (4.2), the Wiener inverse filter has also a zero for those frequencies, hence, those frequencies will not be at its output. In addition, the Figure 2.3 points out that the CTF has periodic zero crossings, which will be subjected to this phenomenon.

$$\lim_{|H(f)|\to 0} W(f) = 0 \tag{4.2}$$

As a consequence, some spatial frequencies are not present in the experimental image to be compared, but they are potentially present in the artificially generated image. This will induce a systematic error when comparing images. However, according to the experiments detailed later, we have not observed any significant influence of this effect.

Search vector description

Image comparisons can be formulated as vector comparisons. From now on we will consider image comparisons as abstract vector comparisons. The easiest way to turn an image into a vector is by flattening all its pixels into a single dimension. For instance, all the rows of a given image could be concatenated to form a vector. In other words, the search vector will have as many dimensions as the number of pixels on an image. A typical image size of 160×160 would require a vector of size 25600. This means that vectors will have a very high dimensionality, with the associated computational and storage cost.

It has been empirically proven that most of the alignment information is below a resolution of $6-8$Å[16]. This means that a downsampled version of the image could be used. Provided that the 160×160 image was captured with a pixel size of 2Å, it could be downsampled with a factor of 2, which would lead to a vector size of 6400.

What is more, because of the orthonormality of the Discrete Fourier Transform (DFT), the comparisons can be performed in Fourier space. This allows to extract coefficients from a disc with the radius of the resolution limit. This disc involves slightly less coefficients due to the fact that it is inscribed inside the downsampled Fourier space. The DC component of the image can also be omitted, as it does not provide any information for alignment. In the previous example, this would reduce the coefficient count to 2512 complex coefficients. For our purposes, complex coefficients can be flattened to a vector twice as large. This leads to a vector of 5024 dimensions, slightly less than the downsampled version of the image.

The Figure 4.4 illustrates this Fourier coefficient extraction. Note that only the left half of the Fourier space is considered, as the Fourier transform of a real valued signal poses conjugate symmetry. Thus, half of the Fourier space is redundant.

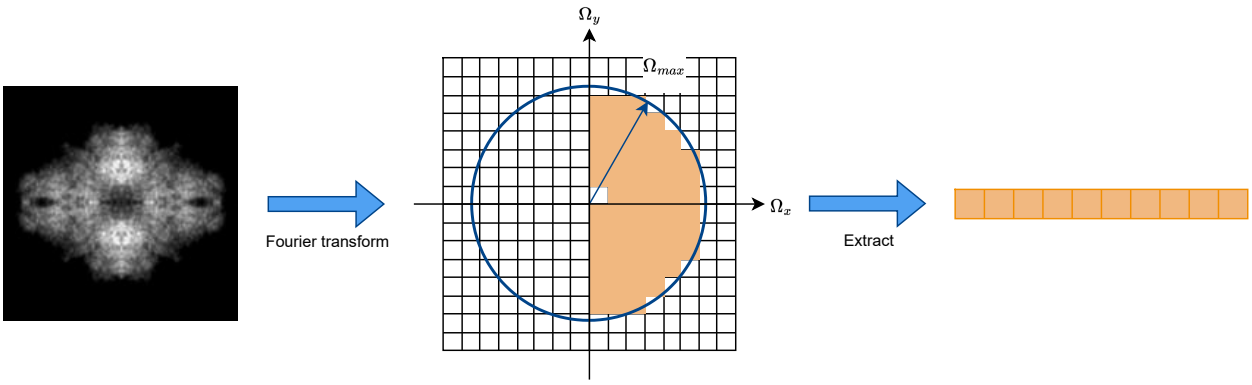


Figure 4.4: Fourier coefficient extraction

We have observed slightly more accurate results when using the Discrete Cosine Transform (DCT) instead of the DFT for a given disc size. However, the cost of computing the DCT is also greater.

As a consequence, we will select a low frequency disc in Fourier space to transform all images into a search vector. As shown in the previous example, the images can be reduced by a factor of $\frac{25600}{5024} \approx 5$ at little accuracy cost. In the results chapter, we will discuss the accuracy loss that can be attributed to discarding high frequency information.

Reference dataset generation

The reference dataset consists in all the in-plane transformations of the reference images. This implementation groups alignment parameters in such a way that intermediary images

can be stored and used repeatedly, avoiding redundant computations. Moreover, the individual operations have been reduced to the bare-minimum, so that their execution time is minimised.

Firstly, each image of the reference gallery is rotated in its own plane by a set of equally distributed angles. At this point, the images are transformed to Fourier space, so that the low frequency disc can be extracted as described earlier. In theory, this last operation could be done before applying the in-plane rotations to the images, which would imply far less repetitions of the Fourier Transform. However, this is not feasible in practice because interpolating Fourier coefficients tends produce incorrect results.

Then, the extracted Fourier coefficients can be multiplied with a set of shift filters defined in Fourier space. The shift filters define a translation in the spatial domain. This operation can be performed in Fourier space because a shift filter is a LTI system, which is applied as a point operation to the Fourier coefficients. Thus, it can be only applied to the coefficients extracted from the disc. The transfer function of the shift filter is defined as (4.3) where $\mathbf{\Omega}$ is the frequency vector and $\mathbf{\Delta}$ is the shift vector in pixels.

$$H(\mathbf{\Omega}) = e^{-j\mathbf{\Omega}\cdot\mathbf{\Delta}} \tag{4.3}$$

The whole process has been illustrated in Figure 4.5. At the end, a vast amount of vectors is generated, representing all possible combinations of in-plane transforms of the reference gallery. Each experimental image will be searched across this collection of vectors to find a best match.

Vector search techniques

At this point, our goal is to find the most similar vector across the previously generated dataset for each experimental image. In the context of ML, this problem is known as kNN. One of the main strengths of this algorithm is the usage of state-of-the-art vector search techniques. In particular, we have used the FAISS library[38], developed by Facebook Research. This library is known to be one of the fastest vector search utilities. Moreover, it supports input from the Torch library, which eases interfacing with the rest of the program.

FAISS allows to build very complex vector databases with a modular structure. Depending on this structure, the expected size and dimensions, it may be suitable to fit entirely in Central Processing Unit (CPU) memory or even GPU memory. This is specially useful for modules designed to take advantage of GPU accelerators. Moreover, this is also beneficial if data is already at the GPU, as it avoids making expensive copies between GPU and CPU memory.

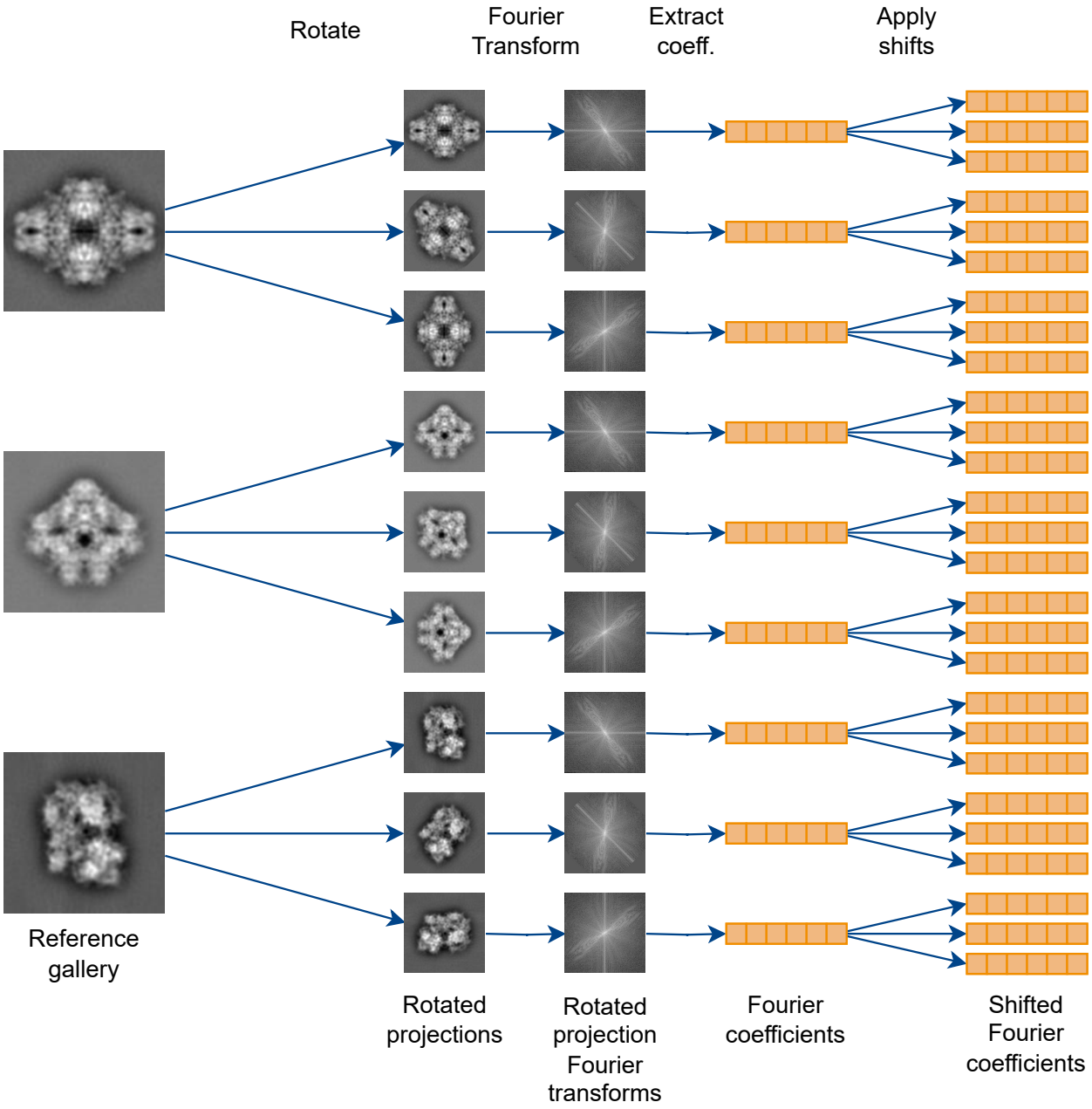


Figure 4.5: Reference dataset generation

We have deduced that the former gallery generation approach can easily reach a size of many million of images. Provided that each image is represented by a couple of thousands coefficients, it can be easily inferred that storing all these coefficients in a raw array would require many Gigabytes. This may fit in high-end server CPU memory, but it is unthinkable to store this amount of information on a GPU.

Therefore, we have chosen to use FAISS's vector quantisation and compression techniques. These methods are designed to compress vectors into a handful of bytes, enabling to store huge datasets in GPU memory. However, these compression techniques come at an accuracy cost. This accuracy loss is highly dependant on the nature of the data itself, as highly correlated data will tend to compress well, whilst independent variables will produce a significant degradation of quality. We will discuss the influence of these compression algorithms with CryoEM data in Chapter 5.

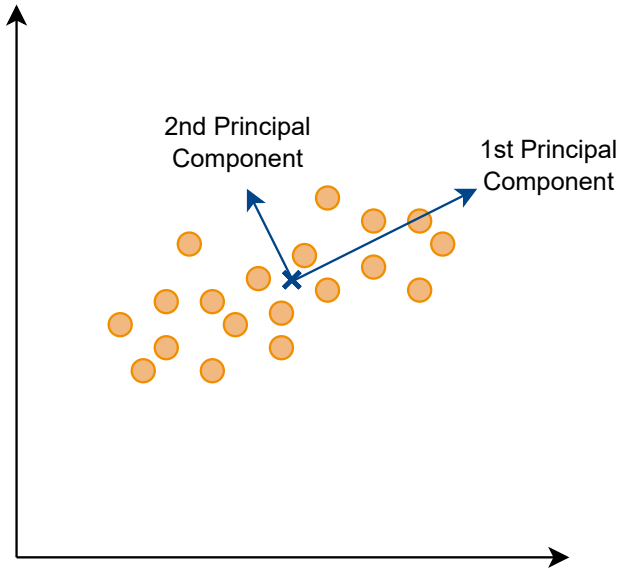
One of the most widely known vector compression techniques is Principal Component Analysis (PCA). PCA can be used to project vectors into a subspace in such a way that the energy lost during the projection is minimal. This projection reduces the dimensionality of the vectors, so more of them can be stored at a given space. Moreover, vector comparisons become more efficient as less components need to be compared.

We have also considered using the database architecture recommended by FAISS guidelines[38] that best represents the magnitudes of our data estimates. This architecture uses the Product Quantisation (PQ) vector compression technique to achieve aforementioned compression ratios. On top of it, it uses a Inverted File (IVF) data structure to accelerate searches. We have also selected these techniques because they have been implemented for GPUs, so that the whole search process can be run in these accelerators. Although these techniques have been already implemented in FAISS, a brief description of them is provided hereafter.

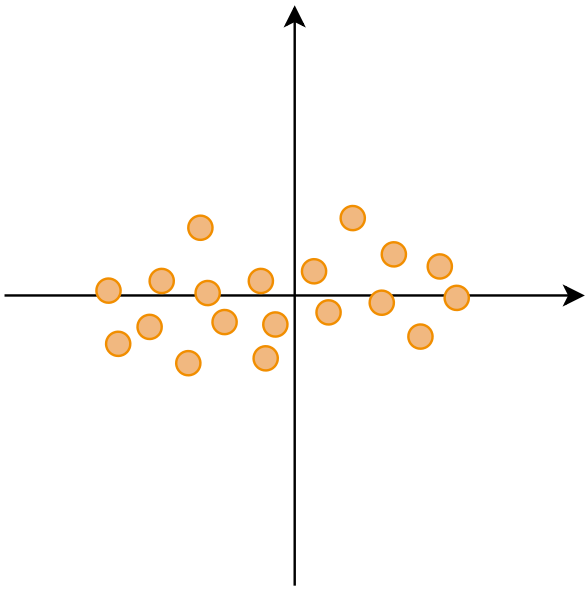
Principal Component Analysis

PCA is a family of techniques to reduce the dimensionality of vectors in such a way that the information lost in the process is minimised[39]. This is achieved with a linear projection of the vectors into a latent space. The basis vectors of this latent space correspond to the axes of the input data where the variance is maximal. For instance, Figure 4.6 shows the axis with maximum variance of a set of 2D points.

The effectiveness of the PCA greatly depends on the covariance between axes. If axes are independent random variables (covariance is zero) there is no principal component and dimensionality reductions will produce a proportional information loss. Contrary to this, if the axes are highly correlated, discarding the least important principal components will produce little to no degradation[39][40].



(a) Principal Component Analysis



(b) Projection into Principal Components

(c) 1D latent space from the 1st Principal component

Figure 4.6: Example of Principal Component Analysis dimensionality reduction

The first step to perform a PCA is to centre samples around the origin of coordinates. To do so, the centroid of the samples is subtracted from them:

$$\boldsymbol{x_i} = \boldsymbol{a_i} - \boldsymbol{\mu} \tag{4.4}$$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{a_i} \tag{4.5}$$

We can define the sample matrix X as the vertical stacking of the mean centred sample vectors:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1}^T \\ \boldsymbol{x_2}^T \\ \vdots \\ \boldsymbol{x_N}^T \end{bmatrix} \tag{4.6}$$

This enables us to compute the covariance matrix of the vectors:

$$\mathrm{Cov}(\boldsymbol{X}) = \frac{1}{N-1}\boldsymbol{X}^T\boldsymbol{X} \tag{4.7}$$

After performing a Singular Value Decomposition (SVD) decomposition of the covariance matrix so that $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{U}^T\boldsymbol{\Lambda}\boldsymbol{U}$ where $\boldsymbol{U}$ is a orthonormal basis formed by the eigenvectors of the covariance matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is a diagonal matrix with the eigenvalues of the covariance matrix. These eigenvalues represent the variance in their corresponding eigenvector's direction. Hence, selecting a handful of eigenvectors with the largest eigenvalues leads to an orthonormal basis where most of the input vector's energy is preserved. This basis can be used to perform a linear projection of vectors into a lower dimension latent space with minimal information loss.

Product Quantisation

Product Quantisation (PQ) is a vector compression technique suitable for performing efficient kNN searches. Unlike other vector compression techniques, it has been designed to work well with large vectors[41][42].

This algorithm splits each vector into fixed sized chunks, which will be individually treated. The last chunk may be padded with zeros if necessary. A example of this partition is shown in Figure 4.7.

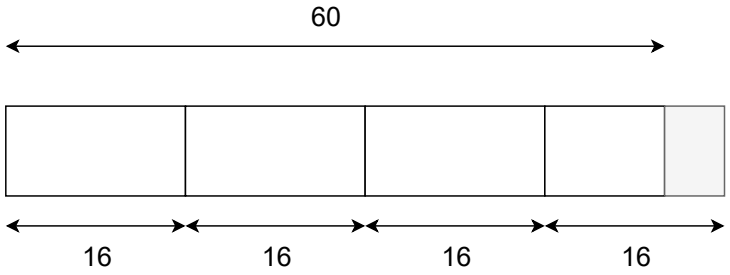


Figure 4.7: Example of vector partitioning for the PQ compression algorithm

At the beginning, the PQ encoder needs to be trained with a representative subset of the dataset. In this training process, a k-means partition computed for each chunk of the vector, producing k centroids. This process is illustrated in Figure 4.8. K-means is a unsupervised clustering algorithm that groups points in such a way that the distance to the centroid of their corresponding class is minimized[43]. K-means requires the training set to be much larger than k in order to be effective.

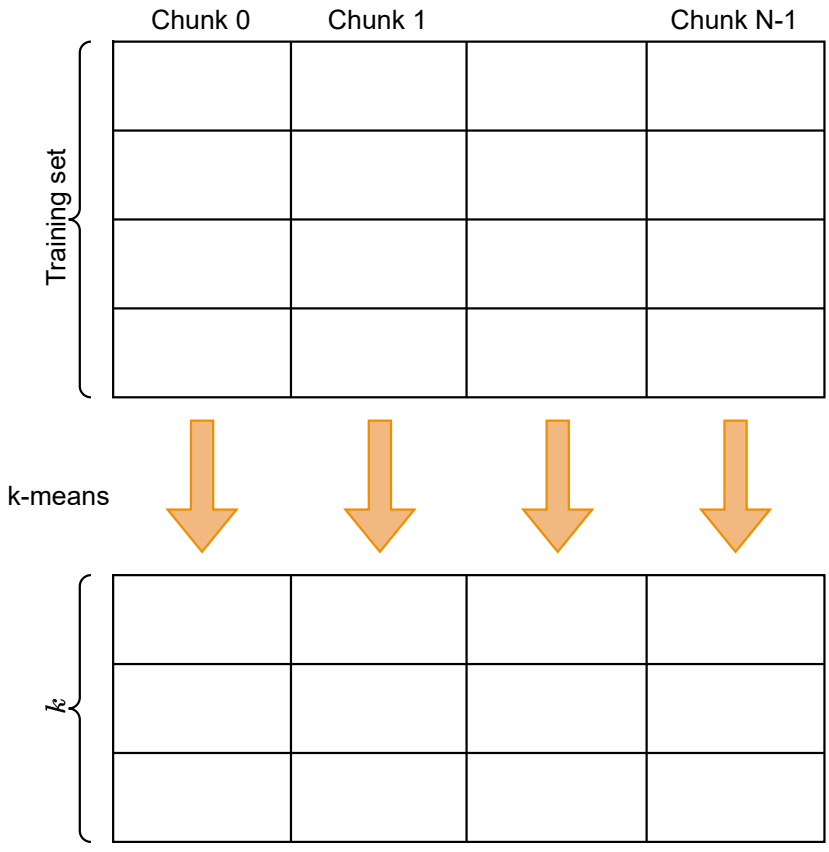


Figure 4.8: K-means usage for PQ vector compression

At this point, the PQ encoder can be used to encode new vectors. The encoding consists in matching the closest centroid for each chunk of the vector. Hence, the vector can be encoded with a set centroid identifiers. This sequence of identifiers is known as PQ-code[41]. Provided that k centroids have been computed for each chunk, $\lceil \log_2 k \rceil$ bits are required to index each chunk. In total, $N \cdot \lceil \log_2 k \rceil$ bits will be necessary to store a vector, N being the

number of chunks[44]. At most k^N unique vectors can be represented with this technique. Beyond this number, code collisions are guaranteed to occur.

