UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA DE SISTEMAS ELECTRÓNICOS

TRABAJO DE FIN DE MÁSTER

Development of novel 3D classification methods for Cryo-Electron Microscopy

OIER LAUZIRIKA ZARRABEITIA

2024

MÁSTER UNIVERSITARIO EN INGENIERÍA DE SISTEMAS ELECTRÓNICOS

TRABAJO DE FIN DE MÁSTER

Título:	Development of novel 3D classification methods 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 2024

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA DE SISTEMAS ELECTRÓNICOS

TRABAJO DE FIN DE MÁSTER

Development of novel 3D classification methods for Cryo-Electron Microscopy

OIER LAUZIRIKA ZARRABEITIA

2024

Resumen

La Microscopía Electrónica Criogénica, tambíen conocido por su acrónimo anglicano, CryoEM, se ha convertido en una técnica relevante para obtener imágenes de alta resolución de muestras biológicas tales como las proteínas. Uno de los desafíos clave en CryoEM es la heterogeneidad de las muestras, que se relaciona con que un solo conjunto de datos puede contener múltiples conformaciones o composiciones de las muestras. Esto es especialmente un problema, ya que la mayoría de los algoritmos de procesamiento de imágenes de Análisis de Partículas Aisladas (SPA), se basan en la suposición de que todas las proyecciones se originan a partir de la misma estructura 3D.

La clasificación 3D es un proceso esencial para abordar este problema. Durante el proceso de clasificación 3D, las proyecciones se categorizan según la estructura de la que emanan, de modo que una vez segregadas, el supuesto de homogeneidad es válido. En este trabajo, proponemos un novedoso método de clasificación 3D que aprovecha algoritmos gráficos para mejorar la precisión y eficiencia de implementaciones del estado del arte.

La mayoría de estas implementaciones modernas son víctimas del problema del sesgo de la solución inicial, un problema bien documentado en la literatura científica. En esencia, estas implementaciones refinan de forma iterativa una solución inicial generada aleatoriamente, corriendo el riesgo de caer en mínimos locales. El método propuesto, proporciona de manera determinista una solución inicial donde las clases están separadas al máximo, de modo que las iteraciones posteriores están sesgados hacia la solución correcta. Además, se necesita de un menor número de estas iteraciones hasta converger, lo que disminuye el tiempo total de ejecución.

Para validar la eficacia de este enfoque, se han eligido varios conjuntos de datos experimentales y se han realizado meticulosas pruebas con ellas. Además, se han realizado estos mismos experimentos con soluciones del estado del arte, permitiendonos obtener comparaciones cualitativas. De hecho, los resultados obtenidos respaldan consistentemente las afirmaciones anteriores.

En resumen, en este trabajo presentamos un nuevo enque a la clasificación 3D que demuestra resultados superiores, tanto en términos de rendimiento como de calidad. Creemos que estos avances en la clasificación 3D pueden generar mejores procesos de procesamiento de imágenes para SPA, aumentando la productividad de los biólogos estructurales.

Palabras clave: CryoEM, Análisis de Partículas Aisladas, Clasificación 3D, Teoría de grafos

Abstract

Cryogenic Electron Microscopy (CryoEM) has emerged as a powerful technique for high-resolution imaging of biological samples such as proteins. One of the key challenges in CryoEM is the heterogeneity of the samples, meaning that a single dataset may contain multiple conformations or compositions of the specimens. This is specially a problem, as most of the Single Particle Analysis (SPA) image processing algorithms rely on the assumption that all projections originate from the same 3D structure.

3D classification is an essential process to address this issue. During the process of 3D classification, projections are labeled according to the structure they originate from, such that once segregated, the homogeneity assumption holds true. In this thesis, we propose a novel 3D classification method that leverages graph algorithms to enhance the accuracy and efficiency of state-of-the-art implementations.

Most of these modern implementations fall victim of the initial solution bias problem, a well documented issue in the scientific literature. In essence, these implementations iteratively refine a randomly generated initial solution, running the risk of falling into local minimas.

The method proposed here deterministically provides an initial solution where classes are maximally separated, so that subsequent iterations are biased towards the correct solution. In addition, fewer of these iterations are necessary until convergence, decreasing the overall execution time.

To validate the effectiveness of this approach, several experimental datasets were carefully chosen and employed in comprehensive testing. Moreover, we have performed the same tests with state-of-the-art solutions, aiming to qualitatively assess the benefits of out approach. In fact, results consistently supported the former assertions.

To sum up, in this work we present a new 3D classification algorithm that has proven to obtain superior results, both in terms of quality and performance. We believe that these leaps in 3D classification can incur in better image processing pipelines for SPA, increasing the productivity of structural biologists.

Keywords: CryoEM, Single Particle Analysis, 3D classification, graph theory

Contents

Contents **ix**

List of Figures **x**

List of Tables **xi**

1 Introduction and objectives **1**
1.1 Heterogeneity in CryoEM and 3D classification 2
1.2 Objectives 3
1.3 Structure of the document 5

2 Single Particle Analysis **7**
2.1 SPA image processing steps 9
2.2 Summary of SPA 14

3 State of the art **17**
3.1 SPA image processing software packages 17
3.2 3D classification in SPA 20

4 Implementation **23**
4.1 Initial partition algorithm 23
4.2 Software architecture 34

5 Results **39**
5.1 Test datasets 39
5.2 Experiments 43

6 Conclusions **51**

7 Future work **53**
7.1 Generalization to multiple classes 53
7.2 Performance improvements 53
7.3 3D classification refinement 54

7.4 Classification consensus 54

Bibliography **55**

A Social, economic, environmental, ethical and professional impacts **61**
A.1 Introduction 61
A.2 Description of impacts related to the project 61
A.3 Conclusions 62

B Economic budget **63**

List of Figures

1.1 SPA image acquisition and structure reconstruction 2
1.2 Example of a 3D classification 4

2.1 SPA workflow 8
2.2 CTF examples 10
2.3 Example of a picked micrograph 11
2.4 Example of 2D classification 11
2.5 30S ribosome with a binding 12
2.6 Typical refinement cycle 13
2.7 Fourier Slice Theorem illustration for 3D 14
2.8 Example of model building 15

3.1 Scipion package usage statistics by type 19

4.1 Overview of the initial 3D classification algorithm 24
4.2 Particle grouping with cones 25
4.3 Representation of the Euler angle convention used in CryoEM 26
4.4 In-plane transformation of the particles 27
4.5 Eigen-image computation process 29
4.6 Example of a Gaussian Mixture Model of two components 30
4.7 Graph embedded on the projection sphere 32
4.8 Global classification 35
4.9 Snapshot of the classification viewer 37

4.10 Snapshot of the 3D graph viewer 38

5.1 TRPV-5 reconstruction 40
5.2 Sample of the TRPV-5 particles 41
5.3 Pre-cathalytic spliceosome reconstruction 42
5.4 Sample of the spliceosome particles 42
5.5 HER-2 reconstruction 43
5.6 Sample of the HER-2 particles 44
5.7 Class distribution across Relion's EM iterations with the TRPV-5 dataset . . . 45
5.8 Slice 127 of the reconstructed volumes of TRPV-5 after classification 47
5.9 Classification experiments with the spliceosome dataset focusing on the helicase subunit 48
5.10 Classification experiments with the spliceosome dataset focusing on the SF3b subunit 49
5.11 Classification experiments with the HER-2 dataset 50
5.12 Execution time comparison 50

List of Tables

B.1 Budget 64

Glossary

BCU Biocomputing Unit. 2, 17

CNB Centro Nacional de Biotecnología. 2, 17, 63

CPU Central Processing Unit. 44

CryoEM Cryogenic Electron Microscopy. vii, 1, 2, 7, 17, 20–22, 35, 39, 51, 54, 61, 62

CryoET Cryogenic Electron Tomography. 17, 18

CSIC Consejo Superior de Investigaciones Científicas. 2, 17, 63

CTF Contrast Transfer Function. 9, 10, 23, 25

CUDA Compute Unified Device Architecture. 18

DNN Deep Neural Network. 22

EM Expectation Maximisation. 3, 7, 20–23, 43–45, 51, 54

EMPIAR Electron Microscopy Public Image Archive. 39–41

FFT Fast Fourier Transform. 14

FOSS Free and Open Source Software. 62

FT Fourier Transform. 13, 14

GMM Gaussian Mixture Model. 29, 30, 37, 38

GPU Graphics Processing Unit. 17, 18, 44, 46, 53

GUI Graphical User Interface. 35, 37

I/O Input/Output. 53

IFT Inverse Fourier Transform. 14

ML Machine Learining. 9, 10, 12

PCA Principal Component Analysis. 21, 22, 28–30, 37, 38

PSD Power Spectral Density. 9

ROI Region of Interest. 20, 28, 35, 42

SDP Semi Definite Programming. 33, 34

SGD Stochastic Gradient Descent. 11, 12

SNR Signal to Noise Ratio. 1, 7, 9, 10, 20, 21

SPA Single Particle Analysis. vii, 1, 7, 9, 14, 17, 18, 27

TEM Transmission Electron Microscope. 1, 7, 9

VAT Value Added Tax. 63

1. Introduction and objectives

Cryogenic Electron Microscopy (CryoEM) is a novel imaging technique that involves Transmission Electron Microscopes (TEMs) to analyze frozen samples. Differing from conventional optical microscopes, TEMs employ an electron beam instead of light, enabling them to capture of images at significantly higher resolutions. Consequently, CryoEM has gained popularity for analyzing biological molecules such as proteins and viruses[1]. In this regard, Single Particle Analysis (SPA) constitutes a set image acquisition and processing techniques that facilitates such a task. In this process, a significant amount of two-dimensional images are utilized to elucidate the three-dimensional structure of the specimen under study.

However, TEMs subject the sample to very extreme conditions, such as near perfect vacuum and high-energy electrons. Thus, the sample is frozen before entering the microscope. This helps to maintain it intact when exposed to such environment. The sample freezing is performed in a matter of milliseconds, in a process known as plunge-freezing. This process avoids the formation of ice crystals, which would diffract the electron beam. This technique was awarded with the 2017 Nobel Prize in Chemistry[1][2].

In SPA, thousands of samples are spread on a copper or gold grid, each of them holding a random orientation. Individually referred to as "particles", a extensive amount of 2D projections can be used to mathematically infer the 3D structure of the specimen[3]. This process is hindered by many artifacts such as the poor Signal to Noise Ratio (SNR) present in CryoEM images, which is in the order of 1/100, this is, noise is much more prominent than the actual signal.

The mathematical models used for reconstruction assume that all projections originate from the same 3D structure. Nevertheless, this does not hold true in many real world cases. In instances where a dataset comprises diverse structures, known as heterogeneous, projections must be categorized based on their corresponding structures, which are unknown. This procedure, known as 3D classification, forms the central theme of this thesis.

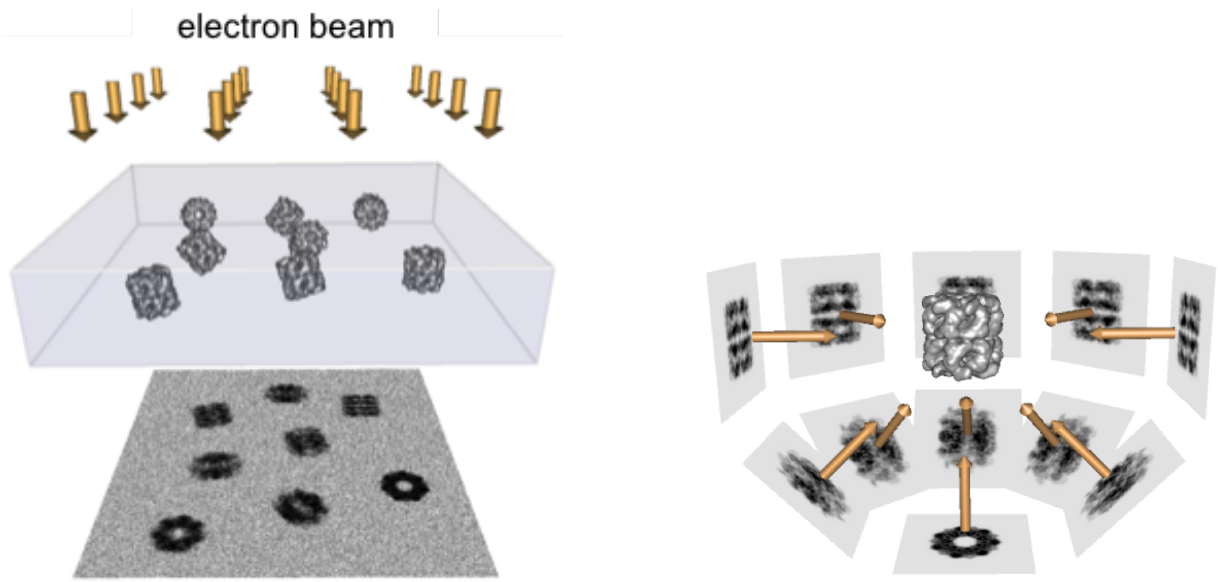
This End of Masters Thesis is the author's second work on CryoEM. Previously, on July 2023, he presented a work on fast image alignment algorithms[5]. Although this project shares

the context (CryoEM) with the previous one, the actual problem is completely different. Nevertheless, some introductory sections were borrowed from the preceding publication.