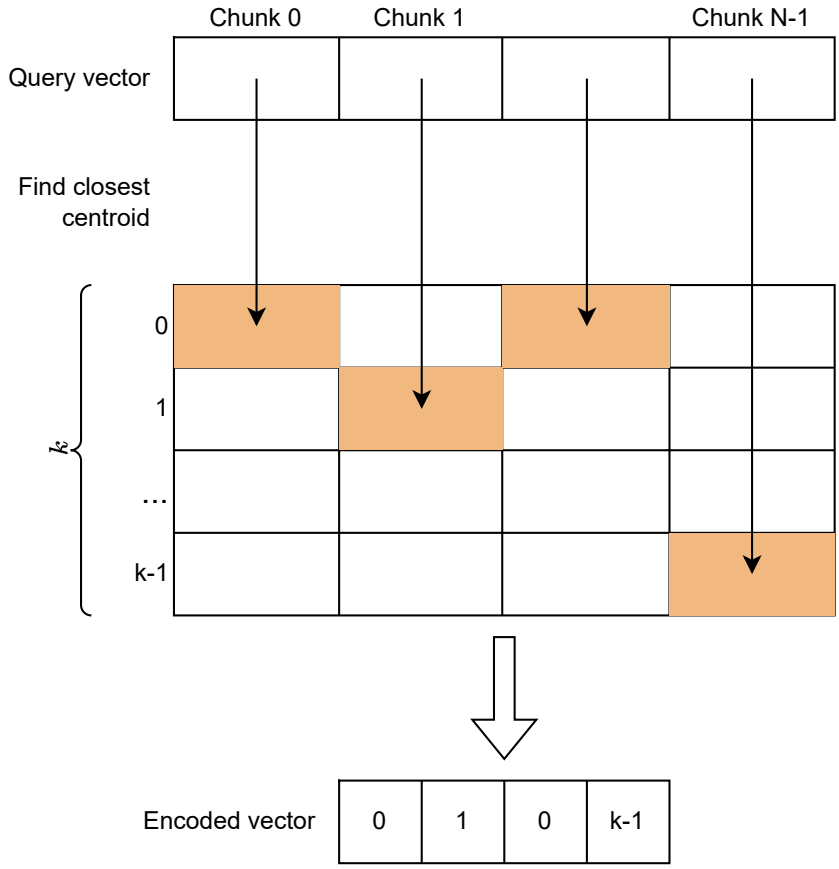
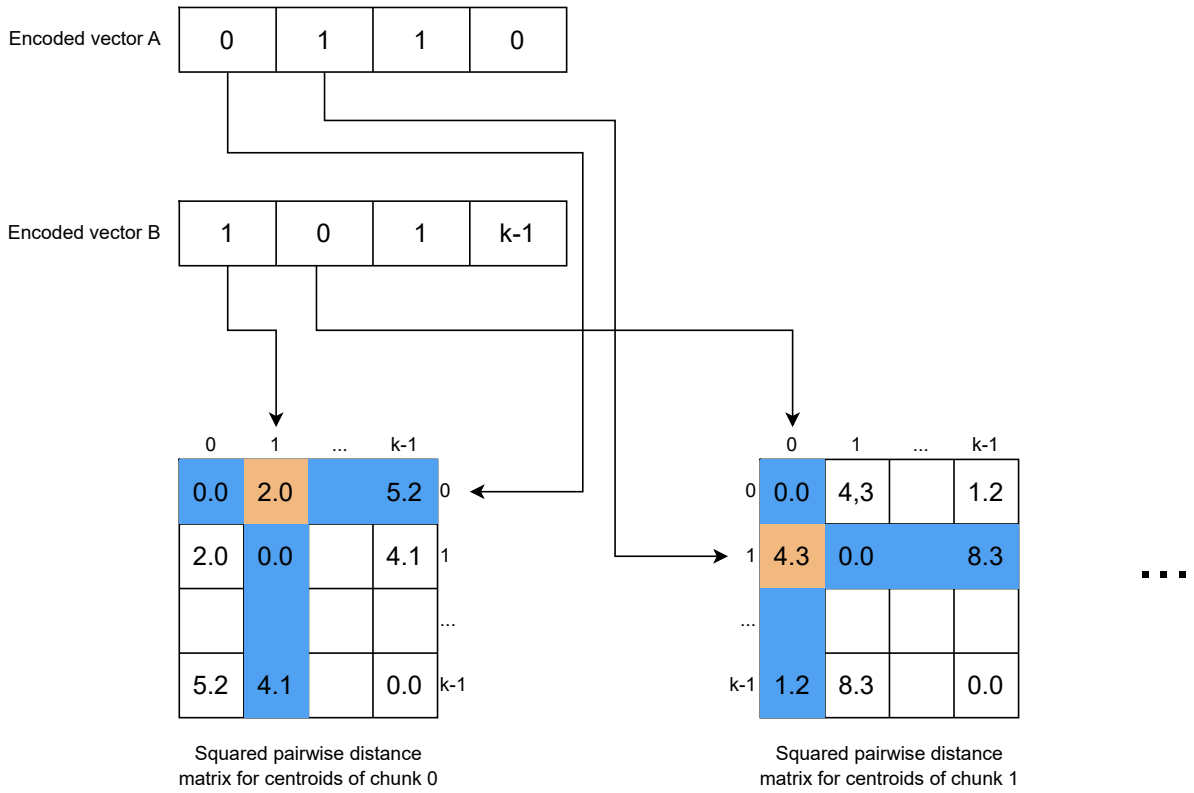


Figure 4.9: Example of PQ encoding

Decoding PQ codes is not necessary for our purpose, but it can be achieved. This process would involve following the inverse path of the Figure 4.9. This is, centroids would be indexed according to the encoded PQ-codes and then concatenated to ensemble a vector.

This technique also allows to efficiently compute distances between a given pair of encoded vectors. To do so, all the pairwise squared euclidean distances between centroids are pre-computed and stored in memory. Then, it is a matter of indexing and accumulating this pre-computed distance for each chunk. Figure 4.10 illustrates this distance computation process for the squared Euclidean distance. However, this idea can be extended for many other distance metrics. Note that the figure illustrates pairwise distances with a matrix. In practice, matrices are not used to store pairwise distances, as they are inefficient due to the symmetry of the distance functions ($\text{dist}(\boldsymbol{A}, \boldsymbol{B}) = \text{dist}(\boldsymbol{B}, \boldsymbol{A})$) and trivial cases ($\text{dist}(\boldsymbol{A}, \boldsymbol{A}) = 0$).

Having a fast distance estimate accelerates kNN vector searches, as these involve comparing a given vector against all the vectors in the dataset. Nevertheless, PQ comes with its own drawbacks. As its name suggests, it is a quantisation procedure, hence, there will be some accuracy loss. This accuracy loss is determined by the dispersion of the data. If the data is structured in a few compact clusters, K-means will be able to identify them and the



$$\text{dist}(\mathbf{A}, \mathbf{B})^2 = 2.0 + 4.3 + \dots$$

Figure 4.10: Example of PQ distance calculation

quantisation error will be minimal. Contrary to this, if the data is random, the centroids may not be representative and a lot of quantisation error will be introduced.

The quantisation loss can be minimised applying a linear transform to the vectors before quantizing them. This linear transform is represented with a rotation matrix, which will rotate vectors in such a way that the quantisation error is minimised. Due to the fact that rotation matrices are orthonormal, distances are preserved after this rotation[45]. This matrix is calculated from a PCA matrix, but its columns are reordered in such a way that it ensures that information is evenly spread across all chunks.

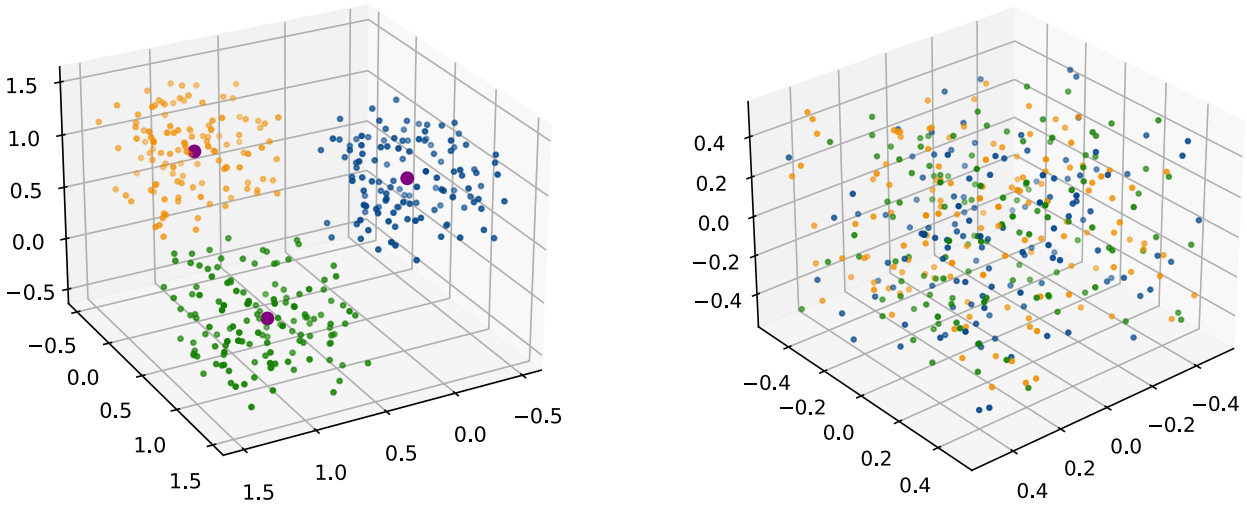
Inverted File

PQ provides a compact way to store vectors and reduce distance computation time. However, it is not effective enough when searching across millions of vectors[46]. Therefore, it is usually complemented with Inverted File (IVF) to significantly reduce search times.

IVF works by limiting the search scope for a particular query vector. To do so, once again K-means is used. This time, K-means is used to partition the entire vector space (instead of a small set of axes). Once again, this partition is done in the training process, so a large

enough represetentive subset of the data needs to be provided. This process will provide a coarse quatization of the vector space in discrete cells, which can be seen as Voronoi cells[46]. These cells can be used to limit the extent of the searches.

Once the database has been trained, vectors are grouped according to their closest centroid. Then PQ will be used to encode the residual vector from the centroid. A example of this partitioning is provided in Figure 4.11. Therefore, a vector will be encoded as the index of its coarse centroid and the PQ-code assotiated to the residual vector[46].



(a) K-means clustering

(b) Residual vectors

Figure 4.11: K-means centroid and residual vector example for IVF searches

When a search needs to be carried out, the query vector will also be assigned to the closest centroid. Then, the residual vector will be searched across all the residual vectors belonging to that group. However, this may pose a problem. If the query vector falls near the cell frontier, it is possible that the closest vector may be in the neighbouring cell. This case is exemplified with the B point in Figure 4.12, which gets assigned to the green region while its closest point is on the blue region. To solve this issue, each query vector is also searched across multiple neighbouring cells[46].

Distance metric

Until this point we have used the distance term without defining a concrete distance function. This was to preserve generality, as several distance functions were implemented. The aforementioned FAISS implementation of IVF-PQ encoding only allows for Euclidean distance minimisation and dot product maximisation. In fact, these metrics are related to one another:

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \langle \boldsymbol{x}, \boldsymbol{y} \rangle \tag{4.8}$$

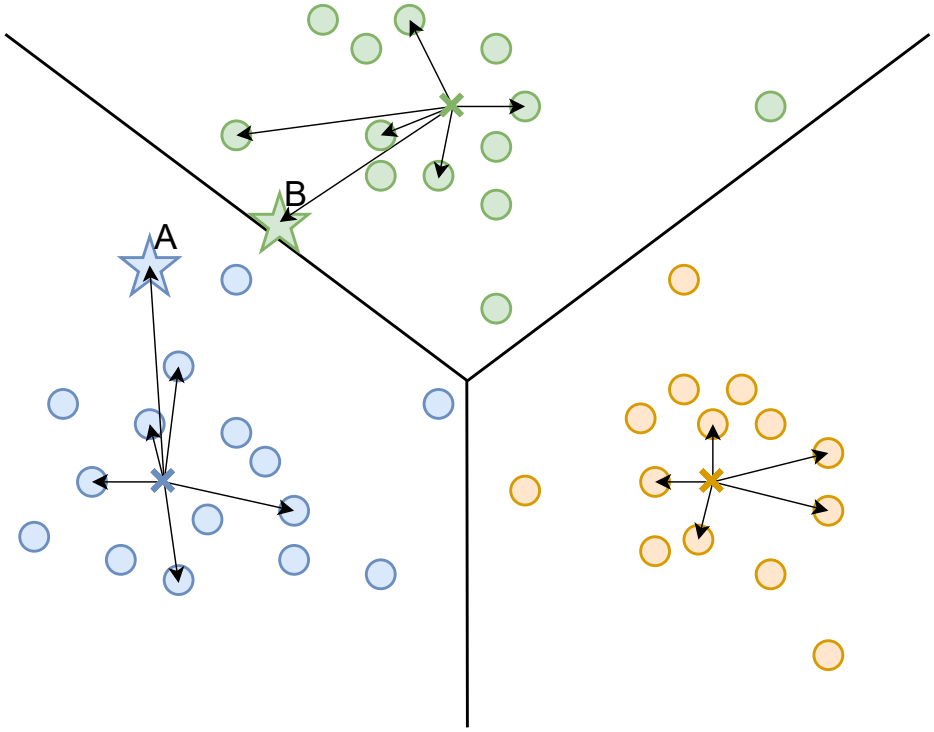


Figure 4.12: IVF search example

Nevertheless, these metrics can be used as a basis to define many other distance metrics. For instance, the cosine similarity can be defined in terms of the dot product by providing normalised vectors.

$$\cos\theta = \left\langle \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \right\rangle \tag{4.9}$$

Similarly, the Pearson correlation can be calculated by subtracting the mean before calculating the cosine distance:

$$\rho = \left\langle \frac{\boldsymbol{x} - \bar{\boldsymbol{x}}}{\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|}, \frac{\boldsymbol{y} - \bar{\boldsymbol{y}}}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|} \right\rangle \tag{4.10}$$

Lastly, the weighted Euclidean distance can be calculated by scaling vector coefficients before computing the Euclidean distance:

$$\|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{w}}^2 = \left\|\tilde{\boldsymbol{W}}(\boldsymbol{x} - \boldsymbol{y})\right\|^2 \tag{4.11}$$

where

$$\tilde{\boldsymbol{W}} = \mathrm{diag}(\sqrt{\boldsymbol{w}}) \tag{4.12}$$

Note that the previous definitions have been stated for real numbers. However, our vectors have been ensembled from complex vectors by interleaving their real and imaginary part:

$$\boldsymbol{x} = \begin{bmatrix} \mathrm{Re}(\hat{x}_1) & \mathrm{Im}(\hat{x}_1) & \mathrm{Re}(\hat{x}_2) & \mathrm{Im}(\hat{x}_2) & \dots \end{bmatrix} \tag{4.13}$$

This vector construct preserves the Euclidean norm from its original complex form and $+/-$ properties. Therefore, the euclidean distance remains valid.

$$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x} = \mathrm{Re}(\hat{x}_1)^2 + \mathrm{Im}(\hat{x}_1)^2 + \dots \tag{4.14}$$

$$\begin{aligned} \|\hat{\boldsymbol{x}}\|^2 = \hat{\boldsymbol{x}}^H\hat{\boldsymbol{x}} = \\ \mathrm{Re}(\hat{x}_1)\cdot\mathrm{Re}(\hat{x}_1) + \mathrm{Im}(\hat{x}_1)\cdot\mathrm{Im}(\hat{x}_1) + \\ \cancel{j\,\mathrm{Im}(\hat{x}_1)\cdot\mathrm{Re}(\hat{x}_1)} - \cancel{j\,\mathrm{Im}(\hat{x}_1)\cdot\mathrm{Im}(\hat{x}_1)} + \dots = \\ \mathrm{Re}(\hat{x}_1)^2 + \mathrm{Im}(\hat{x}_1)^2 + \dots \end{aligned} \tag{4.15}$$

However, in general, the dot product is not preserved from one form to the other because it discards the imaginary part of the result:

$$\boldsymbol{x}^T\boldsymbol{y} = \mathrm{Re}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1) + \mathrm{Im}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1) + \dots \tag{4.16}$$

$$\hat{\boldsymbol{x}}^H\hat{\boldsymbol{y}} = \mathrm{Re}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1) + \mathrm{Im}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1) + j\,\mathrm{Re}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1) + j\,\mathrm{Im}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1) + \dots \tag{4.17}$$

In spite of this, our particular case involves Fourier coefficients, which have symmetry. When considering symmetric pairs of coefficients, the imaginary part of the dot product cancels out. Hence, for the case, the dot product is preserved.

$$\begin{aligned} \hat{\boldsymbol{x}}^H\hat{\boldsymbol{y}} + (\hat{\boldsymbol{x}}^*)^H(\hat{\boldsymbol{y}}^*) = \\ \mathrm{Re}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1) + \mathrm{Im}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1) + \cancel{j\,\mathrm{Re}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1)} + \cancel{j\,\mathrm{Im}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1)} + \\ \mathrm{Re}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1) + \mathrm{Im}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1) - \cancel{j\,\mathrm{Re}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1)} - \cancel{j\,\mathrm{Im}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1)} + \dots = \\ 2\cdot(\mathrm{Re}(\hat{x}_1)\cdot\mathrm{Re}(\hat{y}_1) + \mathrm{Im}(\hat{x}_1)\cdot\mathrm{Im}(\hat{y}_1) + \dots) \end{aligned} \tag{4.18}$$

Considering these two facts, the flattened representation of the complex vector leads to numerically identical (and more efficient) results when computing distances and similarities.

4.3 Refinement cycle

Particle alignment is a core problem in the context of CryoEM. In the previous section, we have described our implementation of such an algorithm. However, the alignment problem on its own does not solve any real task. Thus, we have elected to use the 3D refinement problem as a "playground" for testing the alignment algorithm.

As stated in Chapter 1, the 3D refinement is used to iteratively enhance the resolution of the map. On each iteration, the current volume is projected from all directions to form a reference gallery. Then each experimental image is searched across this gallery to determine the direction it was projected from. This allows to use the experimental images to reconstruct a new volume with presumably higher resolution. Then, this volume can be used as the reference volume for the next cycle.

Although this description of the refinement process is intuitive, it is very naive. In practice, many additional steps need to be introduced in-between the previous steps to assure the quality of the results and prevent over-fitting. On this section we will deeply describe the logic implemented on top of the alignment algorithm that allows to perform a refinement iteration.

Reference volume projection

The first step in the alignment process is to generate a projection gallery of the current reference volume. One of the most crucial parameters regarding the projection is the angular sampling rate. This sampling rate describes the angular interval on which projections of the volume are generated.

This sampling directly affects the quality of the results, as the angular assignment of the images will be quantised to this interval. Ideally, the quantisation error has a uniform distribution $\mathcal{U}(-\frac{\Delta\Phi}{2}, +\frac{\Delta\Phi}{2})$. Therefore, in the best case scenario, the absolute error will have an average value of $\frac{\Delta\Phi}{4}$, where $\Delta\Phi$ is the angular sampling rate. This suggests that the angular assignment error is directly proportional to the angular sampling rate. However, there is another limiting factor to the accuracy, the resolution.

Recalling the alignment algorithm, it was stated that the image comparisons are only performed up to a certain resolution. This resolution is determined by the resolution of the reference map, as it does not make sense to compare coefficients beyond the maximum resolution of the map. The Fourier Rotation theorem states that the rotations in the spatial domain corresponds to the same rotation in Fourier space. Note that we are dealing with discrete Fourier space, which has discrete frequency components. Hence, at most we are only able to detect a rotation that would shift a coefficients at the resolution limit. This mini-

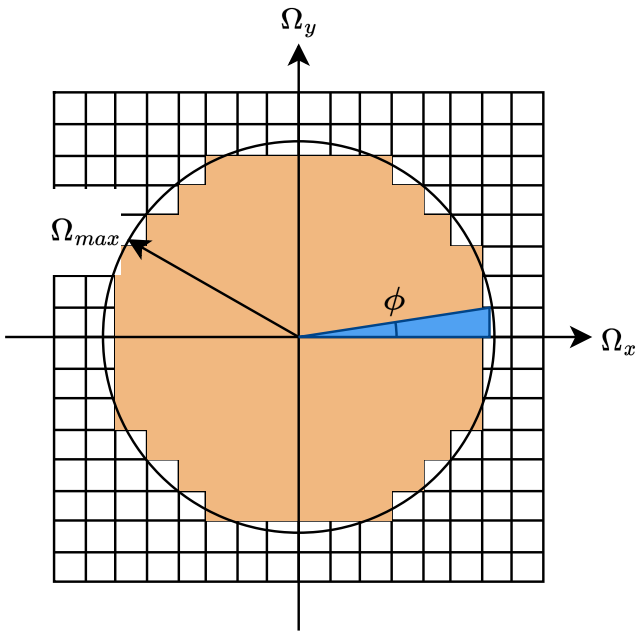


Figure 4.13: Maximum measurable angle at the resolution limit

mally measurable angle is the same for both 2D and 3D cases. The former one is illustrated in Figure 4.13.

Each of the coefficients in Discrete Fourier space is spaced by $\Delta\Omega = \frac{2\pi}{N}$rad where N is the size of the image. Assuming that we can detect changes that involve 1 coefficient shift in the resolution limit, the finest angle shift we can detect is:

$$\Delta\Phi = \sin^{-1}\left(\frac{\Delta\Omega}{\Omega_{max}}\right) \tag{4.19}$$

Knowing that $\Omega_{max} = 2\pi\frac{T_s}{T_{max}}$ where T_s is the pixel size and T_{max} is the resolution limit, the optimal angular sampling rate would be:

$$\Delta\Phi = \sin^{-1}\left(\frac{\cancel{2\pi}\frac{1}{N}}{\cancel{2\pi}\frac{T_s}{T_{max}}}\right) = \sin^{-1}\left(\frac{T_{max}}{NT_s}\right) \tag{4.20}$$

In order to avoid repeatedly sampling in the same directions, a random perturbation is added to the projection directions. Empirical results have shown that a value of $\frac{\Delta\Phi}{4}$ offers good results. This value coincides with the average error of the quantisation.

Once the projection parameters are chosen, the volume is projected from quasi-equally spaced angles using the `xmipp_angular_project_library` program, which was already implemented in Xmipp. This will produce our reference gallery. Prior to this projection, a mask

may be applied to the reference volume, so that the alignment is focused on a particular Region of Interest (ROI).

Additionally, if multiple reference volumes are provided for 3D classification, this step is repeated for each volume. Then all the galleries are combined into a single one, keeping track of the class that each reference image belongs to. in this way, the alignment will not only determine the angular assignment but also the 3D class.