This project was conducted within the Biocomputing Unit (BCU) research group, situated at the Centro Nacional de Biotecnología (CNB) under the Consejo Superior de Investigaciones Científicas (CSIC). This research group is actively involved in the development of two software suites related to CryoEM, namely Xmipp and Scipion. Xmipp focuses on implementing image processing algorithms, while Scipion serves as a framework facilitating seamless interoperability among cutting-edge image processing suites. Consequently, the software developed in this project is incorporated into Xmipp and integrated into Scipion.

1.1 Heterogeneity in CryoEM and 3D classification

Many datasets obtained through CryoEM demonstrate heterogeneity, indicating that a single 3D structure cannot be attributed to the acquisition. Two potential reasons account for this diversity. Firstly, the studied specimen might exhibit flexibility, resulting in projections originating from distinct protein states. This phenomenon is referred as conformational heterogeneity. Secondly, the images may involve a drug binding experiment, with some projections feature a small attached drug while others do not. This last case is known as compositional heterogeneity.



(a) Image acquisition (b) Reconstruction

Images obtained from: [4]

Figure 1.1: SPA image acquisition and structure reconstruction

Regardless of the heterogeneity type, 2D projections need to be categorized according to the 3D structure they belong to, so that each of the classes can be used to reconstruct a disctinct volume homogeneously. This projection classification problem is known as 3D classification. The main difficulty of this process is that the actual variations in the structures are not known. In other words, images need to be segregated according to a criteria that it is not known yet, implying that the partition must be data-driven. This is further hampered by the fact that variations between structures are very subtle and the signal to noise ratio in the data is extremely low. The process is illustrated in Figure 1.2.

Nevertheless, the 3D classification step is provided with some ancillary parameter estimations. For instance, the projection parameters have been estimated by the previous steps in the image processing pipeline. With these parameters, a mixture of the unknown structures can be reconstructed, known as "consensus volume".

1.2 Objectives

Most of the state-of-the-art 3D classification solutions take an iterative approach through Expectation Maximisation (EM). On each iteration, particle projections are compared to multiple structures to find the best fit. Then, these structures are reconstructed with the particles that have been assigned to them.

However, these algorithms need to be provided with a initial solution, which is typically randomly generated. At the same time, it is very well documented that this initial solution will introduce a bias to the EM algorithm[6][7]. This is attributed to the fact that the EM tends to converge to a local minima around the initial solution[8]. Given the random nature of the initial solution, there is a risk that the algorithm may fail to converge to the correct solution.

Moreover, the EM iterations are computationally very expensive, as they involve many image to volume comparisons. As a consequence, the 3D classification algorithms also take a long time to complete.

On this thesis we intend to develop a novel 3D classification method that swiftly provides a solid initial solution. This has two implications: Firstly, subsequent EM iterations will be biased towards the correct solution. Secondly, due to the quality of the initial solution, less EM iterations are necessary until convergence, reducing the total time required for the 3D classification process.

This algorithm has direct implications in the field of structural biology, as researches will be able to obtain quicker and more reliable outcomes from their experiments. This in turn has direct implications on the pace of drug and vaccine discoveries.

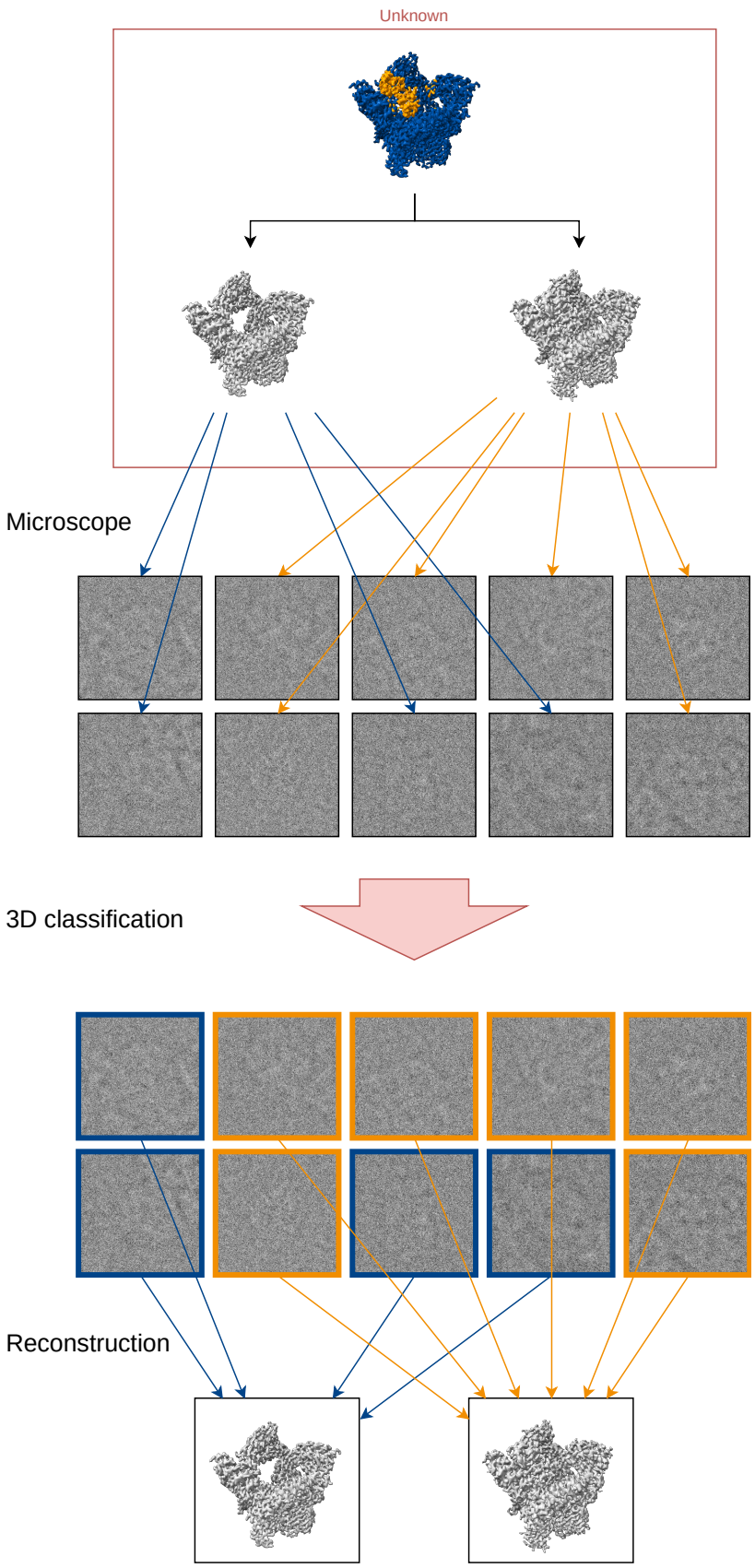


Figure 1.2: Example of a 3D classification

1.3 Structure of the document

The thesis is organized into seven chapters. The first chapter, Introduction and objectives, lays the groundwork by providing an overview of the 3D classification problem, pointing out its complexities and challenges. Following this, the second chapter, Single Particle Analysis, presents a high-level look at the image processing pipeline employed in SPA, setting the stage for subsequent discussions. In the third chapter, State of the art, various methods for tackling the 3D classification problem are examined, highlighting their strengths and limitations. The fourth chapter, Implementation, delves into the specific approach adopted to address the problem, offering a detailed description of the methodology employed. Chapter five, Results, follows providing a meticulous account of the experiments conducted and their corresponding outcomes. The document then progresses to the sixth chapter, Conclusions, where findings are synthesized, and implications are discussed. Finally, in the seventh chapter, Future work, potential avenues for further research and development are explored.

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[9]. 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. Nevertheless, the reconstruction process involves several challenges, as the input images have very poor SNR and other artefacts.

The studied sample is prepared on a copper of gold grid, which is inserted into the 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 2.1 and it will be detailed hereafter.

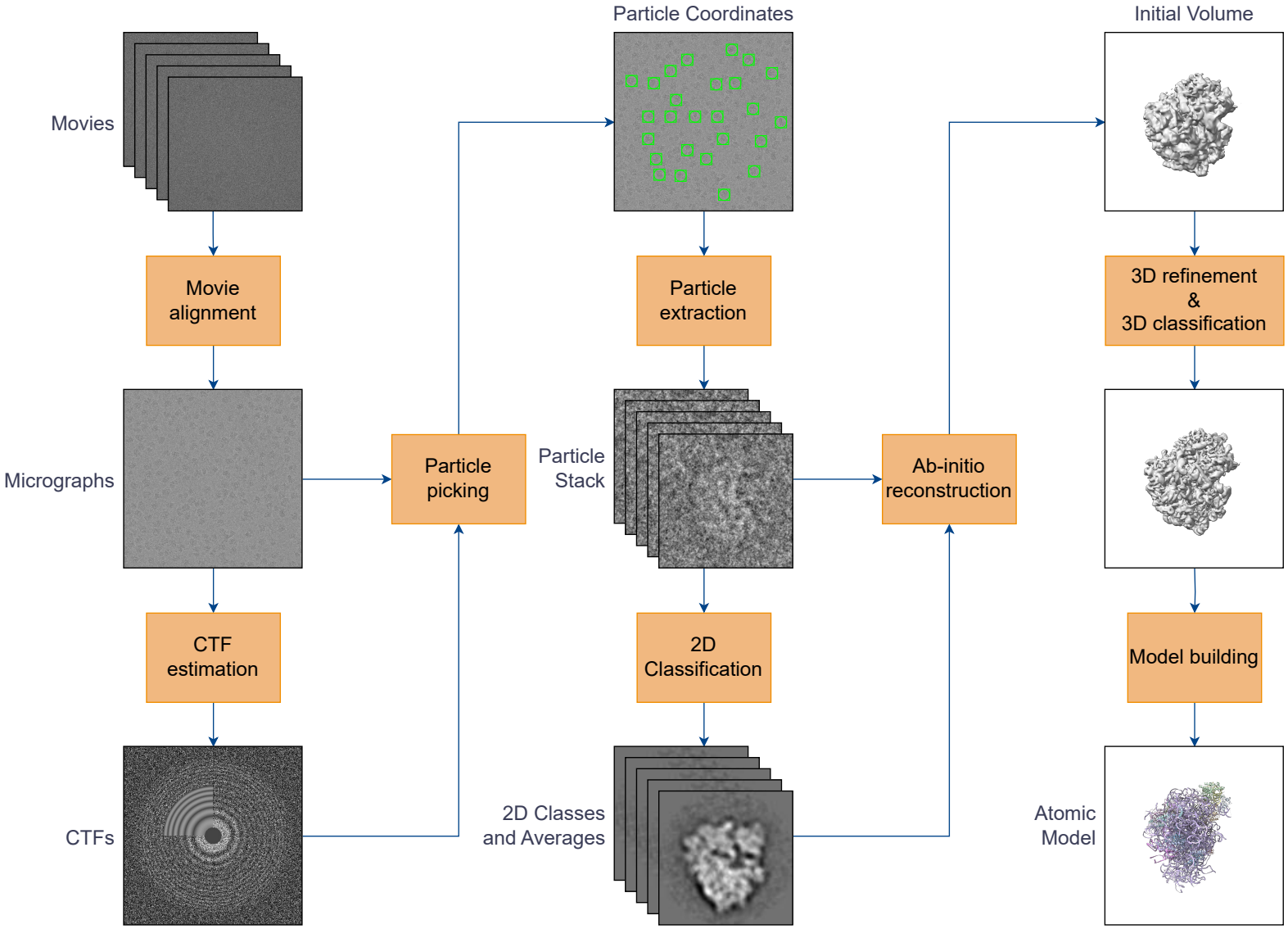


Figure 2.1: SPA workflow

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[3].

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.