Training

Once a reference gallery is obtained, it will be used to train the vector database used in searches. To do so, `xmipp_swiftalign_train` program is used, which augments the reference gallery by applying random in-plane transformations and uses this augmented dataset to train the FAISS database.

Alignment

At this point, the previously generated reference gallery and database can be used to align experimental images. This is achieved using `xmipp_swiftalign_query` program. This program will efficiently produce all the in-plane transformations of the reference gallery and then query each of the experimental images in the database. As a consequence, the sampling of the in-plane transformations also needs to be deduced. The rotational sampling estimated for projections still applies for in-plane rotations. Therefore, only the shift sampling rate needs to be computed.

This sampling rate is calculated assuming that for Nyquist a shift of 1px can be detected. As a consequence, this sampling can be scaled proportionally to the actual resolution limit:

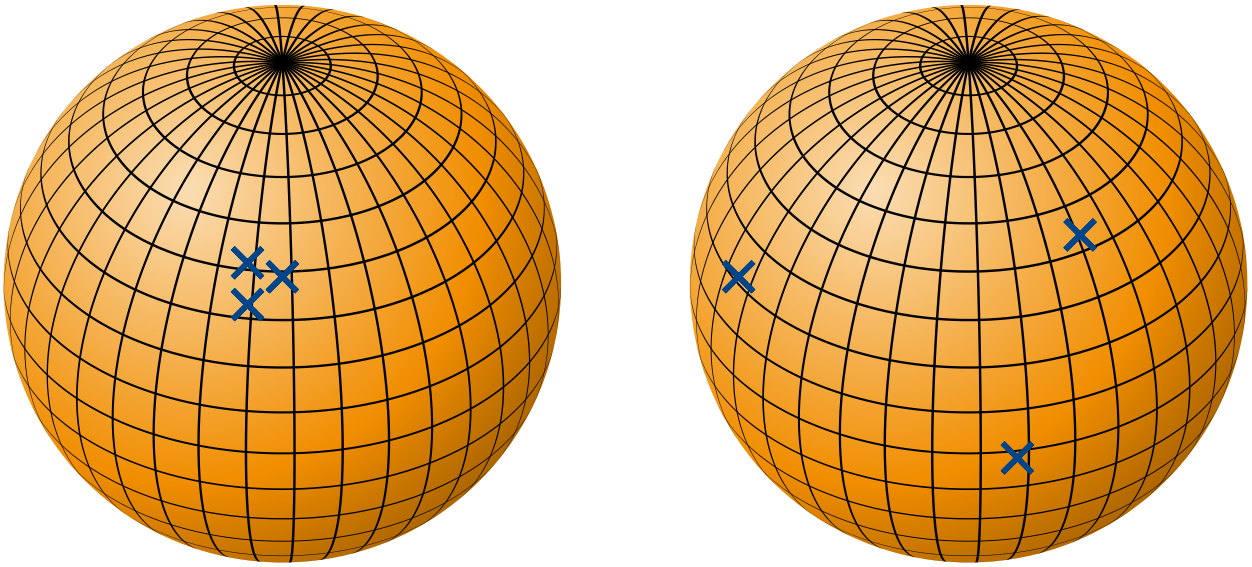
$$\Delta s = 1\text{px}\frac{T_{max}}{T_{nyq}} = 1\text{px}\frac{T_{max}}{2T_s} \tag{4.21}$$

Alignment consensus

We can leverage the speed offered by our alignment method to enhance its outcomes. For this purpose, we can perform multiple alignments to consensuate their outputs, ensuring we retain results that we are confident about. This novel approach enables us to balance speed and accuracy according to the user's requirements. The underlying principle behind the consensus is that we prioritise using fewer particles for reconstruction rather than numerous poorly aligned ones.

Nevertheless, the alignment algorithm is deterministic, meaning that running it multiple times will consistently yield the same results. To overcome this deterministic behaviour, we can generate multiple galleries. As mentioned earlier, each gallery is created with a unique random perturbation. By generating multiple galleries, each with its own distinct random perturbation, we can circumvent the deterministic nature of the algorithm by utilising slightly different galleries for each run.

The aim of the consensus step is to combine the outputs of multiple alignments in such a way that particles that were not aligned properly are discarded. Similarly, particles that coincided in their alignment get their results averaged. An example of such as consensus is displayed in Figure 4.14. This figure illustrates that when all alignments produce similar results, there is consensus among them. Contrary to this, if the 3D alignments have disperse results, there is no consensus and the particle should be discarded.



(a) Consensus (b) No consensus

Figure 4.14: Illustration of angular consensus

Note that the previous example was provided with 3 repetitions for consensus. Nevertheless, the implementation is generalised for any number of classifications. Using this principle, we have elaborated a list of criteria to filter particles. If any of the criteria is not met, the particle will be dropped from reconstruction.

- The angular assignment is validated by calculating the average angular assignment of the runs. Then, the distance from this average to the samples is computed. The angular assignment is considered to be valid only if more than the 50% of alignments are closer than the angular sampling rate ($\Delta\Phi$). If so, this average angular assignment is used for reconstruction.

The averaging of angular assignments must be done in quaternion space. Quaternions are an extension to the complex number system which are useful for representing the orientation in 3D space. Contrary to the Euler angles, which are more intuitive, quaternions are not subjected to the gimbal lock issue, making operations easier. In spite of this, the alignment parameters are usually represented with Euler angles. Hence, the Euler angles of the alignment need to be converted to quaternions.

$$\begin{aligned}
q_0 &= \cos\left(\frac{\phi}{2}\right)\cdot\cos\left(\frac{\theta}{2}\right)\cdot\cos\left(\frac{\psi}{2}\right) - \sin\left(\frac{\phi}{2}\right)\cdot\sin\left(\frac{\theta}{2}\right)\cdot\sin\left(\frac{\psi}{2}\right)\\
q_1 &= \cos\left(\frac{\phi}{2}\right)\cdot\cos\left(\frac{\theta}{2}\right)\cdot\sin\left(\frac{\psi}{2}\right) + \sin\left(\frac{\phi}{2}\right)\cdot\sin\left(\frac{\theta}{2}\right)\cdot\cos\left(\frac{\psi}{2}\right)\\
q_2 &= \cos\left(\frac{\phi}{2}\right)\cdot\sin\left(\frac{\theta}{2}\right)\cdot\cos\left(\frac{\psi}{2}\right) - \sin\left(\frac{\phi}{2}\right)\cdot\cos\left(\frac{\theta}{2}\right)\cdot\sin\left(\frac{\psi}{2}\right)\\
q_3 &= \sin\left(\frac{\phi}{2}\right)\cdot\cos\left(\frac{\theta}{2}\right)\cdot\cos\left(\frac{\psi}{2}\right) + \cos\left(\frac{\phi}{2}\right)\cdot\sin\left(\frac{\theta}{2}\right)\cdot\sin\left(\frac{\psi}{2}\right)
\end{aligned} \tag{4.22}$$

where ϕ is corresponds to the `rot` parameter, θ is the `tilt` parameter and ψ is the in plane rotation.

The average of quaternions is not trivial either. We have implemented the solution described by Markley, Cheng, Crassidis, *et al.* in their paper on Quaternion Averaging. This solution uses a Maximum Likelihood Estimation (MLE) of the average where the average distance from the centroid to the samples is optimised[47]. To do so, the first step is to ensemble a matrix with all the quaternions as rows:

$$\boldsymbol{Q} = \begin{bmatrix} w_1\boldsymbol{q_1}^T \\ w_2\boldsymbol{q_2}^T \\ \vdots \\ w_N\boldsymbol{q_N}^T \end{bmatrix} \tag{4.23}$$

where w_i is the weight associated to each quaternion. For our case, all quaternions will be equally weighted with $w_i = N^{-1}$. Then, the average quaternion can be calculated as the principal eigenvector of $\boldsymbol{Q}^T\boldsymbol{Q}$.

Regarding the validation of the average, the quaternion distance is defined as:

$$\Delta q = 2\cos^{-1}\left(\frac{1 - \|\boldsymbol{q_1} - \boldsymbol{q_2}\|^2}{2}\right) \tag{4.24}$$

Therefore the criteria to keep a particle is that at least $\frac{N}{2}$ comply with:

$$2\cos^{-1}\left(\frac{1 - \|\boldsymbol{q_i} - \overline{\boldsymbol{q}}\|^2}{2}\right) \leq \Delta\Phi \tag{4.25}$$

- Similarly to the previous criteria, the centroid of the shift assignment is computed to validate the shift assignment. 50% or more alignments should be less than the shift sampling rate apart from this centroid. When validated, this centroid is used for reconstruction.

  Formally, the average shift is defined as:

$$\overline{\boldsymbol{s}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{s_i} \tag{4.26}$$

  Then the alignments agree only if more than $\frac{N}{2}$ alignments comply with:

$$\|\boldsymbol{s_i} - \overline{\boldsymbol{s}}\| \leq \Delta s \tag{4.27}$$

- When a 3D classification is performed, each alignment will vote for a class. If there is no absolute majority, there is no consensus. Otherwise, the mode class is selected (which has been elected by more than the 50% of alignments).

Reconstruction

At this point the alignment particles are split into two equally sized random subsets. This is done in order to be able to compute the FSC between two reconstructions in the next step. Each of these subsets is used to reconstruct a volume using `xmipp_reconstruct_fourier` program (or its GPU accelerated variant[29]), which has been already implemented in Xmipp. As the name of the program suggests, the reconstruction is performed using the Fourier back-projection algorithm, detailed in Chapter 3. Each of these volumes is known as *half-map*, because they were obtained from half of the particles.

Resolution estimation and post processing

The final step involves comparing both half-maps to determine the resolution of the map. Since particles are aligned independently, each half-map was obtained separately. Consequently, regions that show strong correlation in both maps can be attributed to the signal, while uncorrelated regions are considered as noise.

The correlation between the Fourier coefficients for each frequency is measured using the FSC function, which was described in Chapter 3. This function provides a correlation function in terms of spatial frequency.

Typically, the FSC function exhibits a low-pass behaviour. The frequency at which this function crosses a specific threshold (usually 0.5 or 0.143) indicates the resolution of the map.

This number provides an insight about the level of detail that can be reliably represented in the volume.

Once the map's resolution is determined, both half-maps are averaged and then subjected to low-pass filtering up to the previously obtained resolution. This filtering step is performed to prevent overfitting caused by noise, as it effectively reduces the number of parameters to be estimated[48]. The resulting filtered average volume can be used as the reference volume for the next iteration or as the final output volume.

5. Results

Throughout the development of the image alignment algorithm, some compromises were made to optimise alignment times. The purpose of this Chapter is to assess the impact of these trade-offs on accuracy, comparing them against the performance benefits obtained. The alignment tests were carried out with the 3D alignments, as this was the main focus of the project. Nevertheless, results can be extrapolated to other CryoEM problems where image alignment is used as a basis.

For this evaluation, we will utilise multiple datasets, including both simulated datasets where all parameters are controlled, as well as real datasets. By incorporating real data into the analysis, we can gain insights into the algorithm's usefulness.

5.1 Test datasets

The alignment algorithm has been thoroughly tested using both synthetic and experimental datasets. Synthetic data is valuable because the ground truth values of the alignment parameters are known. As a consequence, any discrepancy in the output can be attributed as an error of the alignment algorithm. In spite of this, the synthetic data is usually generated using a naive model of the microscope. Therefore, results obtained with synthetic data may not reflect the real performance of the algorithm. To address this issue, the algorithm has also been tested with experimental data.

The datasets used for our tests were selected to reflect proteins with a diverse set of characteristics, such as symmetry, size, pixel size, presence of membrane, flexible parts... In total, 4 proteins were elected. Each of these proteins will be tested with both simulated and experimental data.

Experimental Data

The experimental images used in these tests were obtained from Electron Microscopy Public Image Archive (EMPIAR). EMPIAR is a public archive maintained by the European Molec-

ular Biology Laboratory (EMBL)-European Bioinformatics Institute (EBI) which provides open access to raw CryoEM images[49]. Among other things, this initiative enables testing CryoEM image processing algorithms with a wide variety of real data.

Some of the selected datasets provided extracted particles, which are the starting point of the alignment algorithm. However, some others only provided micrographs or movies. In those cases, data needed to be processed before being suitable for our use case. This processing was done inside the Scipion framework. Depending on the dataset, movie alignment, CTF estimation, particle picking and particle extraction steps needed to be carried out to obtain the desired particles.

To address the fact that the alignment information of the experimental images is not known, these particles were aligned twice using Relion[16] or Cryosparc[18]. Then, their outputs were consensuated so that only particles that coincided below a threshold are considered. This gives some amount of confidence to the estimated alignment of the particles, as two independent algorithms coincided in the result. Nevertheless, these parameters do not need to be the ground truth.

Synthetic Data

EMPIAR datasets have one or more atomic models associated to them in the Protein Data Bank (PDB) archive. The synthetic datasets used to test the algorithm were generated from those atomic models using the Scipion framework[19]. This process will mimic the behaviour of the TEMs.

Atomic models describe the the protein structure with atom coordinates. Therefore, the first step is to render a volume from this model. This can be easily achieved using the `xmipp3 - convert a PDB` protocol. Then, this volume is projected from all directions using `xmipp3 - create gallery` protocol, leading to a set of clean projections of the volume. However, these projections do not reflect experimental images due to the absence of noise. Therefore, some amount of Addititive Gaussian White Noise (AGNW) is added to the images through the `xmipp3 - add noise particles` protocol to simulate ice particles in the sample. This noise has zero mean and its standard deviation will be selected in such a way that the SNR of the image is -10dB. Lastly, a CTF is applied to the images using `xmipp3 - simulate CTF` protocol.

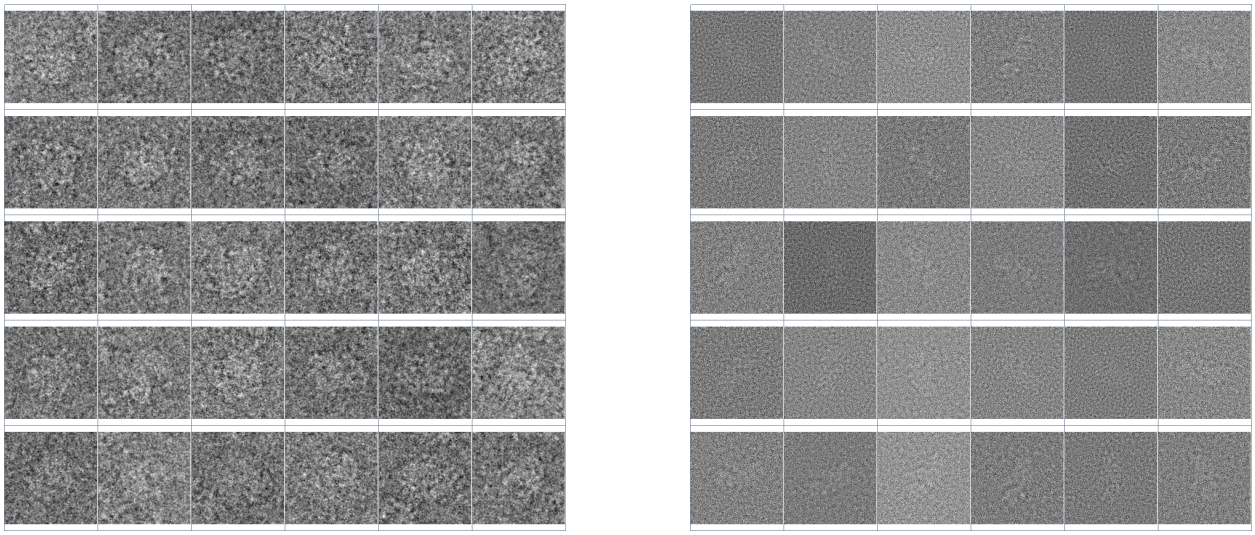
In total, 41219 projections will be generated from equally spaced orientations. This ensures that the algorithm will be tested with all possible orientations of the protein. Additionally, the particles will be shifted in-plane by a normal distribution of $\sigma = 6$px. Therefore, 95% of the images will contain a shift of less than 12px, which is a reasonably high value. Similarly, the defocus parameter of the CTF will be uniformly chosen for each particle in the range of

5000Å to 25000Å (typical CTF defocus values).

Test proteins

EMPIAR-10028

This acquisition is related to the Plasmodium falciparum protein, which is present in the human malaria parasite[50][51]. The dataset is widely used when testing CryoEM algorithms. In fact, as of May 2023, it has been cited by 28 publications related to new CryoEM methods. A visual sample of the particles is displayed in Figure 5.1.



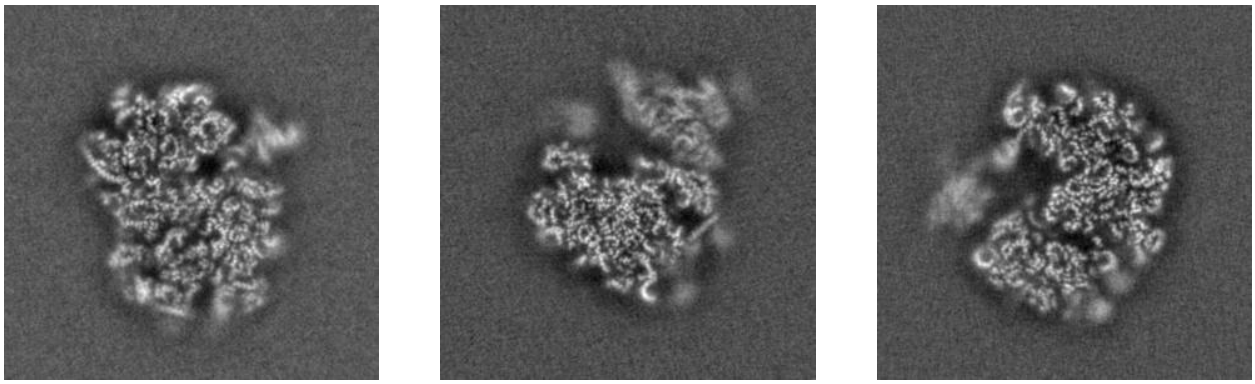
(a) Experimental images (b) Simulated images

Figure 5.1: Visual aspect of EMPIAR-10028 data

Parts of this protein are very flexible. This makes particles difficult to align, as the particles can not be fitted properly to the entire volume at once. Moreover, the protein is quite large, requiring particle images to be also large (300px wide).

As stated earlier, alignment algorithms use particles and a volume as the starting point. Although individual particles are provided in the dataset, these particles were obtained several years ago. Thus, we preferred to obtain the particles from scratch using modern methods so that we can take advantage of state-of-the-art algorithms.

After performing the angular assignment consensus between two independent Relion refinements, 60210 experimental particles were left with a discrepancy lower than 1° in the angular assignment and 0.5px in the shift assignment. These refinements reached a resolution of 4.06Å in their final volumes. The reconstructed volume can be observed in Figure 5.2. We also followed this procedure with Cryosparc (Homogeneous and Non-Uniform refinements) but we were unable to obtain good results.



(a) XY (b) XZ (c) YZ

Figure 5.2: Central slices of the reconstructed EMPIAR-10028 experimental dataset

EMPIAR-10061

The EMPIAR-10061 dataset is an acquisition of the β-galactosidase protein[52][53]. Similarly to the prior dataset, this dataset is also widely used when assessing CryoEM algorithms, having been cited in 23 publications related to CryoEM algorithms.

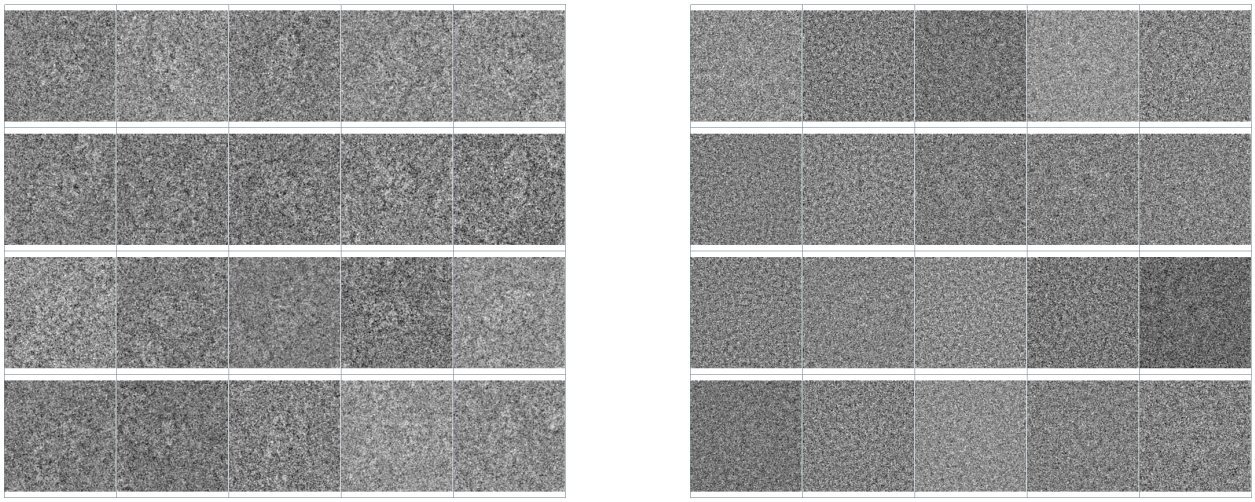
The particularity of this dataset is that it has $D2$ symmetry. This is important when aligning, because it reduces the number of possible projection angles. Moreover, each experimental image can be used to fill multiple planes in Fourier space when reconstructing, which usually increases SNR and resolution.

Contrary to the previous example, Relion's refinement introduced some artefacts. Therefore, Cryosparc's result was used for reference. The refinement reached a resolution of 2.6Å, which is close to the theoretical resolution limit imposed by Nyquist for its pixel size. After consensuating two independent Cryosparc runs, 39682 particles were kept with a discrepancy lower than 0.2px in shift assignment and 0.5° in angular assignment. The resulting reconstructed volume can be observed in Figure 5.4.

EMPIAR-10256

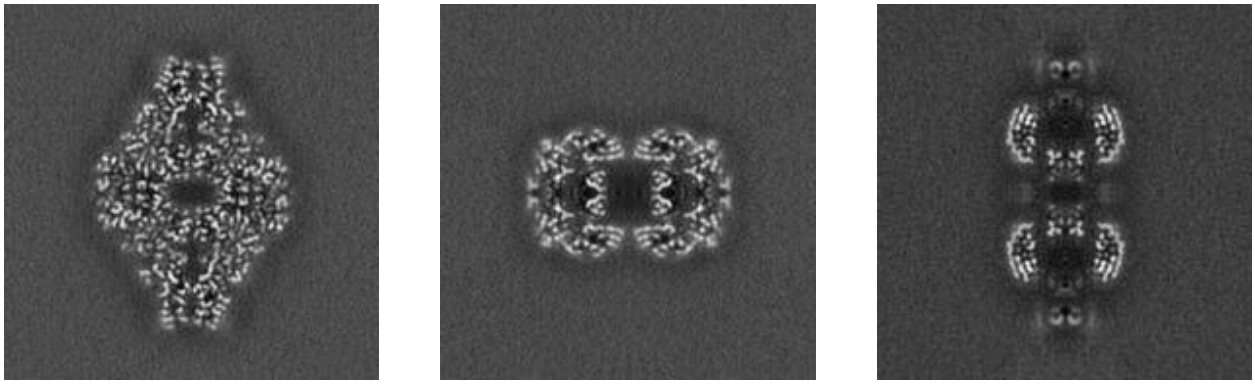
This dataset is a TRPV5 with calmodulin bound CryoEM acquisition[54][55]. This protein is present in the walls of the cells to exchange calcium with the outside. Due to is nature, the protein is embedded on a membrane, which makes it difficult to align. This is because the membrane is highly flexible and does not have a fixed pattern across particles. To make matters worse, the TPRV5 protein has C4 symmetry, but the calmodulin is not bound symmetrically. Therefore, the ensemble is pseudo-symmetric, producing a set of highly similar but disctinct views.

The dataset is provided either in the form of aligned particles or micrographs. For our purposes, the particles were elected as the starting point. These particles were refined using



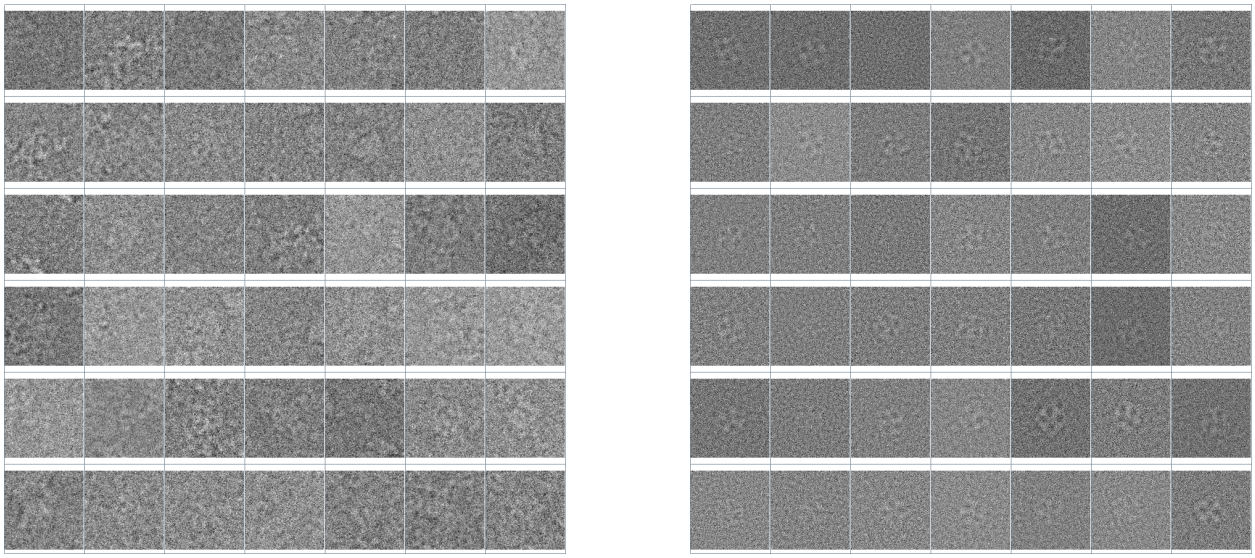
(a) Experimental images (b) Simulated images

Figure 5.3: Visual aspect of EMPIAR-10061 data



(a) XY (b) XZ (c) YZ

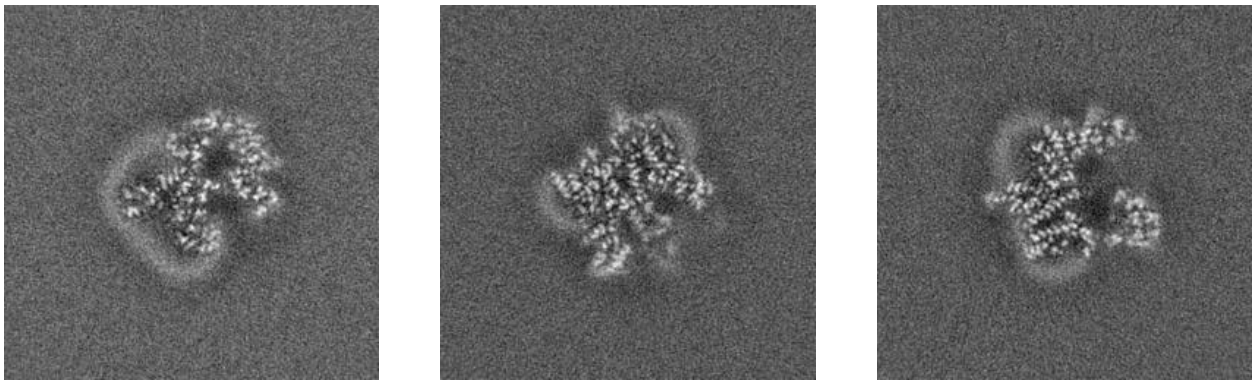
Figure 5.4: Central slices of the reconstructed EMPIAR-10061 experimental dataset



(a) Experimental images (b) Simulated images

Figure 5.5: Visual aspect of EMPIAR-10256 data

two runs of Cryosparc's Non-Uniform refinement. The angles produced by these refinements were consensuated to obtain 53964 particles with a discrepancy lower than 1° in angular assignment and 0.5px in shift assignment. The refinements reached a resolution 3.21Å on its last iteration and slices of the reconstructed volume can be observed in Figure 5.6. The same procedure was used with Relion but the results were worse.



(a) XY (b) XZ (c) YZ

Figure 5.6: Central slices of the reconstructed EMPIAR-10256 experimental dataset

EMPIAR-10391

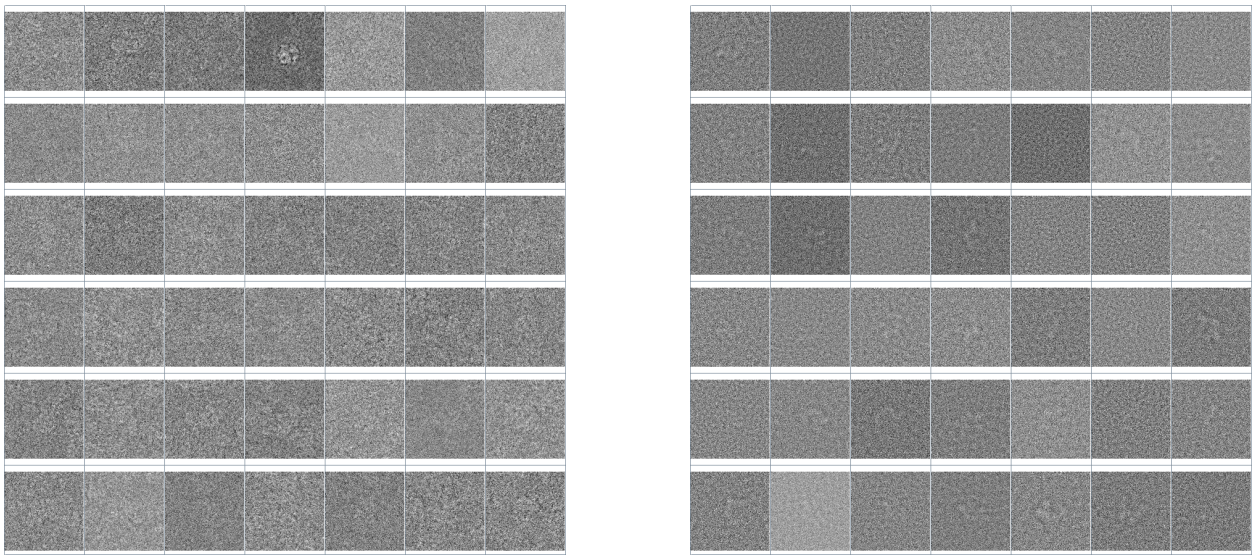
This dataset refers to a Arabinofuranosyltransferase AftD from Mycobacteria CryoEM acquisition. This protein is responsible of causing the tuberculosis decease, which kills over 1 million people every year[56][57].

Similarly to the previous example, the dataset is distributed either in the form of movies or a particle stack. As mentioned earlier, the later one suits better our needs, as the input for the alignment algorithm is a particle stack. Figure 5.7 shows the visual aspect of the dataset.

The aim of the experiment was to test a drug binding to the protein. Hence, the dataset is heterogeneous, this is, some particles may originate from the clean protein and some others may originate from the protein with the drug bound. As the dataset reflects two structures, two atomic models were fitted into it. Consequently, these particles can be used to test if the algorithm is able to distinguish discrete 3D classes.

Another peculiarity of this experiment is that the protein is embedded in a membrane. The membrane structure is highly flexible and does not follow a specific pattern across particles, making it difficult to align.

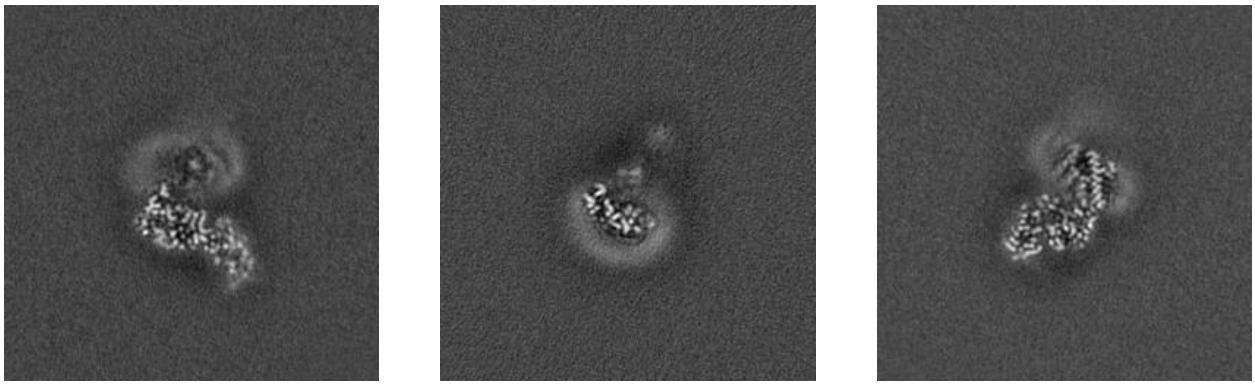
Similarly to the prior datasets, two independent Cryosparc refinements were run. Their angular assignments were consensuated to obtain a estimation of the ground truth values. At the end 96256 particles were kept. These Cryosparc refinements produced a volume with



(a) Experimental images (b) Simulated images

Figure 5.7: Visual aspect of EMPIAR-10391 data

a resolution of 2.9Å. Slices of this volume can be viewed in Figure 5.8. The same procedure was used with Relion but the results were worse.



(a) XY (b) XZ (c) YZ

Figure 5.8: Central slices of the reconstructed EMPIAR-10391 experimental dataset

5.2 Alignment performance

Influence of the Wiener CTF correction

As mentioned in Chapter 4, the approach to tackle the CTF of the experimental images consists in correcting them with a Wiener filter. The issue of this approach is that not all frequencies can be recovered due to bad SNR or zero gain at the CTF. Therefore, comparing those frequencies with the reference image may induce an artificial error. The aim of this section is to asses how this phenomenon affects the alignment accuracy.

To do so, several experiments will be carried out. Firstly, simulated images will be aligned with no CTF being applied to them (only noise). Obviously, experimental images can not be evaluated without CTF, as this is an artefact of the microscope. In any case, this experiment will be useful to observe the accuracy loss that can be attributed to the presence of the CTF. Moreover, the ground truth alignment parameters of the simulated images are known. Therefore, once the CTF has been applied to them, a reconstruction with these images is attempted. This gives an insight on the maximum achievable resolution with the simulated set of images.

In the second trial, experimental images will be clustered by their CTFs, so that the CTF can be assumed to be constant across all images of a given group. This can not be easily done with simulated images, as these do not originate from a micrograph. Thus, simulated images will be left out from this experiment. Once the particles have been clustered according to their CTFs, each of these image groups can be aligned against a reference set filtered with the representative CTF of that group. This approach, similar to the one followed by current refinement packages, will establish the baseline for the alignment accuracy comparison with CTF.

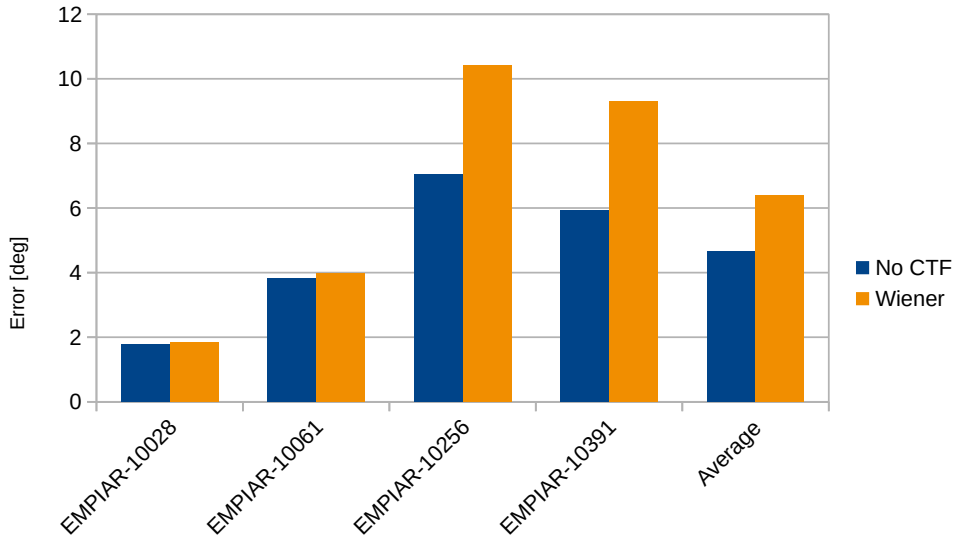
Lastly, the CTF will be corrected with a Wiener filter, both for experimental and simulated images. Then, these images will be aligned against a clean set of reference images. This will give an insight about the alignment quality degradation induced by aligning with Wiener corrected experimental images.

Earlier, it was stated that most of the alignment information is contained below the resolution of 8Å. However, this alignment method targets the initial cycles of the refinement loop, where the reference volume has much less resolution. Therefore, these experiments will be carried out with a resolution limit of 15Å, so that the algorithm is evaluated on its operational range. At this resolution, typical CTFs have one or two zero crossings. To ensure that the alignment errors can be attributed to the usage of the presence of the CTF and the Wiener filter, no vector compression techniques will be used in these tests.

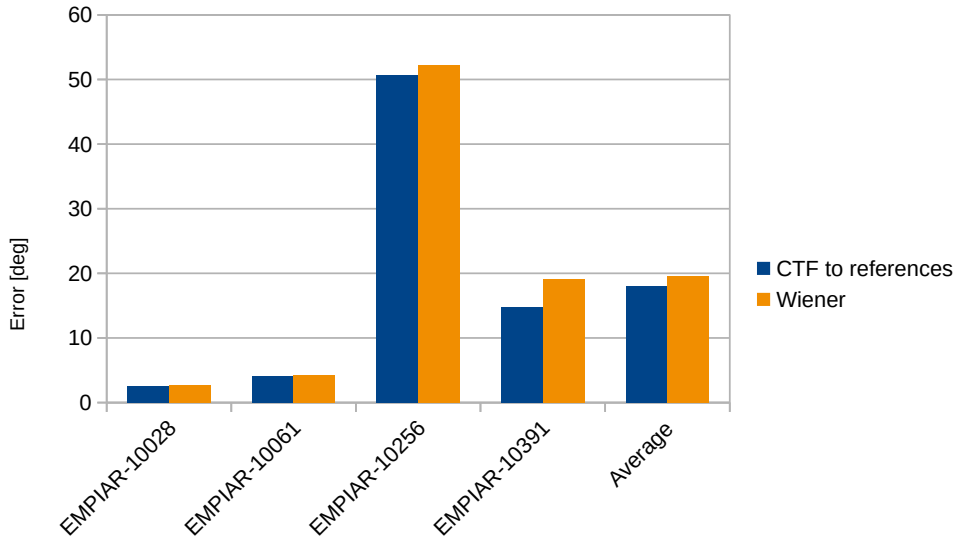
The tests have proved that using a Wiener filter to correct the CTF of the images does not pose a significant penalty respect to applying the CTF to the reference images. While it may introduce notable errors in certain simulated datasets (EMPIAR-10256 and EMPIAR-10391), the baseline for simulated datasets were images without CTF. As a result, these errors can not be only attributed to the CTF correction method but also the presence of the CTF itself. Indeed, the experimental images were assessed against the conventional CTF correction method. Figures 5.9 and 5.10 show that for experimental datasets, the angle and shift assignment error increase induced by the Wiener CTF correction is around 8%.

It can be noted that the angular assignment error for EMPIAR-10256 is considerably higher than the rest. This is because the calumnium bound to the TPRV5 protein exhibits mis-

matched symmetry. This means that there is a predominant symmetry which is not followed across all the regions of the protein. In this case, the TPRV5 protein has C4 symmetry but the calumnium is bound off-centred, breaking ensemble's symmetry. As a consequence, it is not easy to distinguish between symmetrical views of the TPRV5. This angular error will be a trend for the rest of the tests conducted with this dataset.