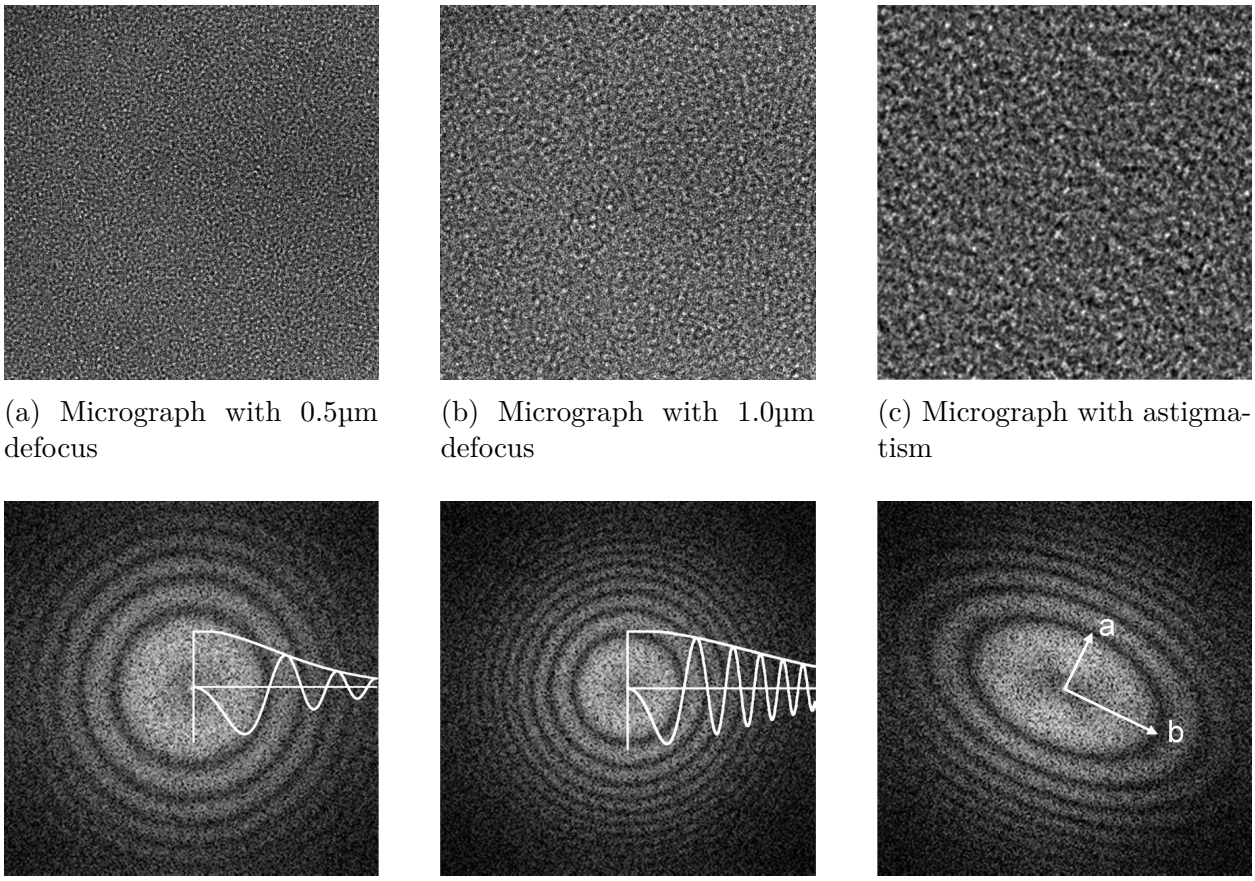
CTF estimation

TEMs do not have a planar frequency response. Instead, they "colour" the images in frequency space that causes a characteristic Power Spectral Density (PSD) pattern known as Thon rings. This transfer function has a sinusoidal appearance, with decreasing periodicity and a overall tendency to attenuate higher frequencies[3]. In addition, the rings may have elliptical shape, being wider in some axis. This is known as astigmatism. A example of a TEM Contrast Transfer Function (CTF) is shown in the Figure 2.2.

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[3]. The characterisation of the CTF is accomplished by calculating the 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: [3]

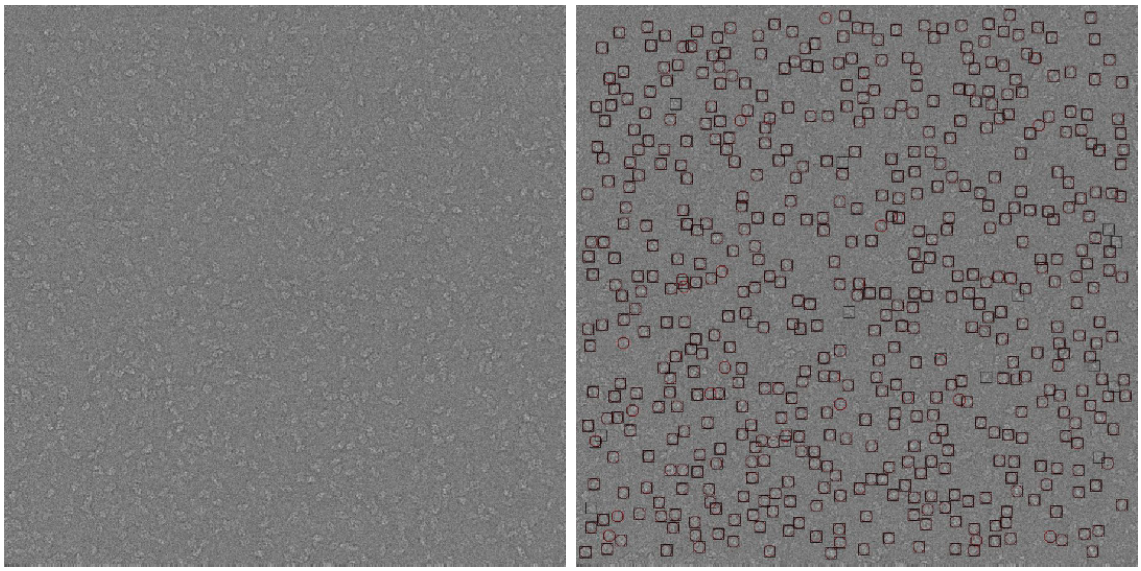
Figure 2.2: CTF examples

using supervised ML algorithms to automate this process. An example of a picking is shown in the Figure 2.3

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 2.4.

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 2.3: 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.

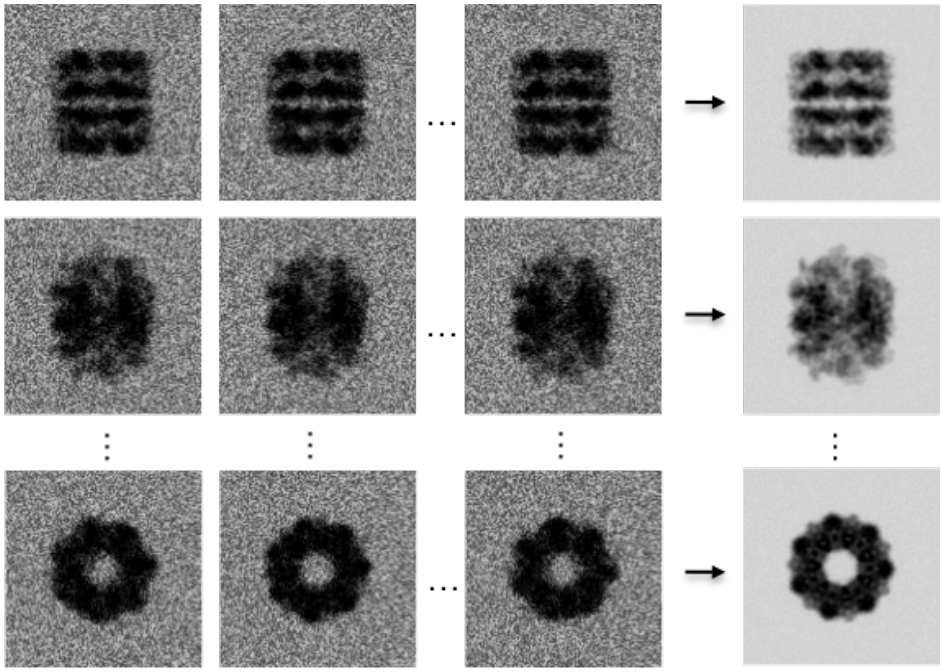


Image obtained from: [4]

Figure 2.4: Example of 2D classification

Ab-initio map reconstruction

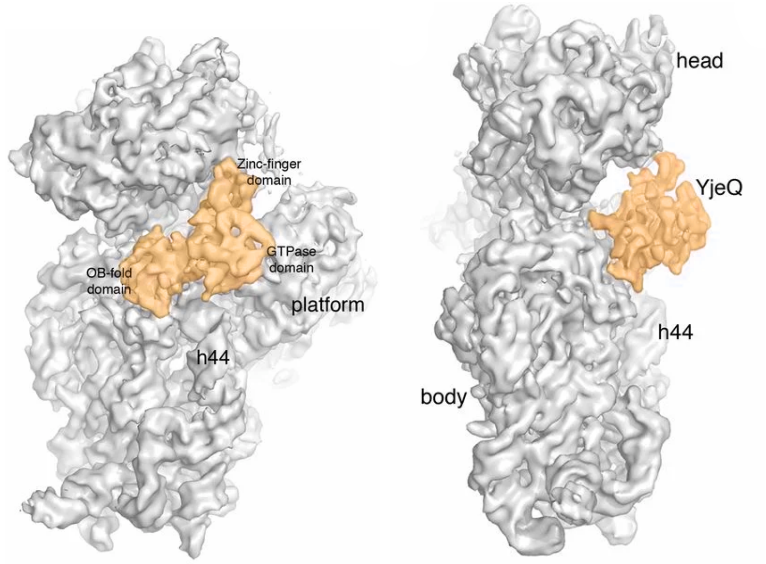
3D reconstruction is usually a Stochastic Gradient Descent (SGD) algorithm which iteratively improves a 3D electron density map of the protein under study. Therefore, choosing

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[10].

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[11]. Finally, there are some novel approaches that make use of unsupervised ML methods to learn a map from the particles[12].

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 2.5 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: [13]

Figure 2.5: 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 2.6.

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[10]. When the loop stops producing significant changes or a desired resolution is achieved, the cycle should be stopped.

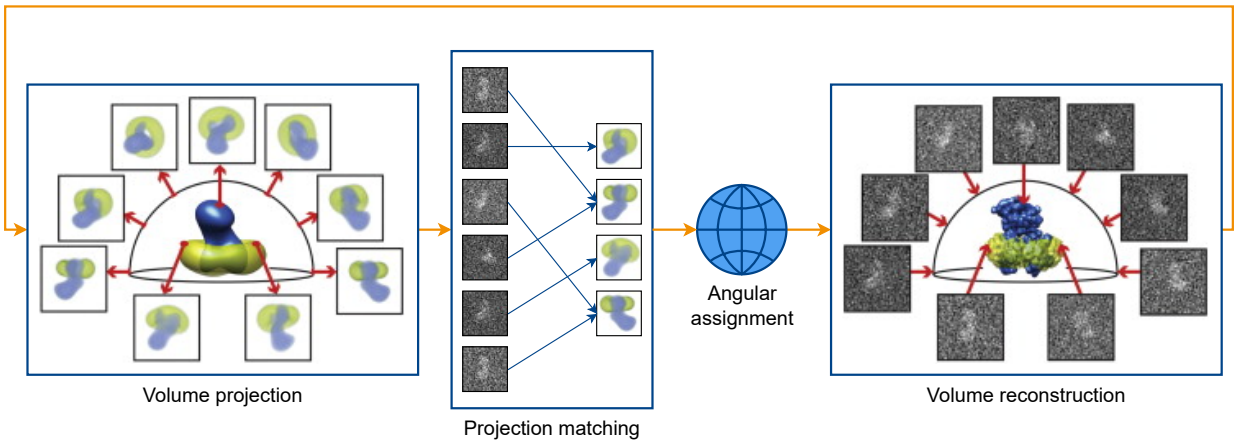


Diagram figures from: [14]

Figure 2.6: 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 2.7. Most reconstruction algorithms leverage

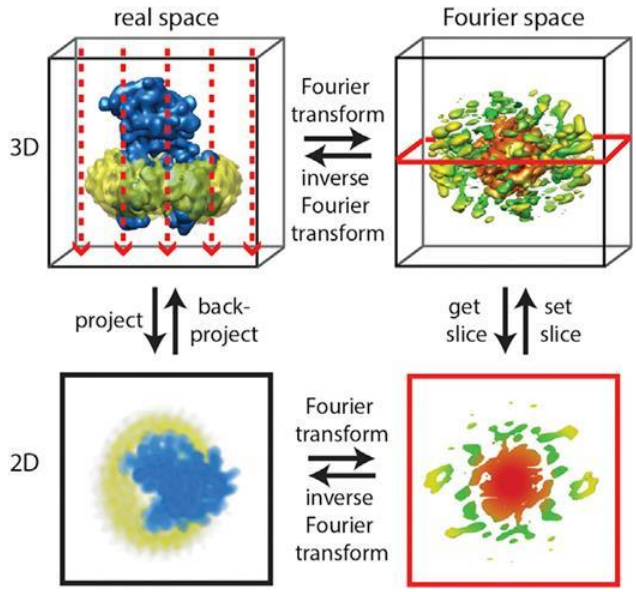


Image obtained from: [14]

Figure 2.7: 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 2.7 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 2.8

2.2 Summary of SPA

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: [15]

Figure 2.8: 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[16], Imagic, Eman[17], Cistem[18], Relion[19] and Xmipp[20]. In 2016 the introduction of CryoSPARC[21] was disruptive due to its significant performance improvements. Closely related to this, Scipion[22] 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.

Xmipp

Xmipp is an 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 in 1996, although it has taken 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) and random conical tilt[20].

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)[23], 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[24] or CryoET[25].

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. As of 2024, there are more than a 100 plugins available for Scipion. 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 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

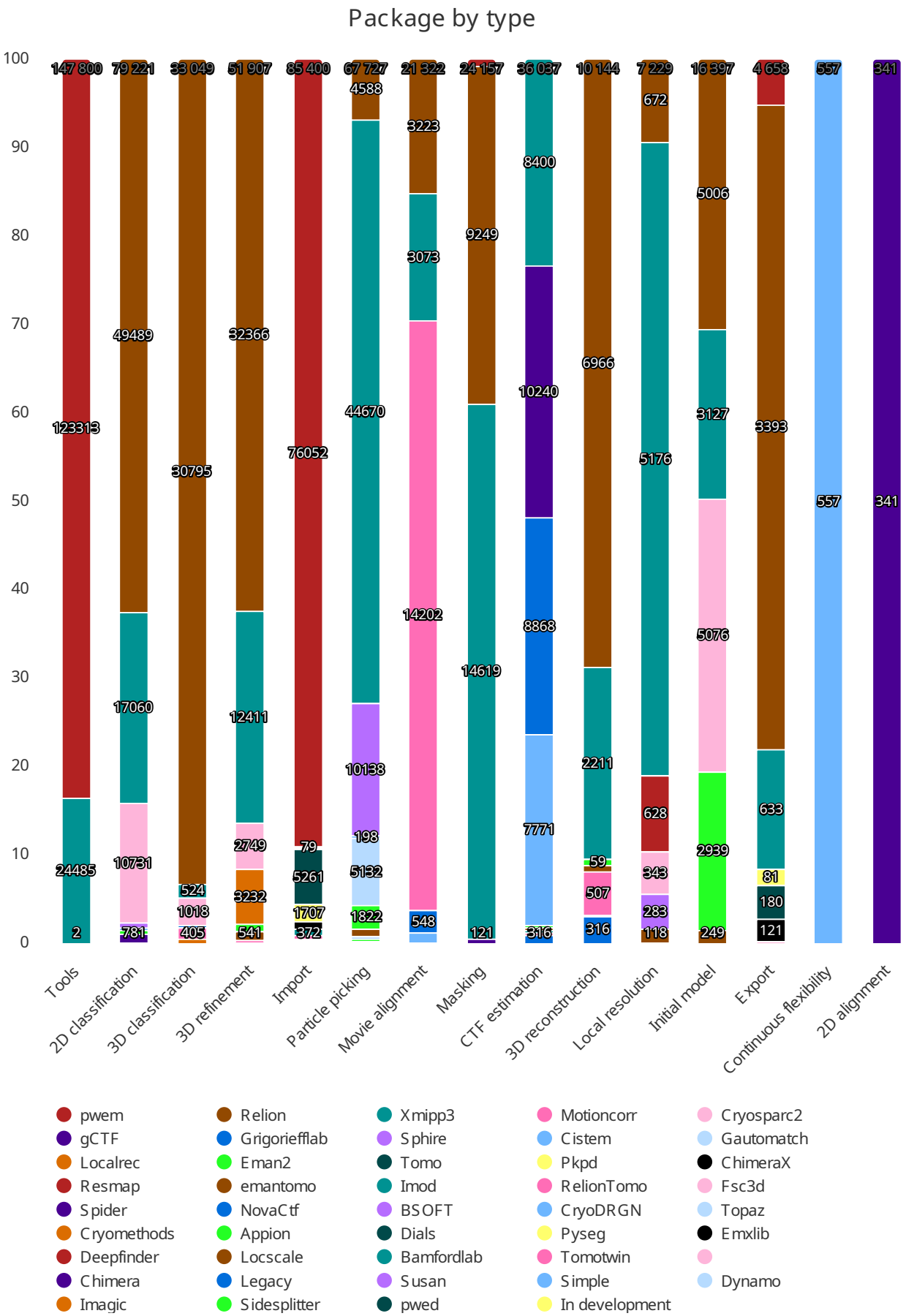


Figure 3.1: Scipion package usage statistics by type

feedback loop enables microscope operation with little human intervention, significantly reducing costs.

3.2 3D classification in SPA

One of the key challenges in CryoEM is the heterogeneity of biological samples, where multiple conformations or compositions may coexist within a single dataset. 3D classification emerges as a crucial strategy to address this challenge.

Unlike 2D classification, which clusters particles based on their similarity, 3D classification leverages the 3D nature of the data to classify particles according to the structure they originate from. However, this is not a trivial task, since these structural variations are unknown. Consequently, particles must be categorized according to a criteria that is hidden. Additionally, the data poses a very low SNR, which is in the order of 1/100, further complicating the task[14].

More often than not, 3D classification is executed after a 3D refinement. This means that particles have their orientations estimated. Similarly, a volume reconstruction comes implicit with these angular estimations. Due to the fact that a single volume was reconstructed from a presumably heterogeneous dataset, it will show features from multiple states. Therefore, this volume is named as "consensus volume". Usually, structural features that remain invariant across states can be reconstructed at high resolution, but non consistent regions will be significantly degraded[26].

This suggests that heterogeneity information is local. Thus, 3D classification is usually performed in a focused manner. To do so, 3D classification algorithms can be provided with a 3D mask that selects a Region of Interest (ROI) on which the classification focuses. If the ROI is unknown, the whole protein is can be selected.

Regarding the algorithmic implementation of the state-of-the-art solutions, these usually take an iterative EM approach to the problem. However, several alternatives have arised in the last decade. These approaches will be detailed hereafter.

Expectation Maximization algorithms

EM is an iterative method to find a local estimate for an unknown parameter of an statistical model[27]. When EM is used for 3D classification in CryoEM, on each iteration, each particle is compared to a set of volumes to find the likelihoods of having been projected from each one of them (expectation). Then, these volumes are reconstructed with the most likely particles (maximisation). These newly reconstructed volumes are used as reference for the

next iteration. Successive iterations of this process are expected to reinforce distinctive features on each of the volumes[21].

In spite of this, an initial solution is required for the first iteration of the algorithm. It's important to recognize that EM algorithms to converge to a local solution[27], implying that the convergence is influenced by the chosen initial solution[6][7]. Therefore, the selection of the starting point plays a crucial role in the outcome of the EM algorithm. At the same time, many existing algorithms depend on randomness to generate these initial solutions, ultimately leading to highly non-deterministic results.

For instance, as stated in Relion's 2016 paper on the topic of 3D classification, "Unsupervised classification is achieved by initializing multireference refinements from a single, low-resolution consensus model and assigning a random class to each particle in the first iteration"[28]. Similarly, Cryosparc offers "simple" initialisation which is equivalent to the previous one.

Another common pitfall of the EM algorithms is the so called "Attraction problem", which relates to a class gathering increasingly more elements on each iteration[6][29]. This can be attributed to many factors, but in the context of CryoEM it is usually provoked by a class that has slightly better SNR than the other classes. Then, this class will correlate better with many elements[30][7], albeit these are not being correctly classified. This in turn will lead to more particles being averaged on that class, further increasing it's SNR[31].

In general, EM based 3D classification requires a prior knowledge of the number of classes in the dataset[7]. This information is not always available to the user, specially when the protein is flexible and the concept of class does not exist (instead there is a continuum of states). Nevertheless, there are numerous approaches that circumvent this limitation such as performing hierarchical 3D classifications[29][32].

3D Variance Analysis

3D variance analysis takes advantage of many reconstructions from randomly selected subsets in the data. Then, these random reconstructions are compared using variability analysis techniques such as Principal Component Analysis (PCA) so that varying regions can be recognized[33]. Usually, 3D variance analysis is used as a method to obtain a initial solution for subsequent EM interactions.

Contrary to the EM approaches, 3D variance analysis does not require a prior knowledge about the number of classes, as these can be inferred from the densities in the latent space[7]. Nevertheless, many hyper-parameters such as the number of particles per reconstruction, the number of random reconstructions and latent space dimensionality need to be chosen in advance.

A widespread implementation of this approach is in Cryosparc, which offers a PCA initialization before performing online EM iterations[21]. This PCA approach is also available in the `split volumes` protocol from Xmipp[23].

Flexibility analysis

Until this point, we have coped with conformational heterogeneity by classifying the particles in numerous discrete classes. This method reconstructs various states from particle images under the assumption that there is a defined number of discrete conformational states explored by the specimen. While this approach has proven successful in many cases, these discrete states limit the information that can be extracted regarding the actual motion of the protein. As a consequence, recent leaps in hardware and software have allowed to extract this motion information from individual particles[34].

Indeed, this is a very novel field in the field of CryoEM and several unique approaches have arised in the last few years. Although the particular implementations vary greatly, they all converge by the fact that they are based on Deep Neural Networks (DNNs)[35].

Nevertheless flexibility analysis is not a substitute for 3D classification, as it relies on the fact that atomic mass remains invariant across states. For obvious reasons, this does not hold true for compositional heterogeneity, where a compound may or may not be present.

4. Implementation

In this project we aim to develop a novel method for determining a reliable initial 3D classification. Doing so, many of the inconveniences related to the current EM and 3D variations analysis approaches are avoided. The algorithm detailed on this chapter provides a high quality initial solution to the subsequent EM iterations. Doing so, the EM algorithm's convergence will be biased towards the correct solution, avoiding local minima and attraction problem. This increased quality of the initial solution also implies that fewer EM iterations are demanded, which in conjunction with the computational efficiency of the algorithm, allows for a significant speed up on the 3D classification process. Lastly, very few hyper-parameters are required to be tuned to achieve correct results.

4.1 Initial partition algorithm

The 3D classification problem is specially challenging due to the fact that images need to be classified according to unknown structures. Once these structures have been partially discovered, 3D classification becomes much easier, as particles can be considered to belong to their most similar 3D structure by the means of projection matching.

In this section we aim to describe a new approach that can be used to discover initial solutions without any prior knowledge, except for the 3D alignment of the particles, which implicitly come associated to a consensus volume. A consensus volume is a volume that has been reconstructed with all the particles, disregarding any possible heterogeneity. Thus, it is expected that it expresses a mixture of features from the underlying structures.

Our initial 3D classification approach leverages the fact that particles projected from similar directions should resemble to one another. Thus, grouping particles by their projection angle, should enable us to perform a image classification in 2D. Then, using graph theory, we can relate neighbouring groups, leading to a global classification. An overview of this process is displayed in Figure 4.1.