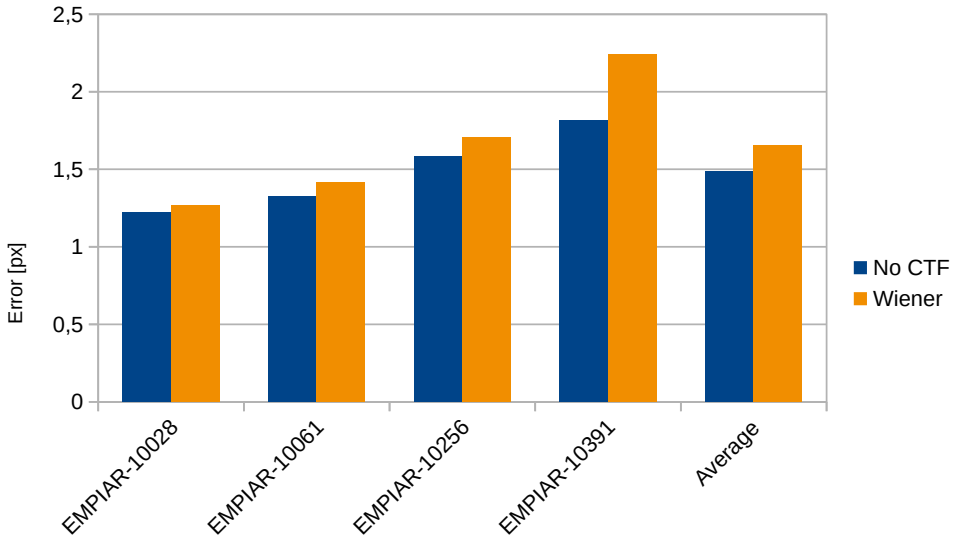
(a) Simulated images



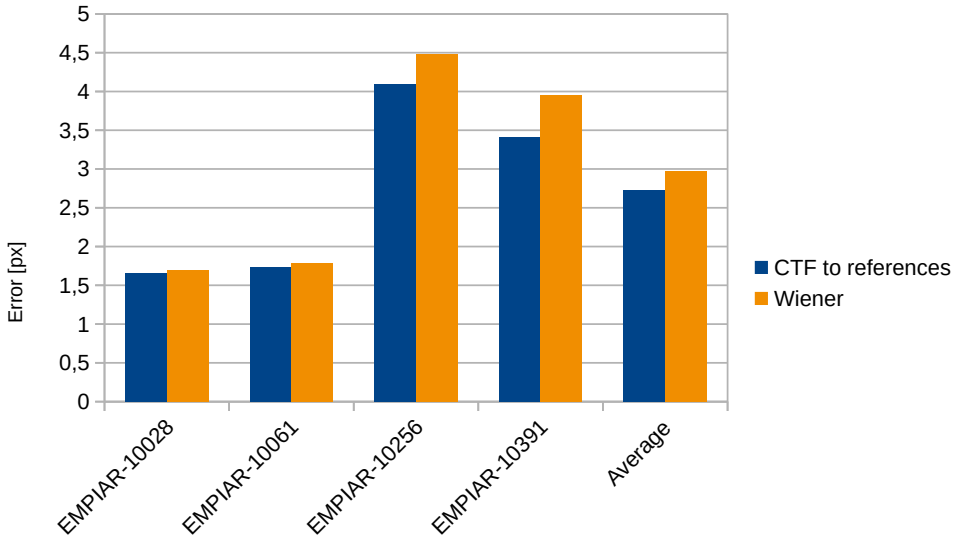
(b) Experimental images

Figure 5.9: Angle accuracy for different compression methods

Regarding the reconstruction resolution of the particles, the results are even better than the previous ones. Comparatively, the resolution degradation associated to the introduction of the Wiener filter is around 3% for experimental images. Furthermore, the high angle assignment errors of the EMPIAR-10256 do not correlate with a loss in resolution, as the



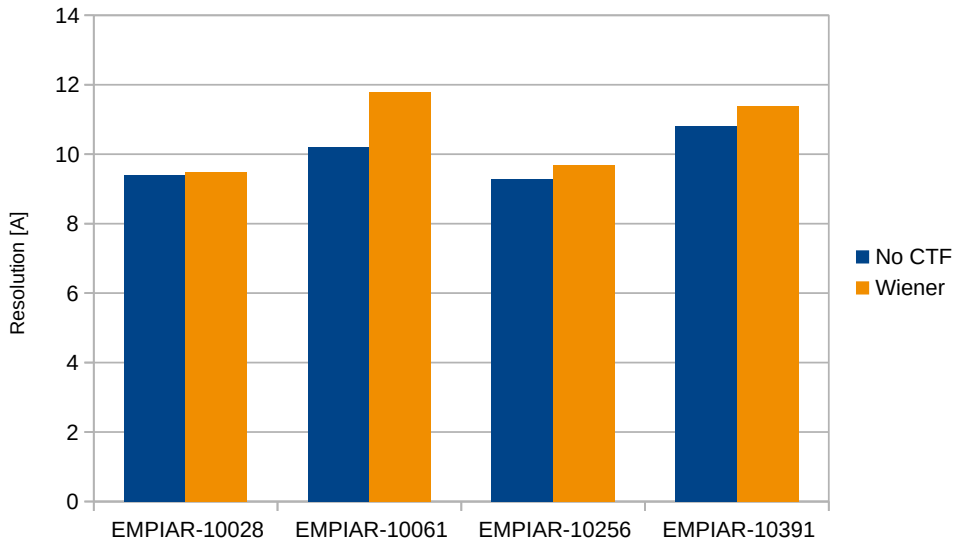
(a) Simulated images



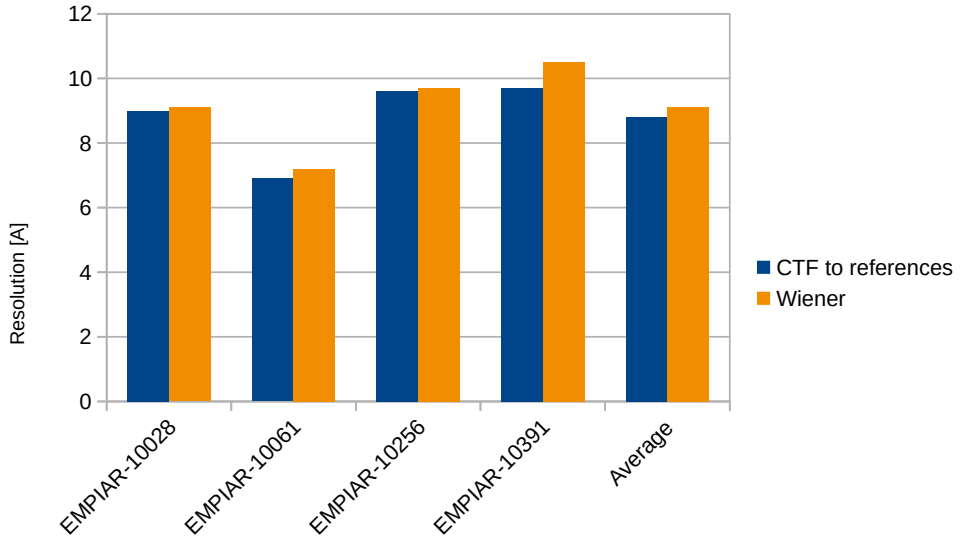
(b) Experimental images

Figure 5.10: Shift accuracy for different compression methods

resolution obtained for this dataset is similar to the one obtained for the rest. This is because the incorrectly assigned views are still highly compatible with the reconstructed volume.



(a) Simulated images



(b) Experimental images

Figure 5.11: Reconstruction resolution for different compression methods

Influence of the vector compression

One of the key innovations of this algorithm is that it performs efficient vector searches using vector compression techniques. The aim of this section is to assess the effectiveness of such compression and evaluate its influence on the quality of the results.

Several compression techniques will be considered in order to select the best compromise between speed and accuracy. Similarly to the previous experiments, the tests will be car-

ried out with a resolution limit of 15Å. For reference, an alignment without any vector compression will also be considered. The raw input vectors have on average around 1300 components, which require 5200B to be stored in 32 bit floating point format.

The first evaluated compression technique will be a PCA. As stated in Chapter 4, this compression technique reduces the dimensionality though a linear projection in such a way that most of the signal energy is kept. In this case, input vectors will be reduced to 128 and 64 components, leading to an approximate compression rate of x10 and x20, respectively. These vectors will have a constant storage cost of 512B and 256B. Additionally the PCA projection matrix needs to be stored, which has a inappreciable size when a large quantity of vectors is used.

Secondly, the IVF-PQ vector compression technique described in Chapter 4 will be tested. In this case, each block of the vector will be quantised into 256 cells, so that a single byte can be used to represent it. The vector will be divided either in 48 or 32 blocks, so that 48B or 32B are needed to store it. In this case, the compression ratios are in the order of x100.

The Figure 5.12 illustrates the storage costs for each of the vector compression techniques. Compressed vectors have a size that is agnostic of the dataset. However, the size of the raw vector depends on the image size and its sampling rate. Therefore, the plot shows bars for each of the datasets, although only the first bar varies across acquisitions. Note that the vertical axis of the graph is logarithmic, suggesting that there is an order of magnitude of difference between not using any compression, using PCA compression and using IVF-PQ compression.

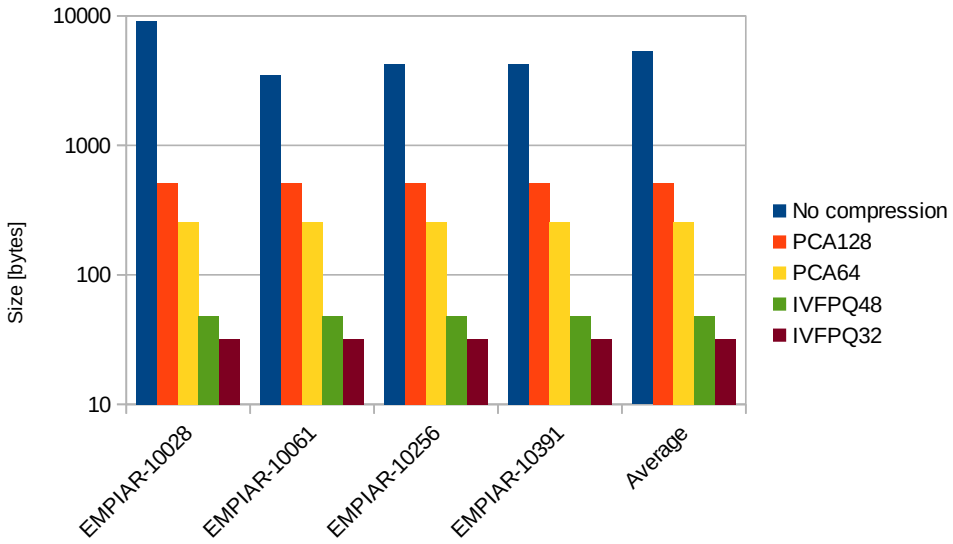


Figure 5.12: Vector storage size comparison between vector compression techniques

Accuracy

Regarding the empirical results, we have been able to measure some degree of accuracy degradation directly related to vector compression. In general, there is correlation between the accuracy loss and the compression ratio employed for storing the vectors. However, the relation is not linear, as the compression ratio raises by orders of magnitude while the decrease in precision is slight. Moreover, the performance increase is considerable.

Figures 5.13 and 5.14 point out that in the case of experimental images, the this degradation is considerably lesser. On average the results of PCA-64 compression are worse than the results of of IVF-PQ compression techniques, which offers a higher compression ratio. In any case, results with simulated images have a more pronounced influence of the compression. Once again, the bad results obtained with EMPIAR-10256 make the average results very bad.

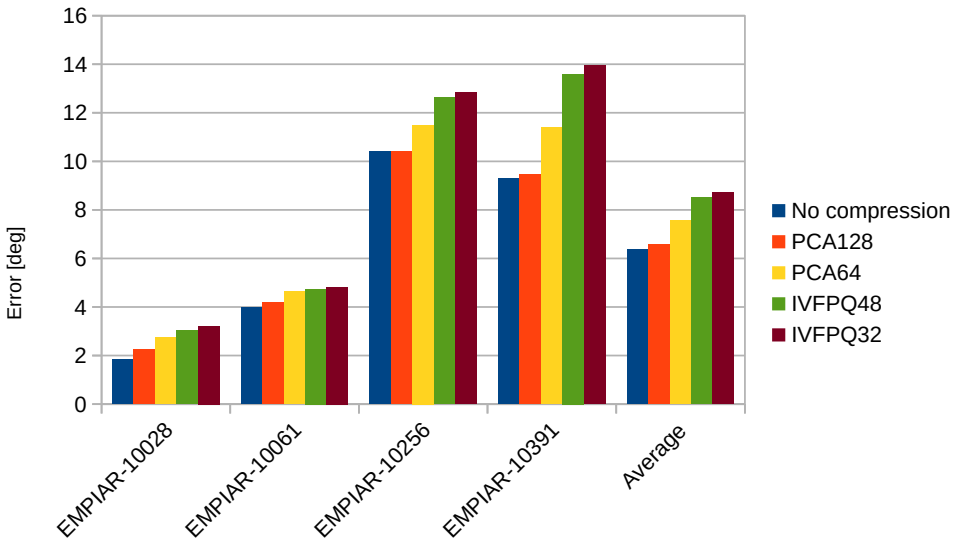
When employing these particles to reconstruct a new volume, the resolutions obtained are shown in Figure 5.15. It can be observed that simulated images produce similar reconstruction resolutions, regardless of the compression method used, but accuracy significantly varies across compression methods. This suggests that the alignment errors of these datasets are induced by very similar views of the volume, as the reconstruction resolution is not affected. Concerning the experimental images, the alignment errors measured earlier correlate with lower reconstruction resolutions.

Performance

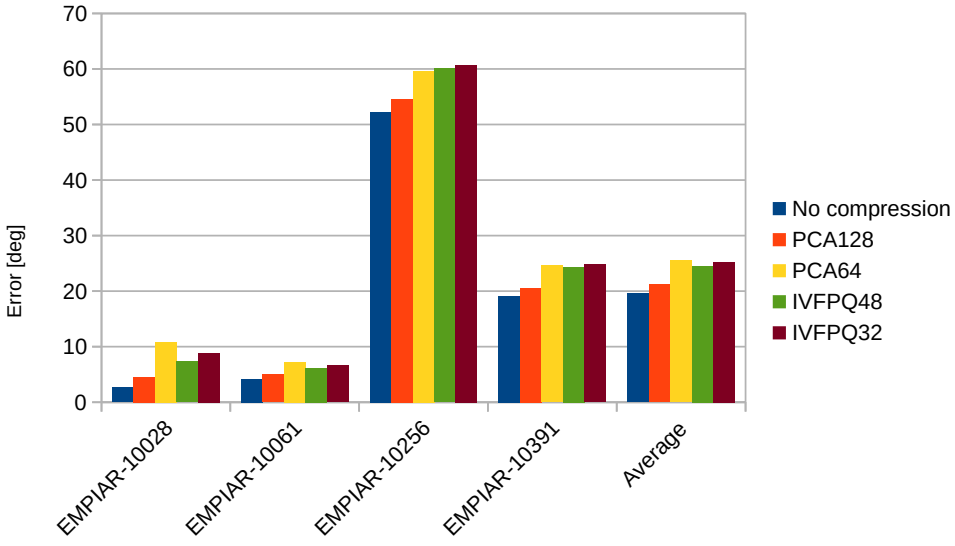
The compression ratio is also highly correlated with the alignment time. Higher compression ratios involve faster memory access times and also quicker distance computations. Thus, the alignment times with vector compression are considerably faster. In this analysis we have distinguished two parts regarding the alignment times. The first one is related to the constant part, which is related to the database training and population processes. This time remains invariant regardless of the number of particles to be aligned. The second part is the time spent on the alignment process itself. This time scales linearly with the number of particles, thus, its measurement is provided as the time per particle (the slope of the linear relationship).

Figure 5.16 proves that the training process required for compression is amortized from the first particle, since the saved time in populating the database compensates this additional step. In any case, the differences in time related to different compression techniques are very subtle. When aligning reasonable amounts of particles, these differences are negligible.

The time required to align individual particles is represented in Figure 5.17, which can be interpreted as the slope of the function representing the total alignment time in terms of

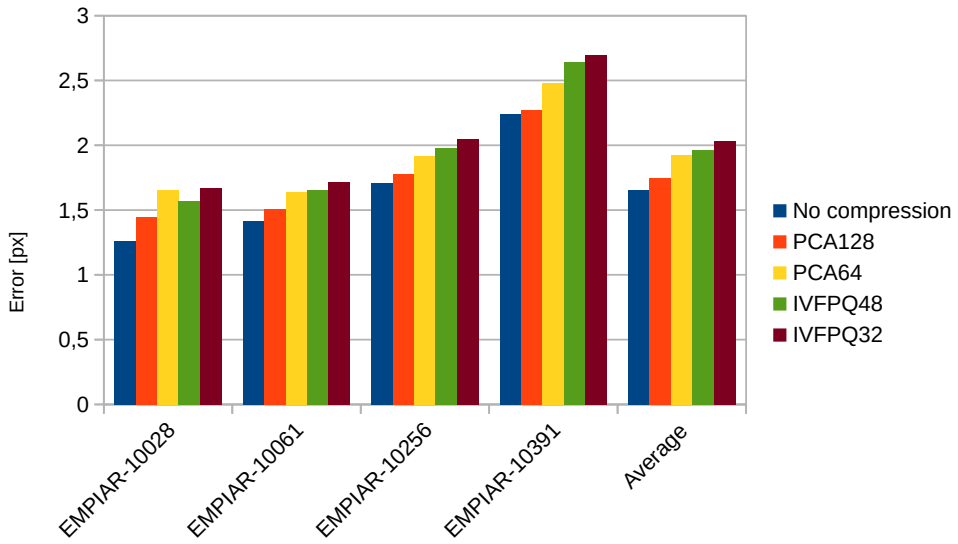


(a) Simulated images

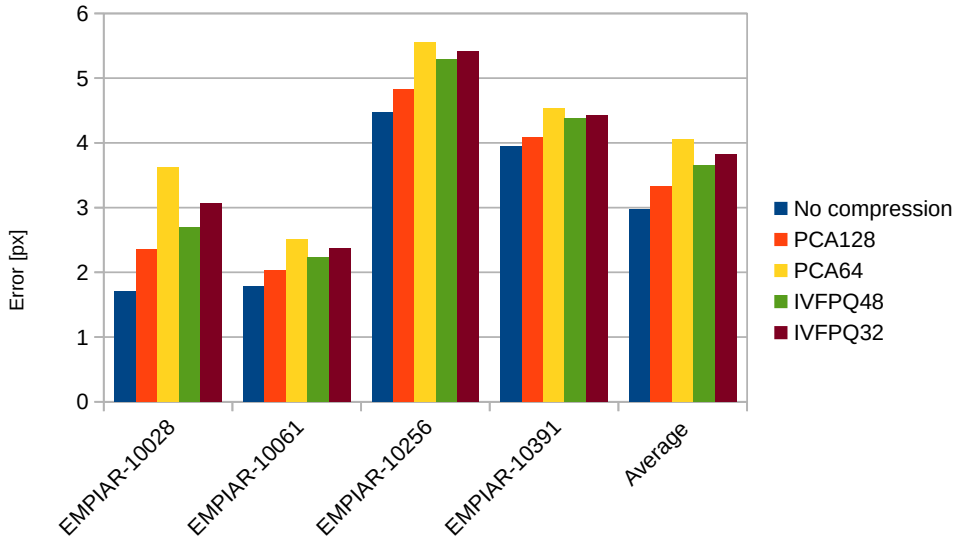


(b) Experimental images

Figure 5.13: Angle accuracy for different vector compression methods

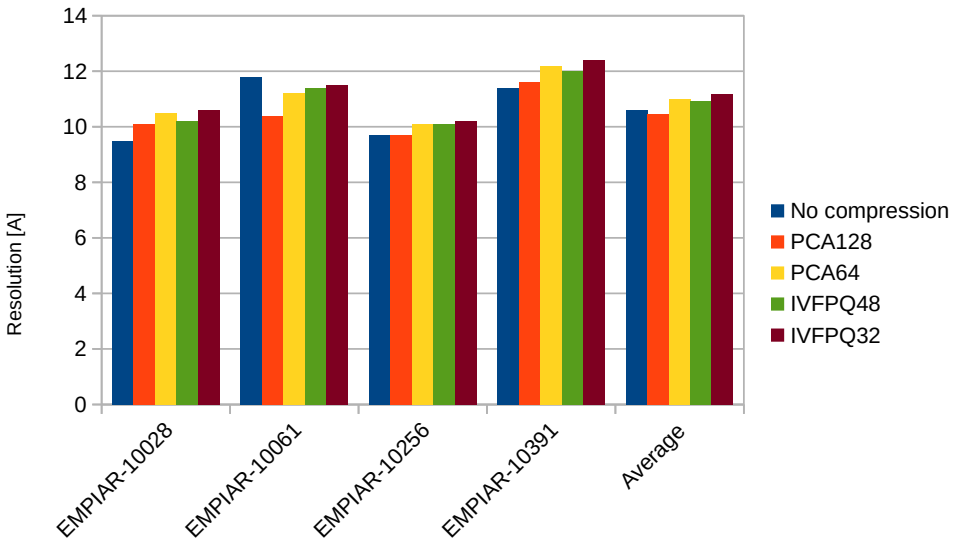


(a) Simulated images

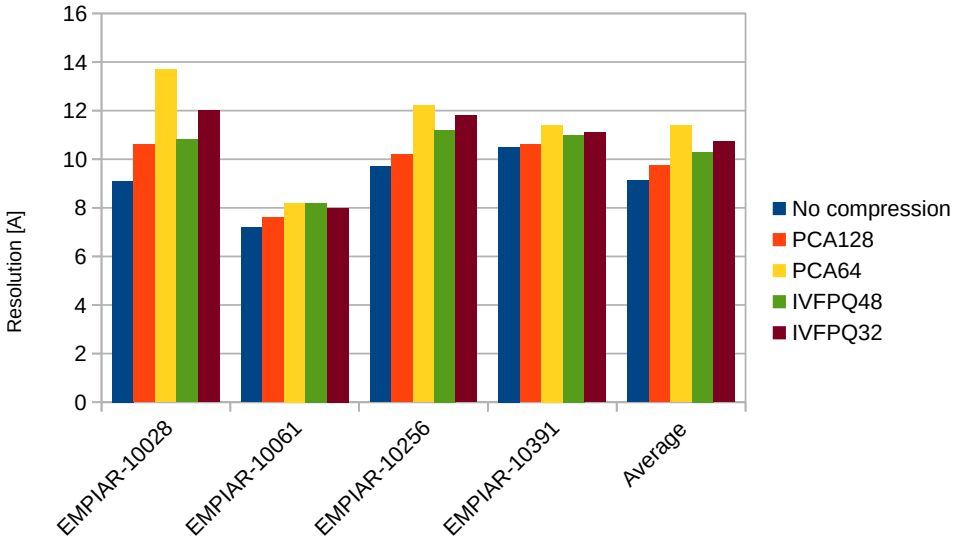


(b) Experimental images

Figure 5.14: Shift accuracy for different vector compression methods



(a) Simulated images



(b) Experimental images

Figure 5.15: Reconstruction resolution for different vector compression methods

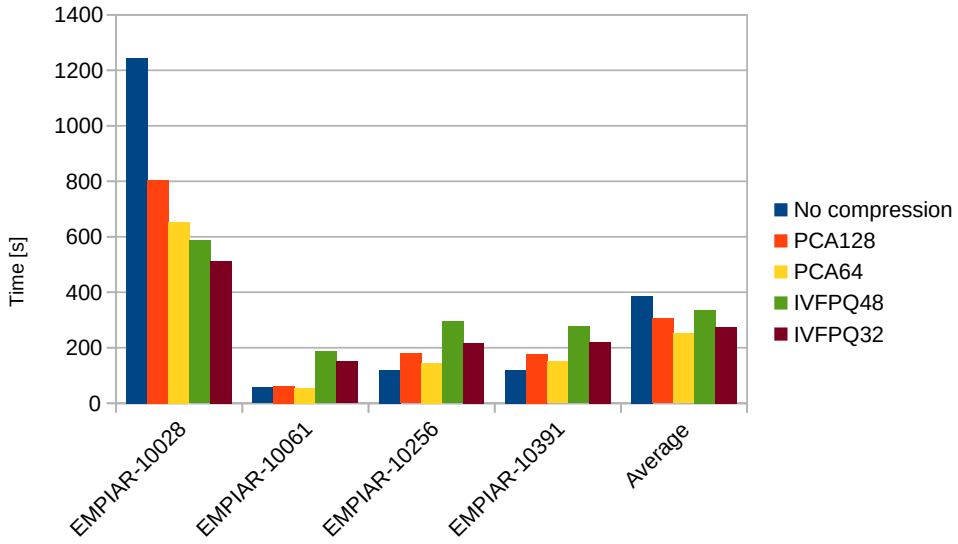


Figure 5.16: Constant time (Training + Populate) for different vector compression methods

the particle count. It is important to note that this graph has vertical logarithmic spacing. Consequently, even slight variations in the graph can result in significant time savings. Considering the significance of this factor for datasets of practical sizes, these numbers serve as the basis for drawing conclusions regarding vector compression techniques.

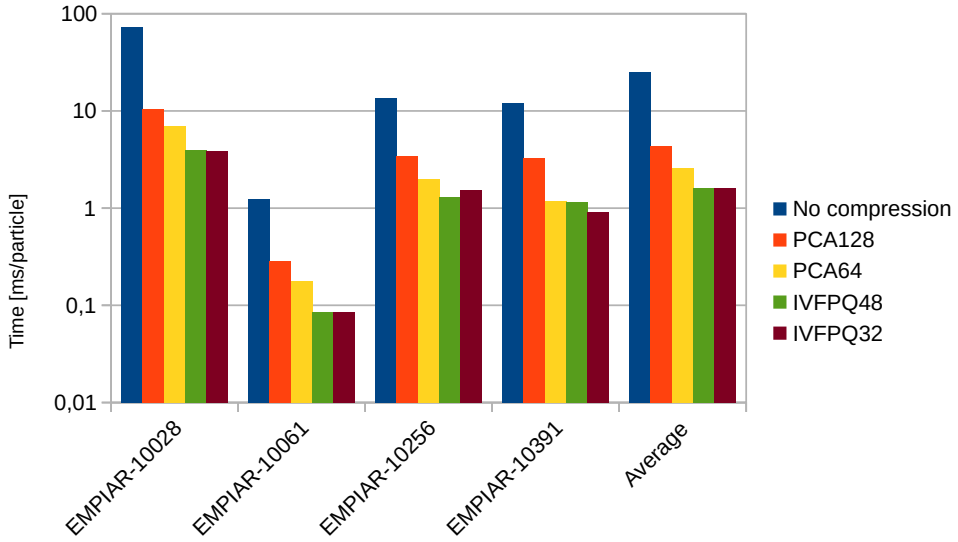


Figure 5.17: Alignment time for different vector compression methods

On average, the usage on any compression method significantly reduces the alignment time. This is specially true for IVF-PQ technique, which obtains a x10 speedup in comparison to the uncompressed vectors. Indeed there is no speedup when varying the number of bytes used with IVF-PQ, so the 48B version should be preferred, as it offers better accuracy results. Similarly, PCA64 provides similar accuracy results but it is slower. As a consequence, PCA64 and IVF-PQ32 compression techniques can be safely dismissed. At the end we have chosen to use the IVF-PQ48 technique, as it is more than twice as fast as the PCA128 option at a

slight accuracy cost. Later we will explore options to trade back this performance gain with better accuracy results.

Influence of the cutoff frequency

Most of the state-of-the art image alignment algorithms compare images only considering low-frequency features of the images. It is widely accepted that information beyond 8Å of resolution is not relevant for the alignment of CryoEM images. In spite of this, all prior tests were done with a resolution limit of 15Å, as this is closer to the operating range of the algorithm. The aim of this section is to explore how the algorithm behaves across different resolution limits, either below and above the previously considered limit.

Accuracy

Figures 5.18, 5.19 and 5.20 show that beyond a resolution limit of 12Å, the alignment accuracy and reconstruction resolution stop improving. This turning point is consistent across all four datasets. However, the actual accuracy value of both angular and shift measurements varies depending on the dataset, likely due to different noise levels.

Curiously, for simulated images the large angular error induced by missmatched symmetry of the EMPIAR-10256 dataset disappears at this resolution limit. However, this effect cannot be observed for experimental images, as it plateaus at a very high alignment error.

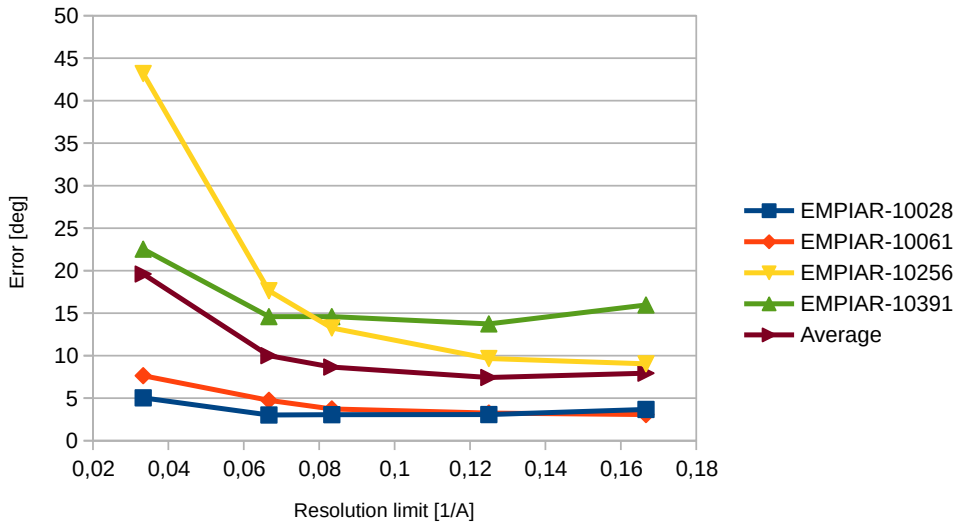
Performance

In contrast to the prior results, the computational cost of the alignment increases rapidly as the resolution limit increases. This is primarily because higher resolution limits involve more reference images. Indeed, we have estimated that the reference count increases with the power of five as the resolution increases.

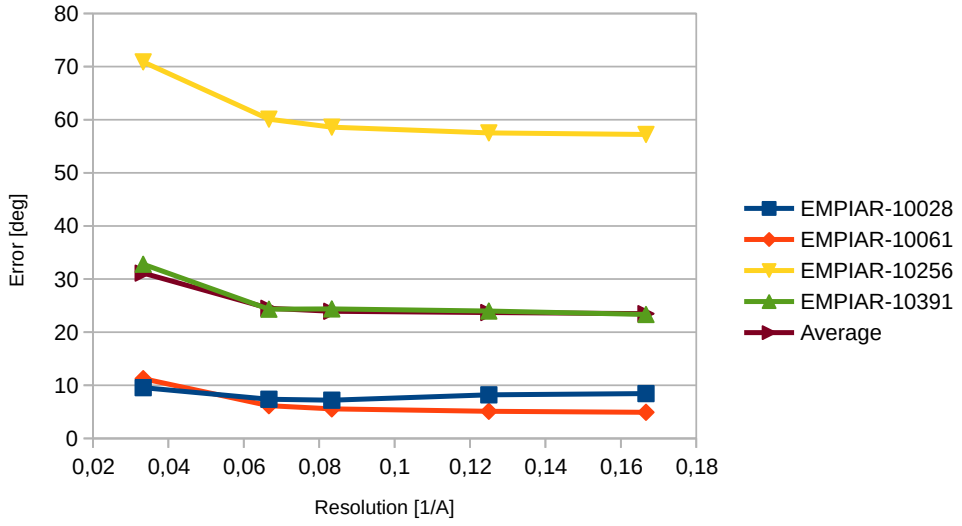
Considering prior measurements, it is important to spare on the resolution limit, as it heavily affects performance but does not provide result improvements beyond a 12Å. In any case, we are researching ways to get closer to the theoretical 8Å limit in a practical way, so that we can obtain better results.

Alignment consensus

In Chapter 4 we described that we can leverage the increased performance of the alignment algorithm to improve its accuracy by consenting the outputs of multiple runs. In this section we would like to analyze the results of this technique. To do so, we have tried repeating the

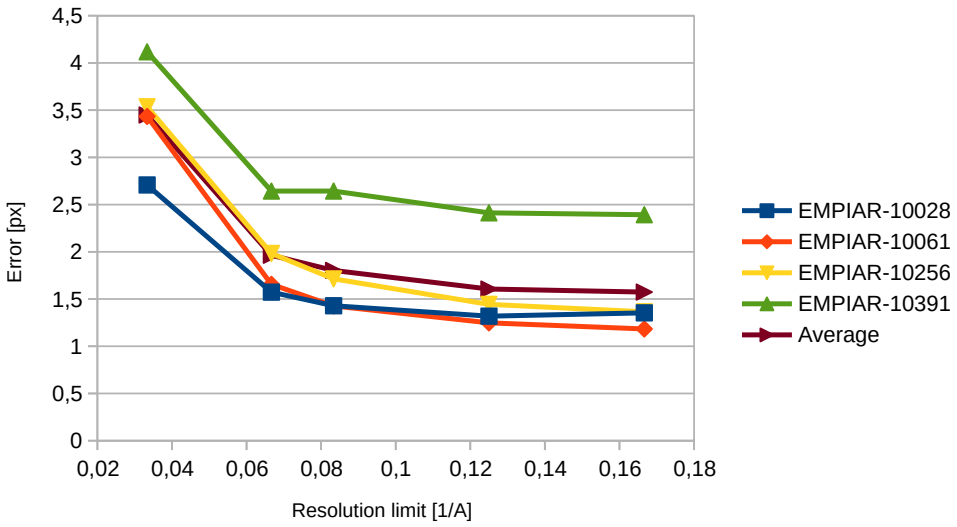


(a) Simulated images

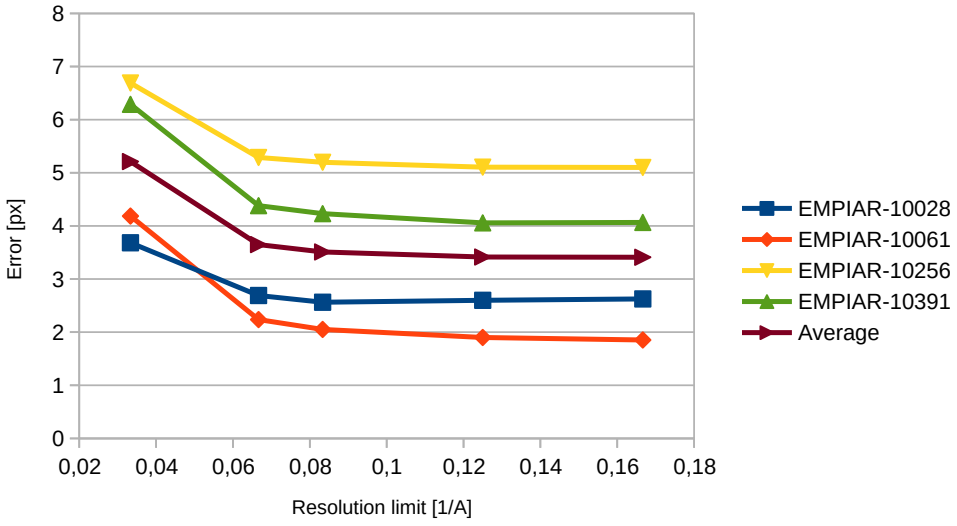


(b) Experimental images

Figure 5.18: Angle accuracy in terms of the alignment resolution limit

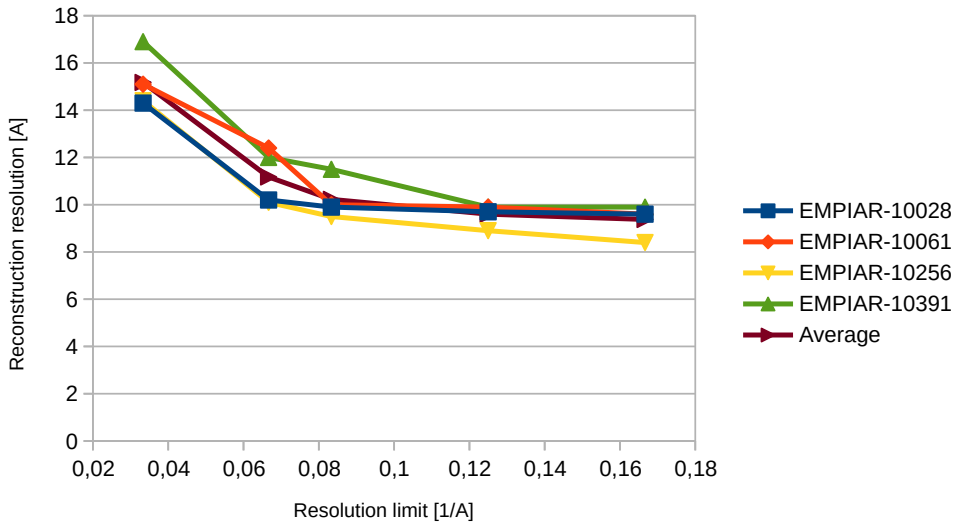


(a) Simulated images

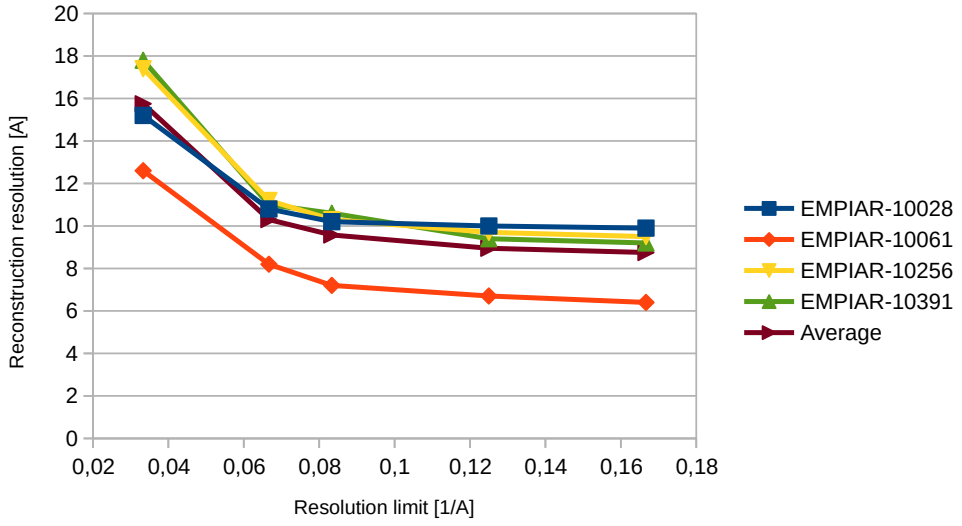


(b) Experimental images

Figure 5.19: Shift accuracy in terms of the alignment resolution limit



(a) Simulated images



(b) Experimental images

Figure 5.20: Reconstruction resolution in terms of the alignment resolution limit

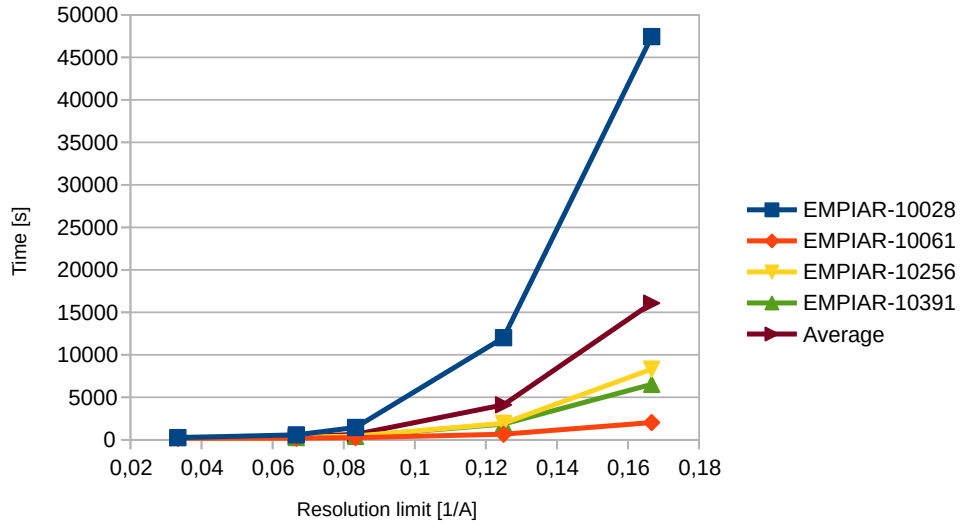


Figure 5.21: Constant time (Training + Populate) in terms of the alignment resolution limit

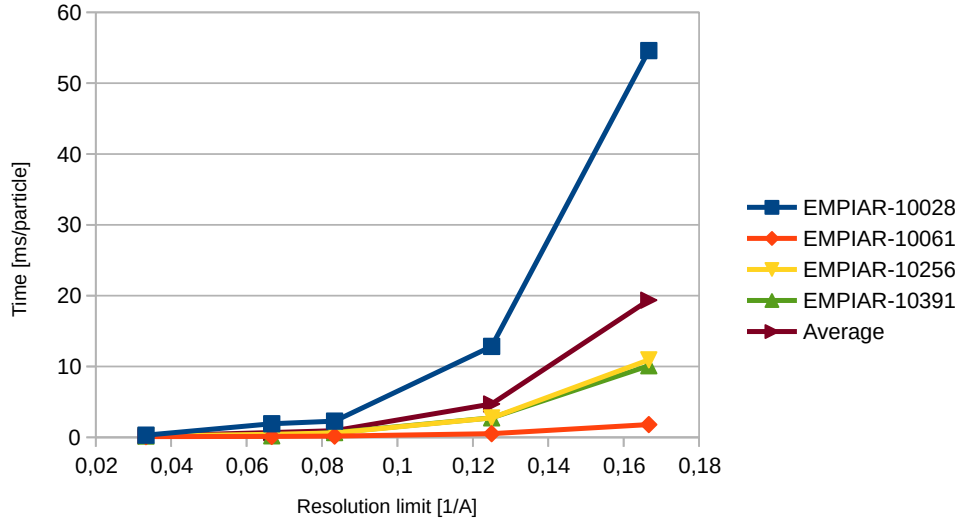
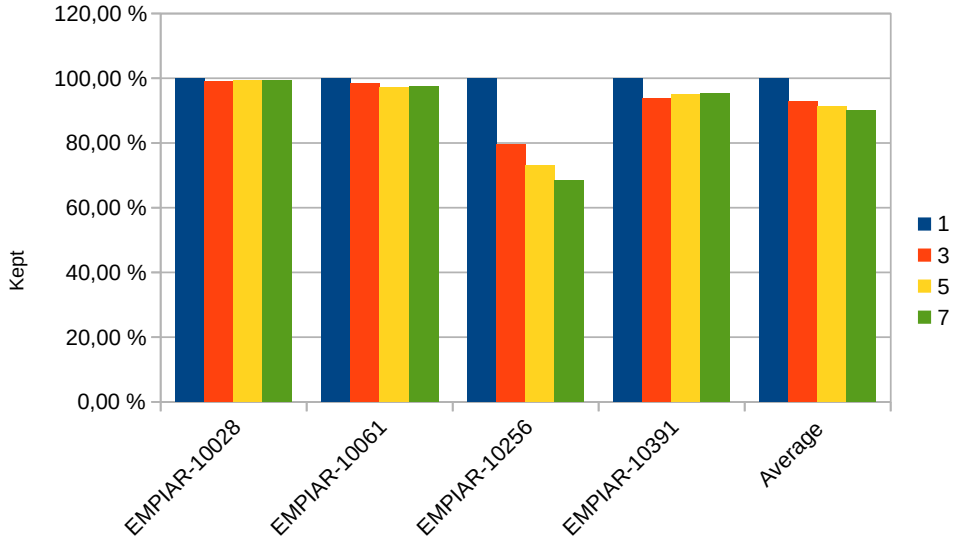


Figure 5.22: Alignment time in terms of the alignment resolution limit

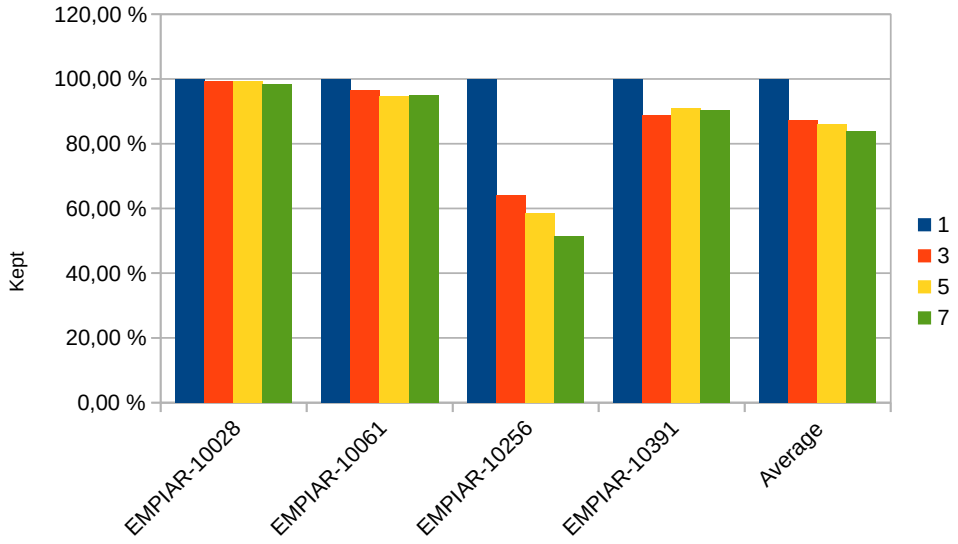
alignment multiple times with slightly different galleries. We expect that good particles produce nearby results consistently, whilst bad particles produce random alignment parameters. Consequently as the number of repetitions increases, the confidence level of the alignment result also increases. The philosophy behind the consensus is that we prefer to use a few images that we have complete confidence in, rather than numerous images that lack our trust.