In this early stage, we will ignore the presence of the CTF and assume that images have

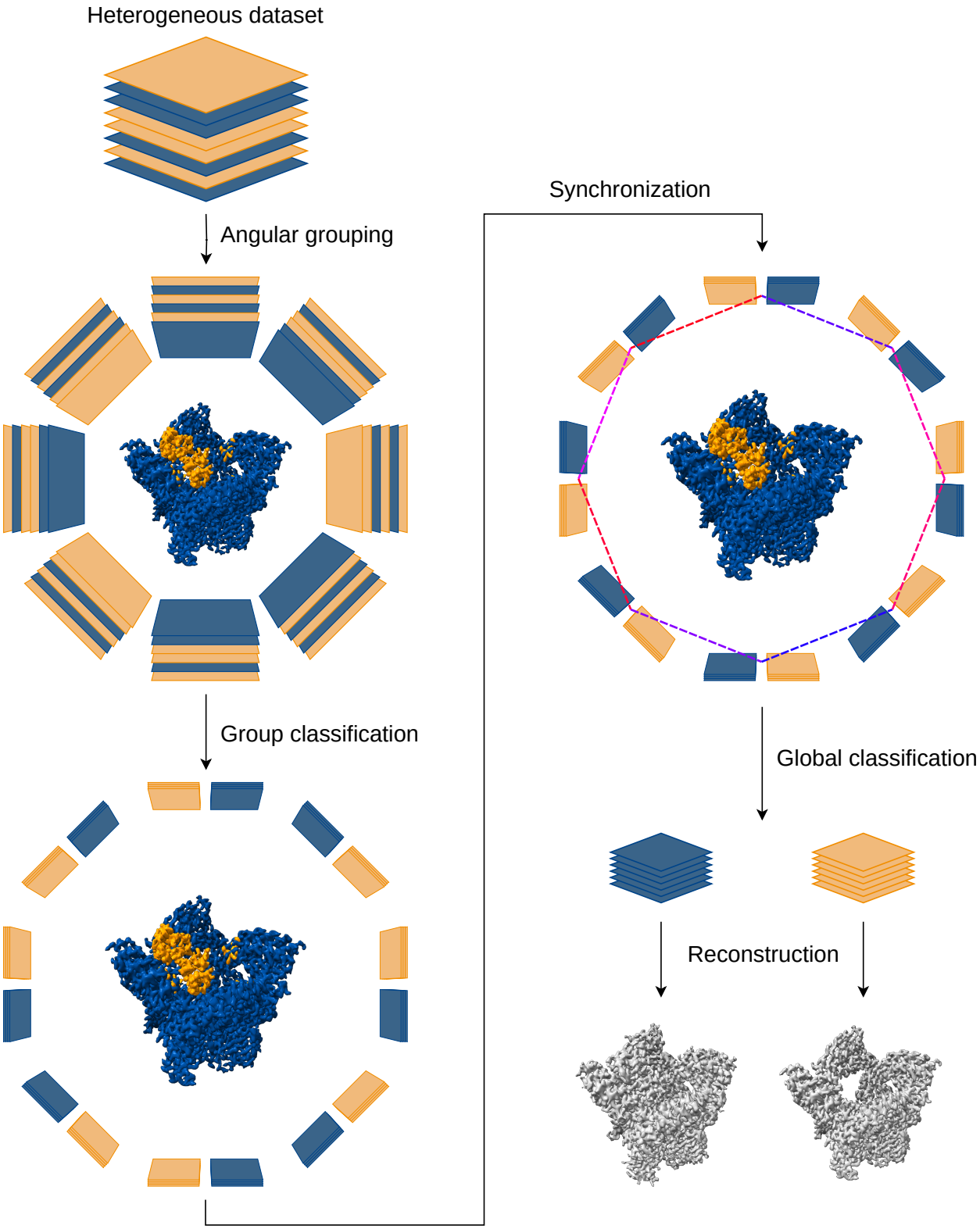
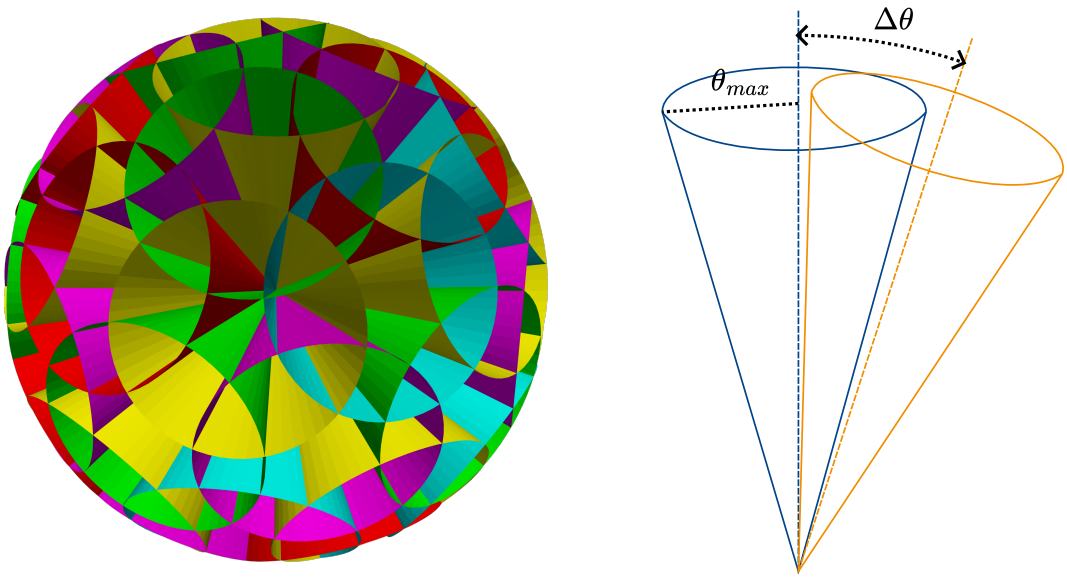


Figure 4.1: Overview of the initial 3D classification algorithm

their CTFs corrected.

Projection grouping

As a first step, our algorithm groups neighbouring projections, as they are considered to be similar enough and only differ in the structural feature we are interested in. These groups are overlapping, this is, a particle may belong to multiple groups. Indeed, for reasons that will be detailed later, some amount of overlap is mandatory. A example of such a grouping is illustrated in Figure 4.2.



(a) Projection sphere divided in overlapping cones (b) Cone spacing criteria

Figure 4.2: Particle grouping with cones

Each of these groups has a representative projection direction, which will be named as $\boldsymbol{r_i} \in S^2$, where S^2 is the unit sphere in $\mathbb{R}^3$. These direction vectors are artificially generated in such a way that they are quasi equally spaced in the unit sphere. These points can be limited to a hemisphere, as projections emanating from antipodes are equivalent, except for a mirror transformation. Additionally, if the protein under study poses symmetry, this can be accounted to further reduce the directions to the unit cell of the corresponding symmetry group.

Regarding the particles themselves, their projection direction is usually represented by a triplet of Euler angles: $(\theta, \phi, \psi)_j$. The physiscal meaning of these Euler angles is represented in Figure 4.3. However, for our purposes, it becomes more handy to represent projection directions of the experimental images with unit vectors $\boldsymbol{\gamma_j} \in S^2$. The Euler angles can be easily converted to vector notation using the expression (4.1). Note that in this conversion the ψ parameter of the Euler angles is discarded, as this does not participate in the projection direction definition. Instead, it is used to define the in-plane rotation as, represented in Figure 4.3.

$$\boldsymbol{\gamma_j} = \begin{bmatrix} \sin(\phi_j) \cdot \cos(\theta_j) \\ \sin(\phi_j) \cdot \sin(\theta_j) \\ \cos(\phi_j) \end{bmatrix} \tag{4.1}$$

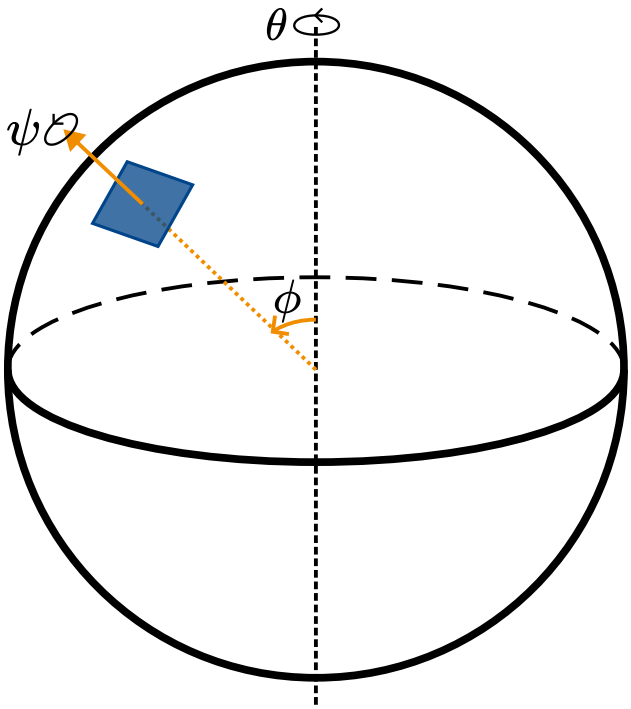


Figure 4.3: Representation of the Euler angle convention used in CryoEM

At this point, the grouping of projections becomes trivial: A particle is considered to be on a group only if the angle between its projection direction and the representative direction of the group is smaller than some arbitrary threshold θ_{max}:

$$G_i = \{j : \mathrm{acos}(\langle \boldsymbol{r_i}, \boldsymbol{\gamma_j} \rangle) \leq \theta_{max}\} \tag{4.2}$$

θ_{max} defines the aperture of the cones used for grouping. Thus, it is important to use a value low enough such that the variability induced by the projection direction is negligible compared to the actual variability in the structure. In our experience, a value of 7.5° provides good results. Similarly, the number of groups needs to be sufficient so that these cones overlap. To do so, we are selecting the number of groups so that they are spaced on average by $\Delta\theta \approx \theta_{max}$ as shown in Figure 4.2.

Classification of neighbouring projections

In the previous step we have classified images according to their projection direction. At this point, our intention is to classify the images from each group into two classes that are maximally dissimilar. This process will be applied for each directional group.

Image alignment

Before attempting any classification, images need to be aligned to one another. To do so, we will leverage the 3D alignment estimation provided by the preceding steps in the SPA image processing pipeline. The three possible in-plane transformations are represented in Figure 4.4.

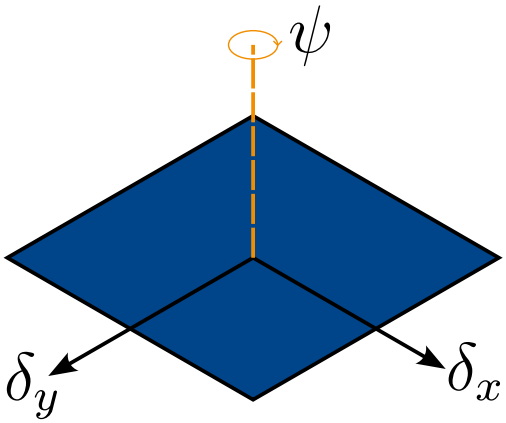


Figure 4.4: In-plane transformation of the particles

First, the center offset of the particles accounted by displacing them by their shift estimate (δ_x, δ_y) in the opposite direction. This ensures that the particle is centered in the image frame.

Secondly, the in-plane rotation of the particles is corrected. To do so, the Euler angle estimates are converted to quaternion space. Quaternions are a four-dimensional extension of complex numbers, commonly utilized to represent rotations in three-dimensional space. Quaternions can also be represented by a vector-scalar pair $\boldsymbol{Q} = (\boldsymbol{v}, w) = (\boldsymbol{u} \cdot \sin\alpha, \cos\alpha)$, this is, a rotation of α angular units around an unitary axis $\boldsymbol{u}$. Note that $\boldsymbol{v} = \boldsymbol{u} \cdot \sin\alpha$ relates to the complex part of the quaternion, whilst $\boldsymbol{w} = \cos\alpha$ involves its scalar component.

The Euler angles can be converted to quaternions using the Equation (4.3). Note that unlike the projection direction $\boldsymbol{\gamma_j}$, the in plane rotation ψ is considered when computing the quaternion representation of the Euler angles.

$$
\begin{aligned}
v_1 &= \cos\left(\frac{\phi}{2}\right) \cdot \cos\left(\frac{\theta}{2}\right) \cdot \cos\left(\frac{\psi}{2}\right) - \sin\left(\frac{\phi}{2}\right) \cdot \sin\left(\frac{\theta}{2}\right) \cdot \sin\left(\frac{\psi}{2}\right) \\
v_2 &= \cos\left(\frac{\phi}{2}\right) \cdot \cos\left(\frac{\theta}{2}\right) \cdot \sin\left(\frac{\psi}{2}\right) + \sin\left(\frac{\phi}{2}\right) \cdot \sin\left(\frac{\theta}{2}\right) \cdot \cos\left(\frac{\psi}{2}\right) \\
v_3 &= \cos\left(\frac{\phi}{2}\right) \cdot \sin\left(\frac{\theta}{2}\right) \cdot \cos\left(\frac{\psi}{2}\right) - \sin\left(\frac{\phi}{2}\right) \cdot \cos\left(\frac{\theta}{2}\right) \cdot \sin\left(\frac{\psi}{2}\right) \\
w &= \sin\left(\frac{\phi}{2}\right) \cdot \cos\left(\frac{\theta}{2}\right) \cdot \cos\left(\frac{\psi}{2}\right) + \cos\left(\frac{\phi}{2}\right) \cdot \sin\left(\frac{\theta}{2}\right) \cdot \sin\left(\frac{\psi}{2}\right)
\end{aligned}
\tag{4.3}
$$

We will use this vector-scalar representation to project the quaternion onto the $\boldsymbol{r_i}$ axis using the expression (4.4). This expression was derived from the twist-swing[36] decomposition, which describes a quaternion with two components: a rotation around a certain axis and the residual part. In this case, we are only interested in the rotation around the projection direction.

$$\hat{\psi} = \text{atan2}(\langle \boldsymbol{v_j}, \boldsymbol{r_i} \rangle, w) \tag{4.4}$$

Note that the ψ component of the Euler angles also expresses the in-plane rotation. However, due to the fact that we are combining images with (slightly) different projection angles, $\hat{\psi}$ leads to a more precise in-plane alignment.