Regarding the particles that pass this consensus, the drop rate is not very large, meaning that most alignment results are consistent. The only exception to this rule is the EMPIAR-10256 dataset, which only preserves around 60% of images after consensus. This is an expected result, as views separated by 90° around the symmetry axis tend to be very similar. Thus, at this resolution-limit, assignments are practically random, leading to a lack of consensus.

Preserved particles are either top views, side views with clear differences or side views where all assignments coincided by chance.



(a) Simulated images

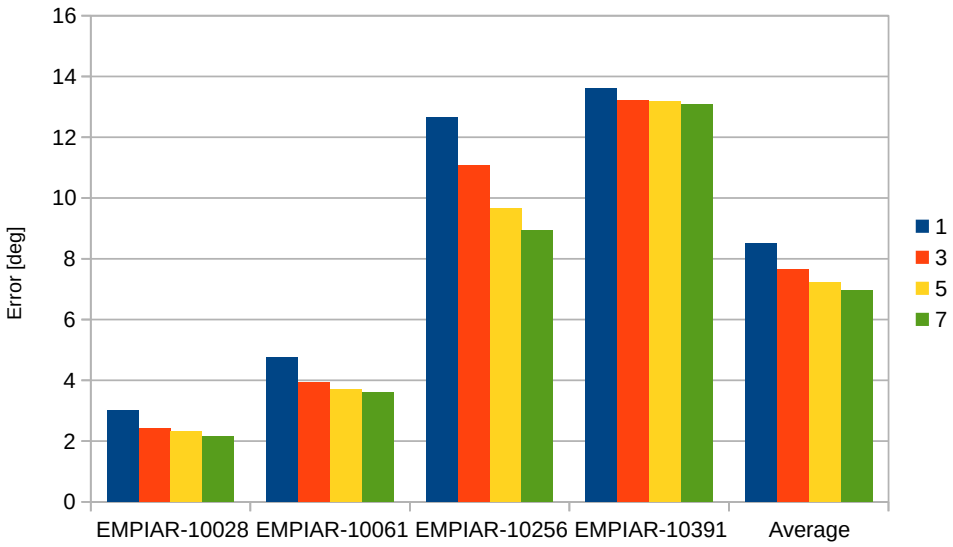


(b) Experimental images

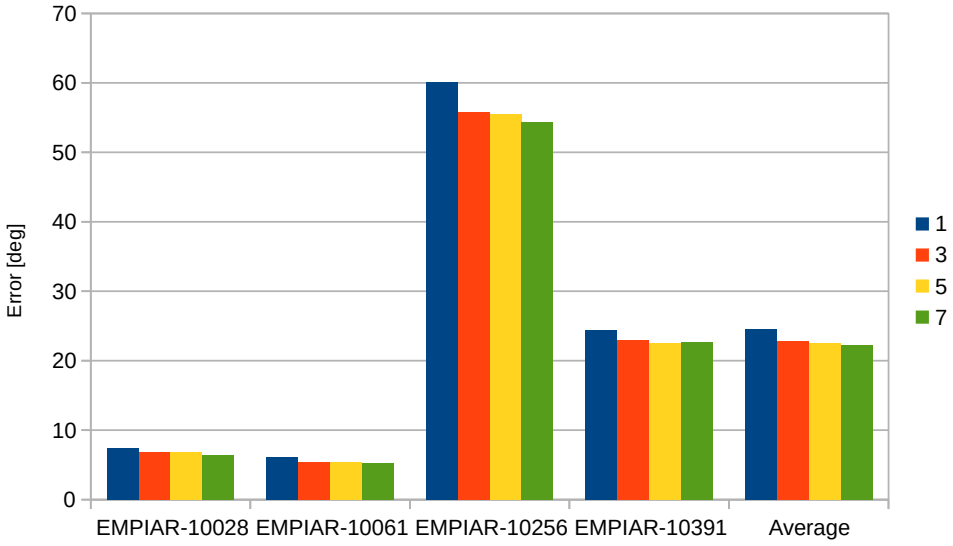
Figure 5.23: Particle dropout ratio for different alignment repetitions

As shown in Figures 5.24 and 5.25, this consensus has a measurable positive outcome on the accuracy of the angle and shift assignments. The most significant improvement is obtained when only 3 repetitions are performed. Further repetitions still help to improve on the results, but accuracy increases are not as significant. Consequently, this technique helps to raise awareness of potential misalignment issues and enhances confidence in the obtained result.

This increased accuracy in alignment results do not always imply an improvement of the

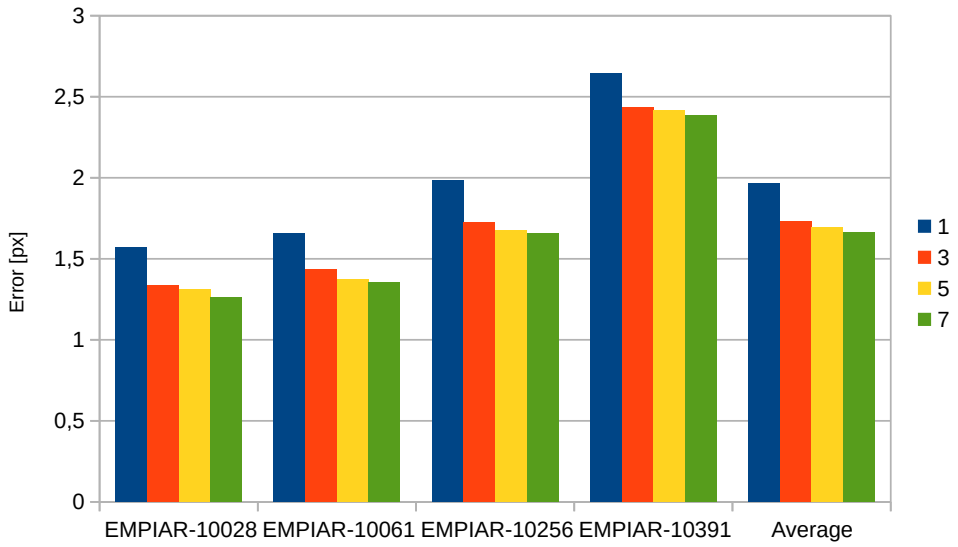


(a) Simulated images

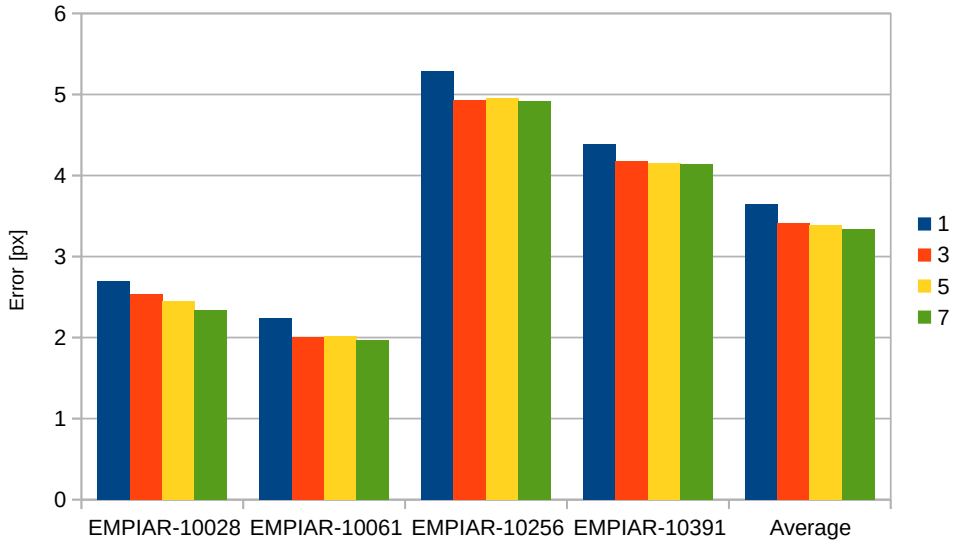


(b) Experimental images

Figure 5.24: Angle accuracy for different alignment repetitions



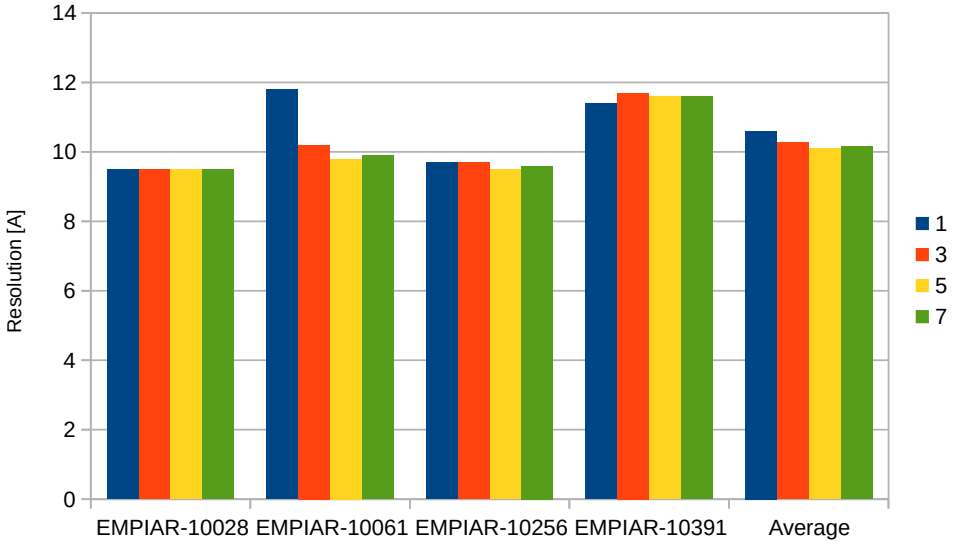
(a) Simulated images



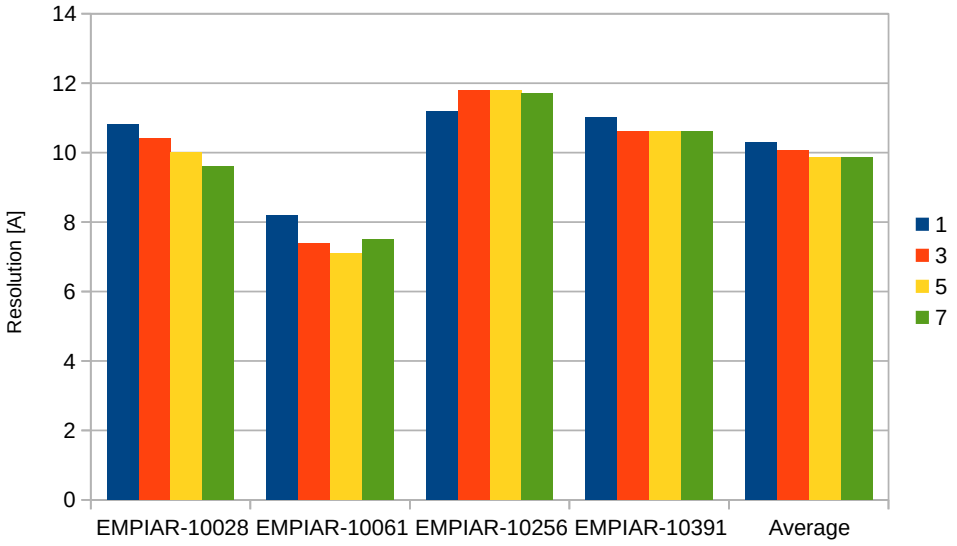
(b) Experimental images

Figure 5.25: Shift accuracy for different alignment repetitions

reconstruction resolution. Indeed, Figure 5.26 shows that in some cases this reduction in the number of particles has a negative effect on the resolution. As described earlier, the alignment errors do not have to be correlated with reconstruction errors, as more often than not, there are many highly compatible views, specially at low resolution. Thus, if a particle "jumps" across several compatible views, we decide that its pose is not conclusive; in spite of the map resolution increase that would involve using it.



(a) Simulated images



(b) Experimental images

Figure 5.26: Reconstruction resolution for different alignment repetitions

Even though the angular consensus does not consistently help to increase the resolution of the reconstructed volume, it provides truthful alignment information to the posterior local alignment. These local alignments are heavily biased by the prior information. Thus, the starting point conditions the capability of local searches to find a global minima. Therefore,

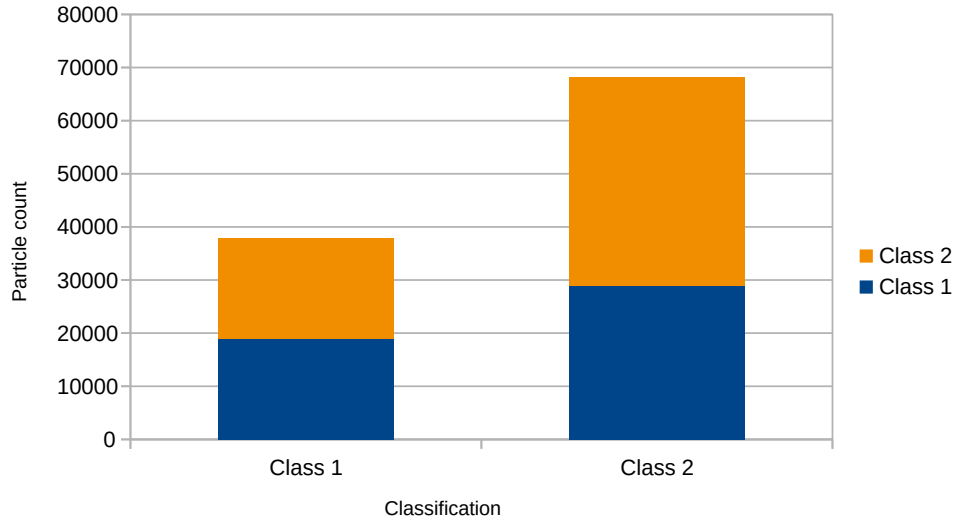
providing them with a reliable initial solution increases the odds of it reaching a local minima or at least reaching a deeper minima. This in turn will help to produce a final volume with a higher resolution.

3D classification

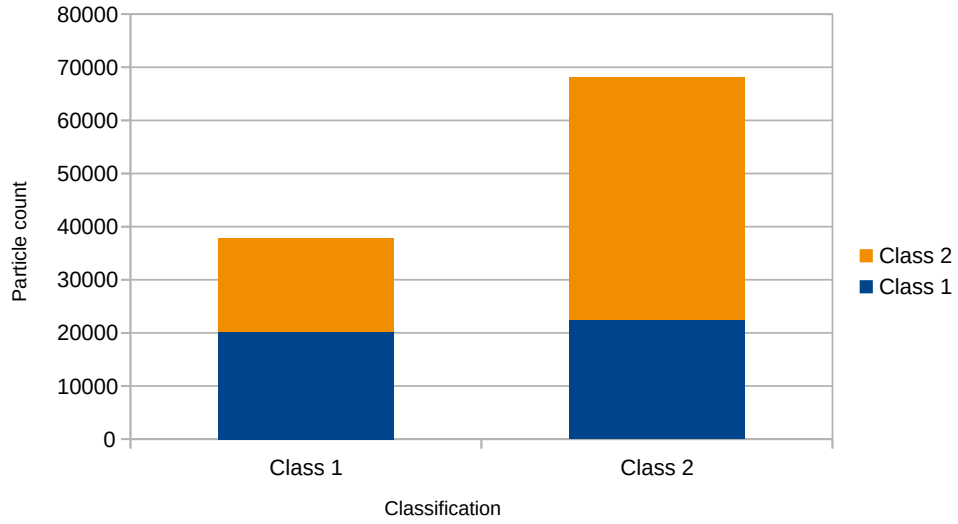
Earlier it was stated that the EMPIAR 10391 dataset exhibits conformational heterogeneity. This means that not all particles come from the same structure. In such cases, the alignment is performed against multiple reference volumes, so that for each particle not only the projection parameters are determined but also the volume that it originated from. In this section, we have used the experimental data of the EMPIAR-10391 dataset to test the effectiveness of our alignment algorithm to separate heterogeneous particles.

We have tested out algorithm using a resolution limit of 15Å and Wiener CTF correction. The results shown in Figure 5.27 proof that the algorithm is able classify particles belonging to the second class, but it messes with the first class (the one without the drug attached), practically assigning images by chance to it. Although these results may not be promising, when compared to a Cryosparc refinement run with the same input data and default parameters, the results are better with our algorithm.

The former results can be potentially improved by performing the classification focusing on ROI defined around the area where the binding appears. This ROI is usually applied by masking the current volumes with a mask designed to leave that region.



(a) Cryosparc



(b) Swiftres

Figure 5.27: Comparison of 3D classifications of the EMPIAR-10391 dataset using Cryosparc and Swiftres

6. Conclusions

In this project, we have developed a CryoEM image algorithm that incorporates vector compression techniques to enhance the analysis of CryoEM images. Our research reveals that the algorithm yields excellent results for low-resolution targets, demonstrating its potential for improving throughput in CryoEM image processing. In spite of this, we observed a decrease in accuracy when applying the algorithm to higher-resolution targets.

The implementation of vector compression allowed for a significant reduction in data storage requirements with little accuracy compromises overall. The compressed vectors retained sufficient information and enabled efficient processing, making them well-suited for low-resolution alignments. This reduction in data size facilitated faster computations and enabled to store larger databases in memory.

Moreover, we have observed that the algorithm is able to resolve 3D heterogeneity even in difficult cases. Similarly, the alignment consensus is able to discard particles with unsure alignments from reconstruction. This allows to keep only images that we are sure about and provide the next steps with high quality alignment estimates.

Despite the positive outcomes observed for low-resolution targets, we observed some limitations in accuracy when requiring higher-resolution targets. The compression algorithm could not reliably retain high-frequency components of the images. Consequently, the results exhibited a plateau in which the accuracy does not increase regardless of the resolution limit used.

To address this challenge and improve the accuracy for higher-resolution targets, further research and development is required. Potential avenues for future exploration include the replacement of the Wiener filter, more effective compression algorithms and the use of weighted comparisons. These solutions will be described in depth in the future work chapter.

Nevertheless, the resolution range limitation does not limit the usage of the algorithm. Typical refinements involve iterative improving the current volume. This algorithm allows to accelerate the first few iterations of this cycle, which do not involve high resolution limits. Indeed, these first iterations are usually quite expensive since they must be global alignments.

Thanks to the alignment consensus, the posterior local alignments will be provided with a very good starting point, reducing the odds of falling into local minimas.

As a consequence, the presented algorithm and its results hold promise for advancing CryoEM and SPA fields. The successful application of vector compression in image alignment, particularly for low-resolution targets, opens up opportunities for more efficient and faster processing of CryoEM acquisitions.

7. Future work

We have observed that our implementation offers exceptional performance in low resolution alignments but struggles to offer superior results for higher resolutions. We would like to focus our future work enhancing the alignment algorithm so that it can offer matchless performance both for low and high resolutions. To do so, we have a couple of ideas that could help to improve the effectiveness at high resolution. Many of the ideas have been already implemented, but still require extensive testing to be conclusive on their results. Additionally, other use cases within the CryoEM context should be explored.

7.1 Weighted distances

The alignment algorithms offers the possibility to compute distances based on some weighting scheme. Other refinement algorithms such as Relion and Cryosparc use a MLE approach which weights each Fourier coefficient with the inverse of the measured noise power (σ^{-2}) for that frequency component.

Even though the weighting is already implemented in the alignment algorithm, we have not been able to determine a proper weighting scheme which improves on the results. Aside from the previously mentioned (σ_N^{-2}) approach, we have also tried using the SNR as a weight source (which also uses the σ_N^2 term in the denominator). The main issue related to them is that we are not able to obtain a reliable estimation of the noise SSNR. Thus, there is some work to be done regarding the determination of the noise model of the dataset and testing the alignment algorithm using weights deduced from this noise model.

7.2 Replacement of the Wiener filter for high resolution

In Chapter 5 we empirically demonstrated that Wiener deconvolution of the CTF is an effective way to tackle the CTF when aligning particles. However, those tests were conducted

at low resolution (15Å). Indeed, the CTF has increasingly more zeros in high frequency. These zeros induce a systematic error on the distance metric, suggesting that this may be one of the causes of problems for high resolution alignments.

Nevertheless, the alignment program offers the possibility of applying the CTF to the reference gallery (instead of correcting it on the experimental images). Although some tests were carried out at high resolution using this alternative approach, the results were not conclusive. Hence, further testing is required to establish a proper approach for tackling the CTF with high resolution alignments. Moreover, this should be done once a reliable weighting scheme has been deduced.

7.3 Local searches

The vector compression techniques described in Chapter 4 can realiably compress a dataset of a couple of million of vectors. In spite of this, the reference dataset size grows rapidly as the resolution increases. Thus, for high resolution alignments the vector quantisation techniques impose a restriction. A possible solution to this issue is to reduce the reference dataset size by performing local alignments.

Assuming that prior information about the alignment of the experimental particles is known, we can reduce our search space to a reduced range around the prior alignment parameters. A first approach to do so would involve only generating in-plane transformations of the reference images in a limited range. Then, each experimental image would be oriented and centred according to its prior alignment information.

Nevertheless, local alignments need to be taken with care. As its name suggests, it searches for local minimas, not global ones. Hence, if the prior information is in a "valley" that leads to a local minima, then the algorithm will not be able to find the global minima. A practical example of this issue would be the EMPIAR-10256 dataset explored in Chapter 5. We observed that at low resolution we are not able to distinguish among the pseudo-symmetrical views of the protein. Hence, if this information were provided as the starting point for a local alignment, this algorithm would not be able to "escape" those incorrectly assigned pseudo-symmetrical views.

7.4 Applications of the image alignment algorithm

The importance of the image alignment in CryoEM image processing has been a lemma in this document. Indeed, image alignment is used in many steps of a typical SPA workflow, such as

the 2D classification, ab-initio volume reconstruction, 3D refinement and 3D classification. In this work we have employed the later two use cases as a playground for our tests. Once alignment algorithm has been perfected, we would like expand its applications to other domains.

Current ab-initio algorithms work by frequently aligning and reconstructing with small batches of the dataset. This means that the reference gallery changes very often, making our alignment algorithm less suitable for this application.

However, the 2D classification problem is a perfect target for our alignment algorithm. These algorithms are very similar to the 3D refinement but instead of using projections of the current volume as the reference gallery, it uses particle averages. Then all the experimental particles are aligned to these averages, and used for updating them.

Another promising use case of this alignment algoritm is Sub-Tomogram Averaging (STA). This would diversify the applications of the algorithm towards the emerging field of CryoET. In essence, CryoET is very similar to CryoEM, but instead of acquiring each spot of the sample grid one, it is acquired reputedly with a varying tilt angle. This has the advantage that it allows directly reconstructing a volume without guessing the angles (as the tilt angles are known). However, the exposure of each tilt angle is considerably low, so the SNR of the images is worse. Additionally, not all possible tilt angles are acquired, producing a missing wedge in Fourier space.

To solve these issues, the STA technique is used, where repetitions of a given structure in the tomogram are aligned and averaged to enhance SNR and resolution. In essence, this problem can be seen as the analogy of the 2D classification for 3 dimensional images (volumes).

The STA problem can be approached from a 3D particle alignment point of view. The 3D alignment is a problem that we have solved for performing 3D refinements. As described earlier, the 3D alignment consists in projecting the current volume from all possible directions to form a reference gallery. Then each experimental image is searched across all the images of this gallery to find a best match. In the case of STA, we would use the projections of each subtomogram (that do not involve the missing wedge) for this alignment. Once all the projections have been angular assigned, we could combine their 3D alignment information alongside the projection restrictions to obtain the alignment of the subtomogram itself.

Bibliography

[1] P. Broadwith. "Explainer: What is cryo-electron microscopy". (Oct. 7, 2017), [Online]. Available: https://www.chemistryworld.com/news/explainer-what-is-cryo-electron-microscopy/3008091.article (visited on 07/29/2022).

[2] "The nobel prize in chemistry 2017". (), [Online]. Available: https://www.nobelprize.org/prizes/chemistry/2017/press-release/ (visited on 05/24/2023).

[3] "The nobel prize in chemistry 1982". (), [Online]. Available: https://www.nobelprize.org/prizes/chemistry/1982/press-release/ (visited on 05/24/2023).

[4] "Cryoem 101". (2022), [Online]. Available: https://cryoem101.org (visited on 07/29/2022).

[5] D. Lyumkis, "Challenges and opportunities in cryo-em single-particle analysis", *Journal of Biological Chemistry*, vol. 294, no. 13, pp. 5181–5197, 2019, ISSN: 0021-9258. DOI: https://doi.org/10.1074/jbc.REV118.005602. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021925820355666.

[6] G. Pintilie. "Cryoem". (2010), [Online]. Available: http://people.csail.mit.edu/gdp/cryoem.html (visited on 11/11/2022).

[7] F. J. Sigworth, "Principles of cryo-EM single-particle image processing", *Microscopy*, vol. 65, no. 1, pp. 57–67, Dec. 2015, ISSN: 2050-5698. DOI: 10.1093/jmicro/dfv370. eprint: https://academic.oup.com/jmicro/article-pdf/65/1/57/7953157/dfv370.pdf. [Online]. Available: https://doi.org/10.1093/jmicro/dfv370.

[8] J. Vargas, A.-L. Álvarez-Cabrera, R. Marabini, J. M. Carazo, and C. O. S. Sorzano, "Efficient initial volume determination from electron microscopy images of single particles", *Bioinformatics*, vol. 30, no. 20, pp. 2891–2898, Jun. 2014, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu404. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/20/2891/17146134/btu404.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu404.

[9] A. Levy, F. Poitevin, J. Martel, *et al.*, *Cryoai: Amortized inference of poses for ab initio reconstruction of 3d molecular volumes from real cryo-em images*, 2022. DOI: 10.48550/ARXIV.2203.08138. [Online]. Available: https://arxiv.org/abs/2203.08138.

[10] A. Razi, J. Ortega, and A. Guarné, "The cryo-em structure of yjeq bound to the 30s subunit suggests a fidelity checkpoint function for this protein in ribosome assembly", *PNAS*, vol. 114, Mar. 2017. DOI: 10.1073/pnas.1618016114.

[11] E. Nogales and S. H. Scheres, "Cryo-em: A unique tool for the visualization of macromolecular complexity", *Molecular Cell*, vol. 58, no. 4, pp. 677–689, 2015, ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2015.02.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1097276515001331.

[12] J. Pfab and D. Si, "Deeptracer: Predicting backbone atomic structure from high resolution cryo-em density maps of protein complexes", *bioRxiv*, 2020. DOI: 10.1101/2020.02.12.946772. eprint: https://www.biorxiv.org/content/early/2020/02/13/2020.02.12.946772.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2020/02/13/2020.02.12.946772.

[13] T. Shaikh, H. Gao, W. Baxter, *et al.*, "Spider image processing for single-particle reconstruction of biological macromolecules from electron micrographs", *Nature protocols*, vol. 3, pp. 1941–74, Feb. 2008. DOI: 10.1038/nprot.2008.156.

[14] S. Ludtke, P. Baldwin, and W. Chiu, "Eman: Semiautomated software for high-resolution single-particle reconstructions", *Journal of structural biology*, vol. 128, pp. 82–97, Jan. 2000. DOI: 10.1006/jsbi.1999.4174.

[15] T. Grant, A. Rohou, and N. Grigorieff, "*cis*tem, user-friendly software for single-particle image processing", *eLife*, vol. 7, E. H. Egelman, Ed., e35383, Mar. 2018, ISSN: 2050-084X. DOI: 10.7554/eLife.35383. [Online]. Available: https://doi.org/10.7554/eLife.35383.

[16] D. Kimanius, L. Dong, G. Sharov, T. Nakane, and S. H. W. Scheres, "New tools for automated cryo-EM single-particle analysis in RELION-4.0", *Biochemical Journal*, vol. 478, no. 24, pp. 4169–4185, Dec. 2021, ISSN: 0264-6021. DOI: 10.1042/BCJ20210708. eprint: https://portlandpress.com/biochemj/article-pdf/478/24/4169/926478/bcj-2021-0708.pdf. [Online]. Available: https://doi.org/10.1042/BCJ20210708.

[17] C. Sorzano, R. Marabini, J. Velázquez-Muriel, *et al.*, "Xmipp: A new generation of an open-source image processing package for electron microscopy", *Journal of Structural Biology*, vol. 148, no. 2, pp. 194–204, 2004, ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2004.06.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047847704001261.

[18] A. Punjani, J. Rubinstein, D. Fleet, and M. Brubaker, "Cryosparc: Algorithms for rapid unsupervised cryo-em structure determination", *Nature Methods*, vol. 14, Feb. 2017. DOI: 10.1038/nmeth.4169.

[19] J. M. de la Rosa-Trevín, A. Quintana, L. del Caño, *et al.*, "Scipion: A software framework toward integration, reproducibility and validation in 3d electron microscopy.", *Journal of structural biology*, vol. 195 1, pp. 93–9, 2016.

[20] D. Střelák, J. Filipovič, A. Jiménez-Moreno, J. M. Carazo, and C. Ó. Sánchez Sorzano, "Flexalign: An accurate and fast algorithm for movie alignment in cryo-electron microscopy", *Electronics*, vol. 9, no. 6, 2020, ISSN: 2079-9292. DOI: 10.3390/electronics9061040. [Online]. Available: https://www.mdpi.com/2079-9292/9/6/1040.

[21] D. del Hoyo Gomez, *Scipion chem*, version 3.0.0, Feb. 1, 2022. [Online]. Available: https://github.com/scipion-chem/scipion-chem.

[22] J. Jiménez de la Morena, P. Conesa, Y. Fonseca, *et al.*, "Scipiontomo: Towards cryo-electron tomography software integration, reproducibility, and validation", *Journal of Structural Biology*, vol. 214, no. 3, p. 107 872, 2022, ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2022.107872. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047847722000429.

[23] D. Strelak, A. Jiménez-Moreno, J. L. Vilas, *et al.*, "Advances in xmipp for cryo–electron microscopy: From xmipp to scipion", *Molecules*, vol. 26, no. 20, 2021, ISSN: 1420-3049. DOI: 10.3390/molecules26206224. [Online]. Available: https://www.mdpi.com/1420-3049/26/20/6224.

[24] C. Sorzano, J. Vargas, J. Otón, *et al.*, "A survey of the use of iterative reconstruction algorithms in electron microscopy", *BioMed Research International*, vol. 2017, pp. 1–17, Sep. 2017. DOI: 10.1155/2017/6482567.

[25] C. O. S. Sorzano, A. Jiménez-Moreno, D. Maluenda, *et al.*, "On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy", *Acta Crystallographica Section D*, vol. 78, no. 4, pp. 410–423, Apr. 2022. DOI: 10.1107/S2059798322001978. [Online]. Available: https://doi.org/10.1107/S2059798322001978.

[26] M. Unser, C. Sorzano, P. Thévenaz, *et al.*, "Spectral signal-to-noise ratio and resolution assessment of 3d reconstructions", *Journal of Structural Biology*, vol. 149, no. 3, pp. 243–255, 2005, ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2004.10.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047847704002047.

[27] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. Society for Industrial and Applied Mathematics, 2001. DOI: 10.1137/1.9780898719277. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9780898719277. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9780898719277.

[28] T. Nikazad, "Algebraic reconstruction methods", 2008.

[29] D. Střelák, C. Ó. S. Sorzano, J. M. Carazo, and J. Filipovič, "A gpu acceleration of 3-d fourier reconstruction in cryo-em", *The International Journal of High Performance Computing Applications*, vol. 33, no. 5, pp. 948–959, 2019. DOI: 10.1177/1094342019832958. eprint: https://doi.org/10.1177/1094342019832958. [Online]. Available: https://doi.org/10.1177/1094342019832958.

[30] D. Herreros, R. R. Lederman, J. Krieger, *et al.*, "Approximating deformation fields for the analysis of continuous heterogeneity of biological macromolecules by 3D Zernike polynomials", *IUCrJ*, vol. 8, no. 6, pp. 992–1005, Nov. 2021. DOI: 10.1107/S2052252521008903. [Online]. Available: https://doi.org/10.1107/S2052252521008903.

[31] S. H. Scheres, M. Valle, R. Nuñez, *et al.*, "Maximum-likelihood multi-reference refinement for electron microscopy images", *Journal of Molecular Biology*, vol. 348, no. 1, pp. 139–149, 2005, ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2005.02.031. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283605001932.

[32] C. Sorzano, J. Vargas, J. Otón, *et al.*, "A review of resolution measures and related aspects in 3d electron microscopy", *Progress in Biophysics and Molecular Biology*, vol. 124, pp. 1–30, 2017, ISSN: 0079-6107. DOI: https://doi.org/10.1016/j.pbiomolbio.2016.09.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0079610716300037.

[33] V. Dubach and A. Guskov, "The resolution in x-ray crystallography and single-particle cryogenic electron microscopy", *Crystals*, vol. 10, Jul. 2020. DOI: 10.3390/cryst10070580.

[34] S. Chen, G. McMullan, A. R. Faruqi, *et al.*, "High-resolution noise substitution to measure overfitting and validate resolution in 3d structure determination by single particle electron cryomicroscopy", *Ultramicroscopy*, vol. 135, pp. 24–35, 2013, ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2013.06.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304399113001472.

[35] J. I. Ronda, *Diapositivas análisis de la señal para comunicaciones*, 2022.

[36] G. Wetzstein, *Lecture notes in computational imaging and display*, Oct. 2018.

[37] N. Grigorieff, "Frealign: High-resolution refinement of single particle structures", *Journal of Structural Biology*, vol. 157, no. 1, pp. 117–125, 2007, Software tools for macromolecular microscopy, ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2006.05.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047847706001699.

[38] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs", *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[39] C. Shalizi, *Lecture notes in principal components analysis*, Apr. 2012. [Online]. Available: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf (visited on 06/06/2023).

[40] I. T. Jolliffe, *Principal Component Analysis* (Springer Series in Statistics), en, 2nd ed. New York, NY: Springer, Dec. 2002.

[41] P. Chang. "Product quantization for similarity search". (May 2022), [Online]. Available: https://towardsdatascience.com/product-quantization-for-similarity-search-2f1f67c5fddd (visited on 03/24/2023).

[42] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011. DOI: 10.1109/TPAMI.2010.57.

[43] D. Sontag, *Lecture notes in clustering*, Oct. 2012. [Online]. Available: https://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf (visited on 03/24/2023).

[44] J. Briggs, *Faiss: The missing manual*. Pinecone. eprint: https://www.pinecone.io/learn/faiss/. [Online]. Available: https://www.pinecone.io/learn/faiss/.

[45] T. Ge, K. He, Q. Ke, and J. Sun, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2946–2953. DOI: 10.1109/CVPR.2013.379.

[46] P. Chang. "Similarity search with ivfpq". (May 2022), [Online]. Available: https://medium.com/towards-data-science/similarity-search-with-ivfpq-9c6348fd4db3 (visited on 03/24/2023).

[47] L. Markley, Y. Cheng, J. Crassidis, and Y. Oshman, "Averaging quaternions", *Journal of Guidance, Control, and Dynamics*, vol. 30, pp. 1193–1196, Jul. 2007. DOI: 10.2514/1.28949.

[48] S. Scheres and S. Chen, "Prevention of overfitting in cryo-em structure determination", *Nature methods*, vol. 9, pp. 853–4, Jul. 2012. DOI: 10.1038/nmeth.2115.

[49] A. Iudin, P. K. Korir, S. Somasundharam, *et al.*, "EMPIAR: the Electron Microscopy Public Image Archive", *Nucleic Acids Research*, vol. 51, no. D1, pp. D1503–D1511, Nov. 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkac1062. eprint: https://academic.oup.com/nar/article-pdf/51/D1/D1503/49645431/gkac1062.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkac1062.

[50] W. Wong, X.-c. Bai, A. Brown, *et al.*, "Cryo-em structure of the *Plasmodium falciparum* 80s ribosome bound to the anti-protozoan drug emetine", *eLife*, vol. 3, W. Kühlbrandt, Ed., e03080, Jun. 2014, ISSN: 2050-084X. DOI: 10.7554/eLife.03080. [Online]. Available: https://doi.org/10.7554/eLife.03080.

[51] W. Wong, X.-c. Bai, A. Brown, *et al.*, *Cryo-EM structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine*, May 2015. DOI: 10.6019/empiar-10028. [Online]. Available: https://doi.org/10.6019/empiar-10028.

[52] A. Bartesaghi, A. Merk, S. Banerjee, *et al.*, "2.2 a resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor", *Science*, vol. 348, no. 6239, pp. 1147–1151, Jun. 2015. DOI: 10.1126/science.aab1576. [Online]. Available: https://doi.org/10.1126/science.aab1576.

[53] A. Bartesaghi, A. Merk, S. Banerjee, *et al.*, *Cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor*, Apr. 2016. DOI: 10.6019/empiar-10061. [Online]. Available: https://doi.org/10.6019/empiar-10061.

[54] S. Dang, M. K. van Goor, D. Asarnow, *et al.*, "Structural insight into trpv5 channel function and modulation", *Proceedings of the National Academy of Sciences*, vol. 116, no. 18, pp. 8869–8878, 2019. DOI: 10.1073/pnas.1820323116. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1820323116. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1820323116.

[55] S. Dang, M. K. van Goor, D. Asarnow, *et al.*, *Cryo-EM structure of trpv5 with calmodulin bound*, Feb. 2019. DOI: 10.6019/empiar-10256. [Online]. Available: https://doi.org/10.6019/empiar-10256.

[56] Y. Z. Tan, L. Zhang, J. Rodrigues, *et al.*, "Cryo-em structures and regulation of arabinofuranosyltransferase aftd from mycobacteria", *Molecular Cell*, vol. 78, no. 4, 683–699.e11, 2020, ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2020.04.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1097276520302562.

[57] Y. Z. Tan, L. Zhang, J. Rodrigues, *et al.*, *Cryo-EM structure of arabinofuranosyltransferase aftd from mycobacteria, mutant r1389s*, May 2020. DOI: 10.6019/empiar-10391. [Online]. Available: https://doi.org/10.6019/empiar-10391.

A. Social, economic, environmental, ethical and professional impacts

A.1 Introduction

This work focuses on accelerating a core problem in the context of CryoEM, which is a very powerful tool for structural biologists to research proteins and their interactions. This information is very valuable for discovering new drugs and vaccines, which in turn enhance the quality of life of the citizens.

Due to the direct relation between the project and the pharmaceutical and biotechnology industry, it is necessary to analyse all the possible ethical social, economic and environmental impacts of the project. The aim of this section is to go over each of these aspects and describe how this algorithm could have an influence on them.

A.2 Description of impacts related to the project

Social impacts

The social impacts of this development are primarily related to science. Research groups may be influenced by this project in the ways stated hereafter.

As mentioned earlier, this algorithm is used in CryoEM image processing, which is a tool used in the research process of new drugs and vaccines. Therefore, enhancing the throughput of

the algorithm can reduce the development time of drugs and vaccines. When these medicines reach the market, they improve on the quality of life and longevity of humans.

Additionally, this development, in the same way as the rest of the Scipion and Xmipp framework, is FOSS. This means that anyone can copy, modify and redistribute the source code. As emphasised by the lemma "Open software accelerates science", this directly helps other developers to improve on the state of the art, as they can continue to work on top of an existing and accessible scaffolding.

Environmental impacts

One of the key focuses of this algorithm is the performance improvement related to the image alignment process. Image alignment is a core problem in CryoEM image processing, so a lot of computation time is spent on running this kind of algorithms. Consequently, any performance improvements lead to a more efficient use of computational resources, drastically decreasing the electrical power consumption.

Economic impacts

As a consequence of the previous points, this project poses an economic impact for research facilities. Firstly, the fact that this project is being developed as a FOSS implies that its usage is free of charge. Secondly, the reduction in power consumption has a direct effect on the power bill. Therefore, this two facts will reduce the operational costs of some research groups.

A.3 Conclusions

To sum up, this project benefits the scientific community in many aspects. Firstly, it may help to reduce development time for drugs and vaccines. Moreover it can have a positive environmental and economic impact by reducing the power consumption associated to the image processing pipeline.

B. Economic budget

This project is estimated to last a semester. During this period, a full-time engineer will be hired to carry out all the software development. The estimated cost associated to this position is 30€/hour, taxes included. Considering that during the span of 6 months 37.5h/week of labour will be dedicated to the project, the engineer will work a total of 900h.

The development of the project will be carried out on a laptop for convenience. This laptop must have enough computational power to run small tests, but the intensive testing will be carried out in a high-end workstation. An amortisation time of 3 years was considered for these electronic devices. Additionally, the developer will be benefited from a paid subscription to GitHub Pro. All these expenses make up for the material resources listed in Table B.1. These prices were accounted with the Value Added Tax (VAT) excluded, as this is accounted separately for the entire project.

This work will take place inside CNB-CSIC facilities. This research centre not only provides office space for the worker, but it also provides a data centre with adequate cooling and power management for our computing equipment. These costs were accounted as indirect costs, which are 15% of the direct costs. CNB-CSIC is a non-profit organisation. Thus, no industrial benefit will be applied to the budget.

Finally, according to the Spanish economic framework, a 21% VAT tax was applied to the subtotal. At the end, the budget for this project totals **THIRTY-NINE THOUSAND SIX HUNDRED NINETY AND TWENTY-ONE HUNDREDTHS EUROS** (39690.21€)

Labour (direct cost)			
Position	Hours	Cost/hour	Cost
Engineer	900	30.00 €	27,000.00 €
Total			27,000.00 €

Material resources (direct cost)				
Item	Purchase prize	Usage time	Amortization time	Cost
Dell Precision 7960 Tower Workstation	5,997.87 €	6 months	36 months	999.65 €
Dell P2415Q Monitor	380.12 €	6 months	36 months	63.35 €
MSI Stealth GS66 Laptop	2,617.96 €	6 months	36 months	436.33 €
Github Pro monthly subscription	4.00 €	6 months	1 months	24.00 €
Total				1,523.33 €

Total direct costs		28,523.33 €
Indirect costs	15 %	4,278.50 €
Budget subtotal		32,801.82 €

VAT	21 %	6,888.38 €

Total budget	39,690.21 €

Table B.1: Budget