In practice, both of the alignment operations (rotation and shift) are performed at once with an affine matrix transformation. This matrix is shown in Equation (4.5). Note that the shift is applied before the rotation, so the rotation is also applied to the shift vector.

$$M = \begin{bmatrix} \cos\hat{\psi} & -\sin\hat{\psi} & 0 \\ \sin\hat{\psi} & \cos\hat{\psi} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & \delta_x \\ 0 & 1 & \delta_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \cos\hat{\psi} & -\sin\hat{\psi} & \cos\hat{\psi}\cdot\delta_x - \sin\hat{\psi}\cdot\delta_y \\ \sin\hat{\psi} & \cos\hat{\psi} & \sin\hat{\psi}\cdot\delta_x + \cos\hat{\psi}\cdot\delta_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.5}$$

Image classification

Once all the particles of a directional group are aligned to one another, a classification is attempted. To do so, we have used PCA, a popular dimensionality reduction technique[37]. In short terms, PCA describes a set of multidimensional points in an orthogonal basis where its components are not correlated. Moreover, components of the basis are ordered by the variance they capture[38].

On our approach, we consider a lexicographically ordered version of the image (pixels layed out as a 1D vector), so that PCA can be applied to them. This ordering is performed withing a mask, so that the classification can focused on a ROI. This mask is generated by projecting the input 3D mask in the reference direction.

This PCA procedure is applied in each group to an aligned stack of particles as shown in Figure 4.5. Assuming that the primary source of variability across images is the heterogeneity, the first eigen-image will be representative of this heterogeneity.

In our implementation, we use `pytorch`[39] to perform the PCA analysis and obtain the first eigen-image. Then, the eigen-image can be used as a projection basis for the images. This allows to order particles according to their projection value. Indeed, this can be interpreted as an interpolation weight between two classes that are maximally dissimilar, C_+ and C_-.

$$\rho_i = \langle \boldsymbol{u_1}, \boldsymbol{x_i} \rangle \tag{4.6}$$

where $\boldsymbol{u_1}$ is the first principal component (largest eigenvalue) and $\boldsymbol{x_i}$ is the vector representation of a particle.

We will assume that PCA projection values of follow Gaussian distributions $C_+ \sim \mathcal{N}(\mu_+, \sigma_+^2)$ and $C_- \sim \mathcal{N}(\mu_-, \sigma_-^2)$. Thus, using the projection values we can attempt a classification by fitting a Gaussian Mixture Model (GMM) of two components to their histogram. A GMM is a probabilistic model that represents the probability distribution of a dataset as a mixture of multiple Gaussian distributions.

To avoid numerical stability issues, we enforce the same variance for both GMM components ($\sigma_+^2 = \sigma_-^2$). Consequently, each component of the GMM can be used to estimate the likelihood of their corresponding class. An example of such a fitting is displayed in Figure 4.6.

Computing the log likelihood ratio of the components, we can rank particles according to the class they most likely belong to. In other words, log likelihood ratio's sign determines which of the classes is more probable, while its magnitude reflects the certainty. This ratio is also plotted in Figure 4.6.

$$\log \Lambda(\rho) = \log \frac{\mathcal{L}(C_+ \mid \rho)}{\mathcal{L}(C_- \mid \rho)} \tag{4.7}$$

We will name the log likelihood ratio of a given particle as λ_i:

$$\lambda_i = \log \Lambda(\rho_i) \tag{4.8}$$

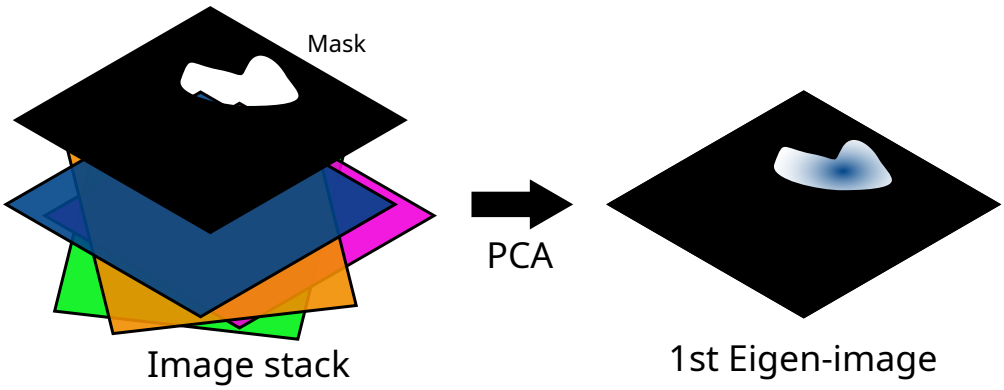


Figure 4.5: Eigen-image computation process

Directional class synchronization

In the previous step we have established two classes for each of the directional groups. It is expected that these classifications have been performed according to the heterogeneity we are interested in. However, this does not mean that classes match across groups. In fact, PCA has two equally valid solutions for each axis, which are opposite to one another. Similarly, the GMM fitting is arbitrary. Thus, class ordering is not deterministic. In other words, C_+ in one group may correspond to C_- in another group and vice versa.

Consequently, a synchronization of classes across directional groups is required. That is why cones need to overlap. Doing so, we can make use of common particles to evaluate if a given pair of adjacent groups have equal or opposite classifications.

The first step to perform this synchronisation is to find the common particles by computing the set intersection of adjacent groups:

$$I_{ij} = G_i \cap G_j \tag{4.9}$$

The dot product of the log likelihood ratios of common particles in each group can be used as a similarity metric to compare their classifications. Positive values indicate that the classifications are aligned. Accordingly, negative values point that the classes are swapped.

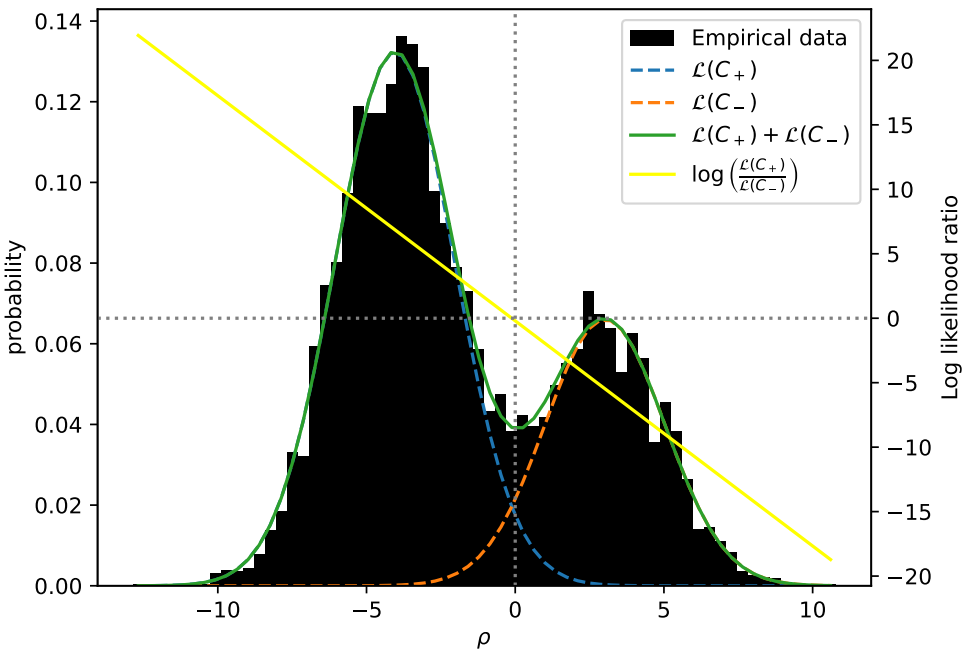


Figure 4.6: Example of a Gaussian Mixture Model of two components

$$w_{ij} = \sum_{p \in I_{ij}} \lambda_{ip} \cdot \lambda_{jp} \tag{4.10}$$

At this point we want to maximize the compatibility of all directions by flipping the classifications of a given set of groups. This swapping of classes translates into applying a negative sign to all their log likelihood ratios. According to the expression (4.10), this would also result in the application of this negative sign to all the weights related to these groups.

Thus, we can express our synchronization problem as the quadratic binary optimisation problem posed in equation (4.11). This is, we want to maximise the total compatibility by flipping some classifications.

$$\begin{aligned} &\max_{\boldsymbol{\sigma}} f(\boldsymbol{\sigma}) = \sum_{i,j} w_{ij} \cdot \sigma_i \cdot \sigma_j = \boldsymbol{\sigma}^T W \boldsymbol{\sigma} \\ &\text{subject to} \\ &\sigma_i \in \{-1, +1\} \Leftrightarrow \sigma_i^2 = 1 \end{aligned} \tag{4.11}$$

where the sign of σ_i relates to a group being swapped or not.

Analogy with magnetic dipoles

Although this optimisation problem may look naive, it is not trivial to solve, as it cannot be approached with Lagrange multipliers. Similar cases to this problem also appear in nature, as is the case of the Hamiltonian of the Ising Model[8].

The Ising Model describes the behaviour of a network of magnetic dipoles where these may freely flip their polarity. The Hamiltonian of the model represents the energy of the system[8]. As in many cases in nature, the energy of a system tends to be minimal, and the Ising Model is not an exception to this rule. Thus, the Ising Model will be on a stable state only if its Hamiltonian is minimal. The expression for Hamiltonian of the Ising Model without external interactions is provided hereafter:

$$\begin{aligned} &H(\boldsymbol{\sigma}) = -\sum_{i,j} J_{ij} \cdot \sigma_i \cdot \sigma_j \\ &\sigma_i \in \{-1, +1\} \end{aligned} \tag{4.12}$$

where J_{ij} represents the magnetic interaction between each pair of dipoles. When $J_{ij} > 0$ the interaction is ferromagnetic (magnet polarities to align) and when $J_{ij} < 0$ the interaction is anti-ferromagnetic (magnet polarities tend to oppose). σ_i represents the spin of each of the magnetic dipoles in the system. This will help us establish a classification-magnet metaphor.

Except for the minus sign, the expression (4.12) is equal to our objective function detailed in (4.11). Indeed, due to this duality, the minimisation of the energy in the Ising Model corresponds to the maximisation of our objective function. Therefore, solutions to the Ising Model also serve as solutions for our problem.

One of the most common ways to approach the Ising Model problem is by finding the maximum cut of the graph defined by the weighted adjacency matrix $\boldsymbol{A} = -\boldsymbol{J}$, which in our case corresponds to the similarity matrix $\boldsymbol{W}$. Once the graph has been bi-partitioned using the maximum cut criteria, magnets (or classes) of one partition are flipped to minimize energy (or maximize compatibility)[8].

An example of a graph of direction similarities is detailed in Figure 4.7, where the edge colours represent the similarity metric between adjacent directions. The vertices are positioned in $\boldsymbol{\gamma_i}$, so that the graph appears to be embedded in the projection sphere.

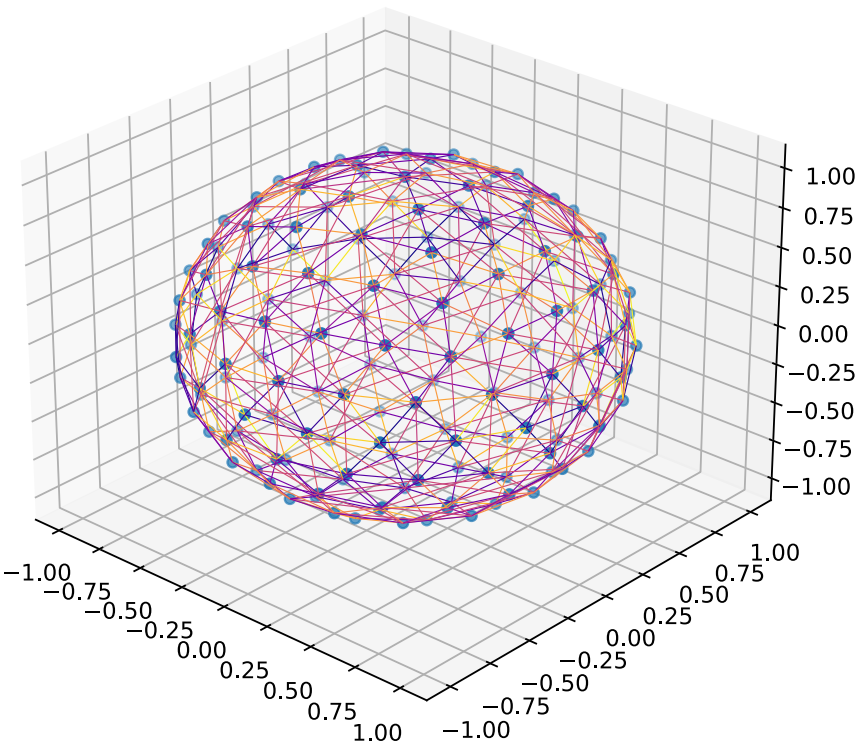


Figure 4.7: Graph embedded on the projection sphere

Graph maximum cut algorithm using SDP

Until this point we have only shifted our optimisation problem into other domains, without giving any concrete solution to it. Nevertheless, we have reasoned that our problem directly maps to the graph maximum cut problem.

A graph cut is a partition of the graph into two disjoint sets of vertices (no edges connecting the two subsets). Thus, the the graph cut size is the sum of the edges that need to be removed to obtain a disjoint bi-partition. Considering the former definition, the graph maximum cut is defined as "A cut whose size is at least the size of any other cut"[40]. However, similar to the previous analogies, this problem is not easy to solve, as it is deemed NP-Hard[41]. Luckily, numerous heuristic approaches exist that provide reasonably bounded solutions.

In our case, we have decided to use an approach based on Semi Definite Programming (SDP). SDP is a mathematical optimization technique that extends linear programming to solve optimization problems involving symmetric matrices, offering powerful tools for addressing a wide range of real-world optimization challenges. Using a SDP approximation to solve the graph maximum cut problem is guaranteed to find a solution that is at least 87% accurate while also being computationally efficient[42].

The SDP approach involves relaxing $\sigma_i \in \{-1, +1\}$ to $\boldsymbol{x_i} \in S^{N-1}$ where S^{N-1} is the unit sphere in $\mathbb{R}^N$. This allows us to define the matrix $\boldsymbol{X}$ of pairwise dot products:

$$X_{ij} = \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle \Leftrightarrow \boldsymbol{X} = \boldsymbol{Y} \cdot \boldsymbol{Y}^T \tag{4.13}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{x_1} & \boldsymbol{x_2} & \dots & \boldsymbol{x_N} \end{bmatrix} \tag{4.14}$$

Due to the commutative property of the inner product, we can deduce that $\boldsymbol{X}$ must be symmetric. Moreover, all its diagonal values must be 1, as these involve the dot product of a unit vector with itself. Last but not least, given the fact that $\boldsymbol{X}$ can be expressed as the outer product of a matrix $\boldsymbol{Y}$ with itself, $\boldsymbol{X}$ is also guaranteed to be positive semi-definite:

$$\boldsymbol{x}^T(\boldsymbol{Y}\boldsymbol{Y}^T)\boldsymbol{x} = (\boldsymbol{Y}^T\boldsymbol{x})^T(\boldsymbol{Y}^T\boldsymbol{x}) = \|\boldsymbol{Y}^T\boldsymbol{x}\|^2 \geq 0 \quad \forall \boldsymbol{x} \tag{4.15}$$

Considering all these constraints, we can pose the SDP problem shown in Equation (4.16). The greatest benefit of this formulation is that there are SDP solvers that run in polynomial time (as opposed to the binary optimisation problem which was NP-Hard).

$$
\begin{aligned}
&\min_{\boldsymbol{X}} f(\boldsymbol{X}) = \sum_{i,j} a_{ij} X_{ij} \\
&\text{subject to} \\
&\boldsymbol{X} = \boldsymbol{X}^T \\
&X_{ii} = 1 \\
&\boldsymbol{X} \succeq 0
\end{aligned}
\tag{4.16}
$$

In our implementation, we use `cvxpy`[43][44] package to solve this SDP problem. To get back the values of $\boldsymbol{x_i}$ from $\boldsymbol{X}$ we use matrix square root function ($\boldsymbol{Y} = \boldsymbol{X}^{\frac{1}{2}}$) available in `scipy`[45]. Finally, we undo the relaxation to extract σ_i from $\boldsymbol{x_i}$[42]:

$$
\sigma_i = \text{sign}(\langle \boldsymbol{v}, \boldsymbol{x_i} \rangle) \tag{4.17}
$$

where $\boldsymbol{v}$ is a random vector in S^{N-1}.

Global classification

As a final step, we combine the information gathered from all the 2D classifications to obtain a 3D classification of the particles. To do so, once all the 2D classifications have been synchronized to one another, all the log likelihood ratios of a given particle are averaged (remember that a particle may belong to several directional groups). Recalling that the sign of the log likelihood ratio represents the class, this sign is used to assign the 3D class of the particle.

After bi-partitioning the particle set according to their 3D class, a volume is homogeneously reconstructed for each of the 3D classes. This is done though the `xmipp_reconstruct_fourier` program, which was already implemented in the Xmipp suite[46].

4.2 Software architecture

In the previous section we described the algorithmic implementation of the 3D classification method. In this section we aim to explain how the code was structured to approach this task. Most of the code was developed as a Scipion protocol, although some specific tasks were delegated to separate programs.

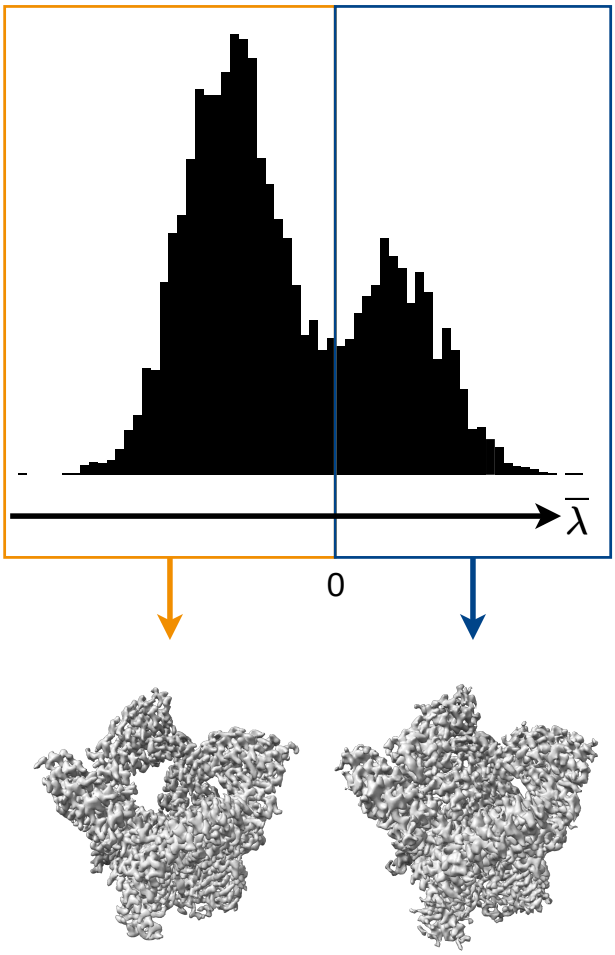


Figure 4.8: Global classification

Scipion protocol

Scipion is a CryoEM image processing platform that integrates many widespread image processing packages through plugins. Each plugin consists of a collection of programs known as protocols. Usually, protocols can be seen as high-level steps of an image processing pipeline. Indeed, a 3D classification can be considered as a protocol. In this project, we have implemented our program as a Scipion protocol named as `split volume` inside the Xmipp plugin.

Typically, a Scipion protocol defines a set of input parameters which are displayed in the Graphical User Interface (GUI) when launching the protocol. The most relevant parameters of our protocol are listed hereafter:

- **Input particles**: The particles analyzed during execution
- **Input mask (optional)**: A binary mask defining the ROI on which the classification focuses. If not provided, an spherical mask is automatically generated

- **Symmetry group**: If the protein under study exhibits a particular symmetry, this can be specified here to reduce projection directions.
- **Resize (optional)**: Controls if the particles are down sampled to a particular size. If not provided, particles are not resized.
- **Angular sampling**: Average spacing between angular groups. Defaults to 7.5 degrees.

Once the protocol is launched, it executes a series of operations known as steps. In the case of our protocol, these steps are linear, meaning that they are executed sequentially. The heavy operations are usually delegated to standalone programs, whilst the straightforward operations are implemented inside the protocol steps. The most relevant steps from our protocol are detailed here:

1. **Convert input**: The input is converted to an appropriate format for our processing. If downsampling is selected, this downsampling occurs here.
2. **Angular neighborhood**: Groups particles according to their projection directions. To do so, the existing `xmipp_angular_neighbourhood` program is used.
3. **Classify directions step**: Each group is classified into two classes as dissimilar as possible. This is achieved with a newly created ad-hoc program named as `xmipp_aligned_2d_classification`, which will be detailed in the next section.
4. **Build graph**: Compares neighbouring classifications to build a graph with their similarities.
5. **Graph optimization**: The maximum cut of the graph is computed so that classifications can be synchronized. `xmipp_graph_max_cut` program was created to perform this task.
6. **Partition**: Once the classifications are synchronized, their log likelihood ratios are averaged and individual particles are classified according to their sign.
7. **Reconstruction**: Each of the classes is used to produce a volume using the already existing `xmipp_reconstruct_fourier` program.
8. **Create output**: Classes are converted to Scipion format.

At the end, the Scipion protocol produces a set of classes and a set of volumes representing the classes. These objects may be used as input for another protocol that requires them, like for instance, a subsequent Relion 3D classification.

Viewer

Scipion protocol viewers are small GUIs that show detailed information about the protocol's execution. Therefore, a protocol viewer was also implemented for this algorithm for diagnose purposes. This viewer is able to present two visualizations related to its execution.

Firstly, directions can be interactively analyzed to ensure that heterogeneity is being captured. To do so, a histogram of the PCA projection values is presented alongside the GMM fitting. By dragging the yellow line, the user can visualize the heterogeneous state represented by the projection value. A snapshot of this representation is shown in Figure 4.9.

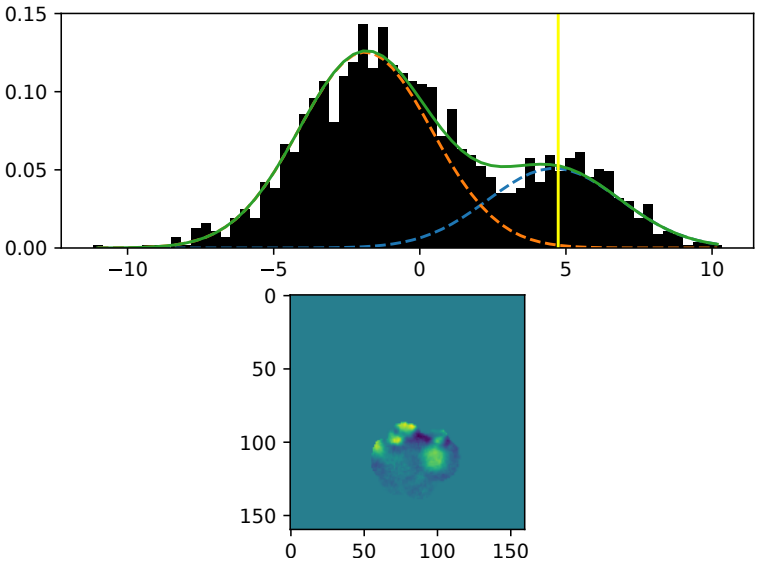


Figure 4.9: Snapshot of the classification viewer

The other visualization involves a graph relating adjacent directions. This graph is embedded on a 3D sphere, which represents the projection directions. The edges of the graph are coloured so that they represent the weight attributed to it. This representation is displayed in Figure 4.10. Note that this graph is also interactive, the user may drag the cursor to change the viewing angle.

Auxiliary programs

As mentioned earlier, computationally intensive operations were segregated from the protocol logic into their own programs. Then, these programs are invoked by the protocol. This separation benefits cluster users, as programs invoked by Scipion protocols can be dispatched to queue engines such as Slurm. In addition, it helps distinguishing the general control logic from the algorithmic nuances.

Aligned 2D classification

This program receives a set of images with their corresponding 3D alignment parameters and a reference direction. Then, it aligns the images to the reference projection direction and computes their PCA. These PCA projection values can be used to evaluate the class of each particle. We use `pytorch` to transform the images and compute the PCA.

Note that in our implementation we fit a GMM model to these PCA projection values. This fitting is performed by the protocol itself and not the classification program.

Graph max cut

Another computationally demanding task is the graph maximum cut. As mentioned in the previous section, we translate the maximum cut problem into a semi-definite programming program, which is solved by `cvxpy`. This program precisely does this translation. It takes a potentially sparse matrix representing the adjacency matrix of a graph and it converts it into a semi-definite programming problem. Once solved, it outputs a two sets of indices expressing the vertices corresponding to each of the partitions of the graph.

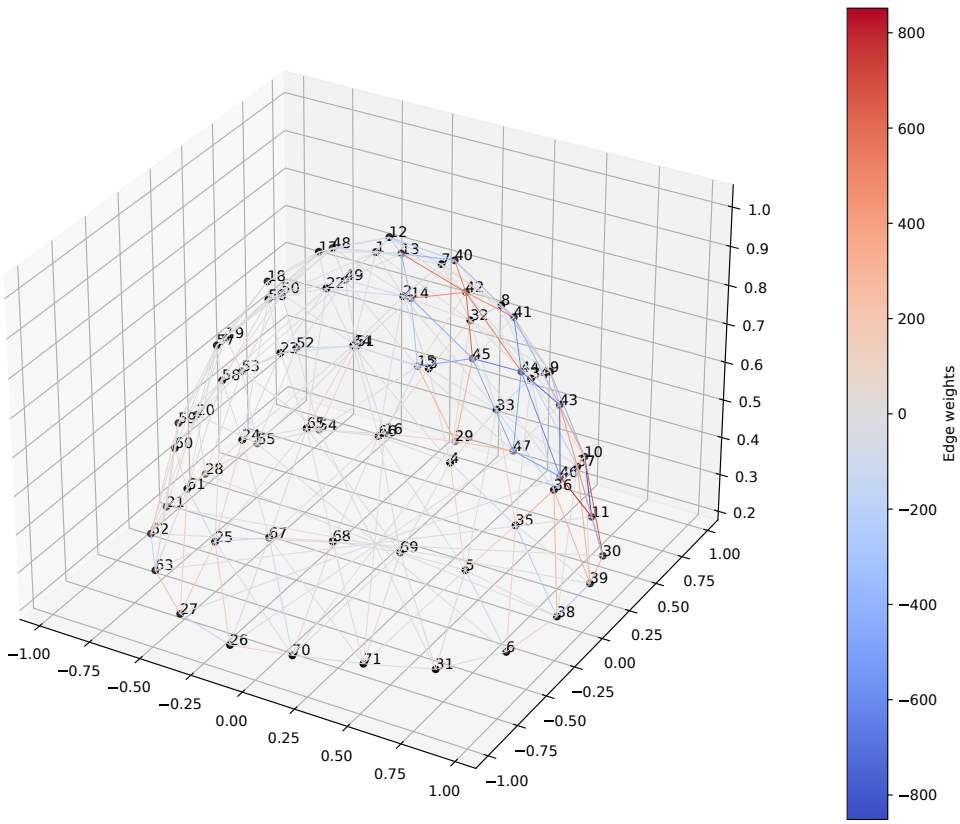


Figure 4.10: Snapshot of the 3D graph viewer

5. Results

This chapter aims to evaluate the algorithm described in this project. The assessment will involve the utilization of multiple datasets to demonstrate its robustness across varying scenarios. Subsequently, a comparative analysis will be conducted, comparing the algorithm against well established state-of-the-art solutions.

5.1 Test datasets

In the algorithm's assessment, three carefully chosen CryoEM datasets have been used. The datasets reflect various conditions that can be found in reality, so that the comparisons shown here are transcendental. One the datasets exhibits compositional heterogeneity, while the other two are highly flexible. In such a way, we intend to assess the performance in both heterogeneity scenarios.

We have preferred to use publicly available datasets from Electron Microscopy Public Image Archive (EMPIAR), so that the results detailed here can be replicated. Nevertheless we have also used a in-house acquisition that is not public yet (although it is expected to become public soon).

TRPV-5

The Transient Receptor Potential Vanilloid 5 (TRPV-5) protein is an ion channel, which plays a crucial role in the regulation of calcium homeostasis within various tissues and cells. This sort of channels are widely distributed in mammalian organisms and are involved in sensory perception, cell signaling, and the maintenance of cellular ionic balance. TRPV-5 is primarily expressed in the renal tubules, where it participates in the reabsorption of calcium ions. The protein's significance in renal physiology underscores its role in maintaining systemic calcium levels, ultimately impacting bone health, neuromuscular function, and overall mineral homeostasis.

Research on TRPV-5 has gained significance in recent years, focusing on its structural fea-

tures, functional properties, and the signaling pathways it engages in. Understanding the molecular behaviour of TRPV-5 provides a basis for developing targeted therapies for disorders related to calcium deregulation.

We have decided to mix two EMPIAR entries where one of them comes from an experiment where calmodulin was added in-vitro (EMPIAR-10253) and potentially binds to two N and C lobes. The other experiment comes from a mutant in a clean buffer, although it had potential to bind endogenous calmodulin (EMPIAR-10256). Unlike the first experiment, the mutation on the TRPV-5 from the second experiment disables the calmodulin binding in the C lobe. Therefore, the difference between the two datasets lies around the binding site in the C lobe[32][47]. These lobes are highlighted in Figure 5.1.

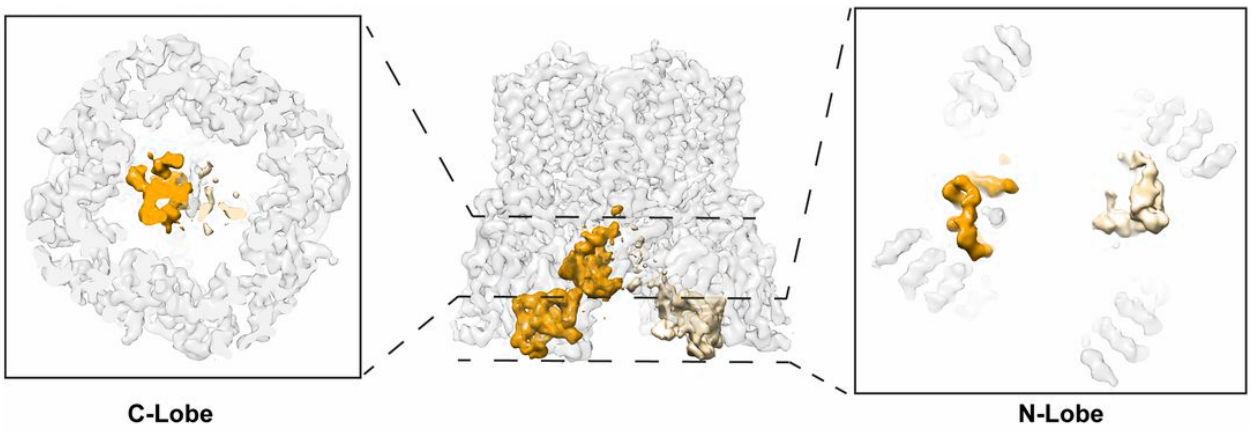


Image obtained from: [47]

Figure 5.1: TRPV-5 reconstruction

A similar classification experiment was proposed in the "Data-driven determination of number of discrete conformations in single-particle cryo-EM" paper. As stated by its authors, "These datasets are challenging because the extra density corresponding to calmodulin is very small and breaks the symmetry of the complex making accurate particle alignments critical to achieve a successful separation"[32].

The two datasets used in these tests are the EMPIAR-10253[48] and the EMPIAR-10256[49][47]. In conjunction, they sum $166,611$ particles, from which 60% originate from the mutant experiment, and the other 40% originate from the in-vitro experiment. These particles have a size of $256 \times 256\text{px}^2$ and they were acquired with a sampling rate of 1.06Å/px. As two datasets were mixed, it is not fair to consider their alignments, as these were estimated in isolation. Thus, all the particles were re-refined to re-estimate the alignment parameters in heterogenous conditions using CryoSPARC's non-uniform refinement[21]. A sample of the particles of these datasets is showcased in Figure 5.2.

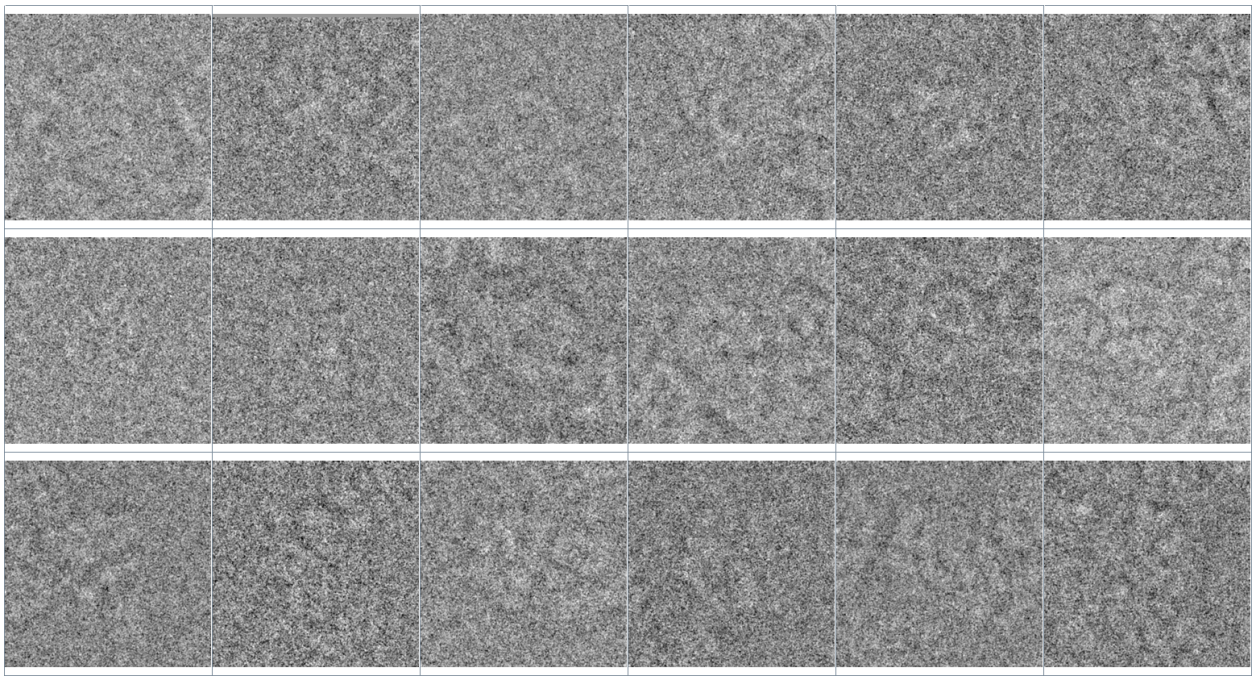


Figure 5.2: Sample of the TRPV-5 particles

Pre-cathalytic spliceosome

The pre-cathalytic spliceosome is a critical component in the process of RNA splicing. RNA splicing is a fundamental cellular mechanism that involves the removal of introns (non-coding regions) from precursor messenger RNA (pre-mRNA) and the joining of exons (coding regions) to generate mature mRNA. The spliceosome is a large and dynamic molecular machine responsible for orchestrating this process.

Understanding the functions of the pre-catalytic spliceosome is crucial for unraveling the molecular mechanisms that govern RNA splicing, which plays an important role in gene expression and cellular function. Researchers investigate these processes to gain insights into various genetic and cellular disorders, as abnormalities in splicing can lead to diseases.

This molecular machine is highly flexible, meaning that it exhibits continuous heterogeneity. Indeed, it contains two independent regions with flexibility, which will be analyzed separately. In these analysis we aim to observe multiple stable states on those regions. In particular, we will focus our analysis on the SF3b and helicase regions detailed in Figure 5.3, which are the most flexible ones.

We have conducted our tests on this macro-molecule using the publicly available a public dataset from the EMPIAR repository, precisely the EMPIAR-10180[51] dataset. This dataset is commonly used as a baseline to assess and evaluate flexibility analysis algorithms[26][50][34]. Due to this continuous heterogeneity, the our aim is to observe the most common states. The EMPIAR-10180 dataset is provided as a set of $327,490$ aligned particles of size $320 \times 320\text{px}^2$ at a sampling rate of 1.699Å. Additionally, the authors of the

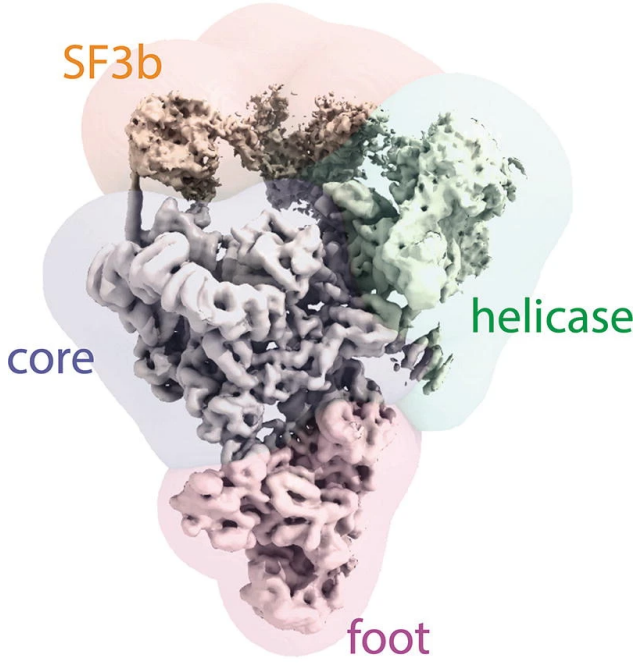


Image obtained from: [50]

Figure 5.3: Pre-cathalytic spliceosome reconstruction

dataset also provide masks enclosing the ROIs of each of the flexible regions. A sample of the particles of this dataset is displayed in Figure 5.4.

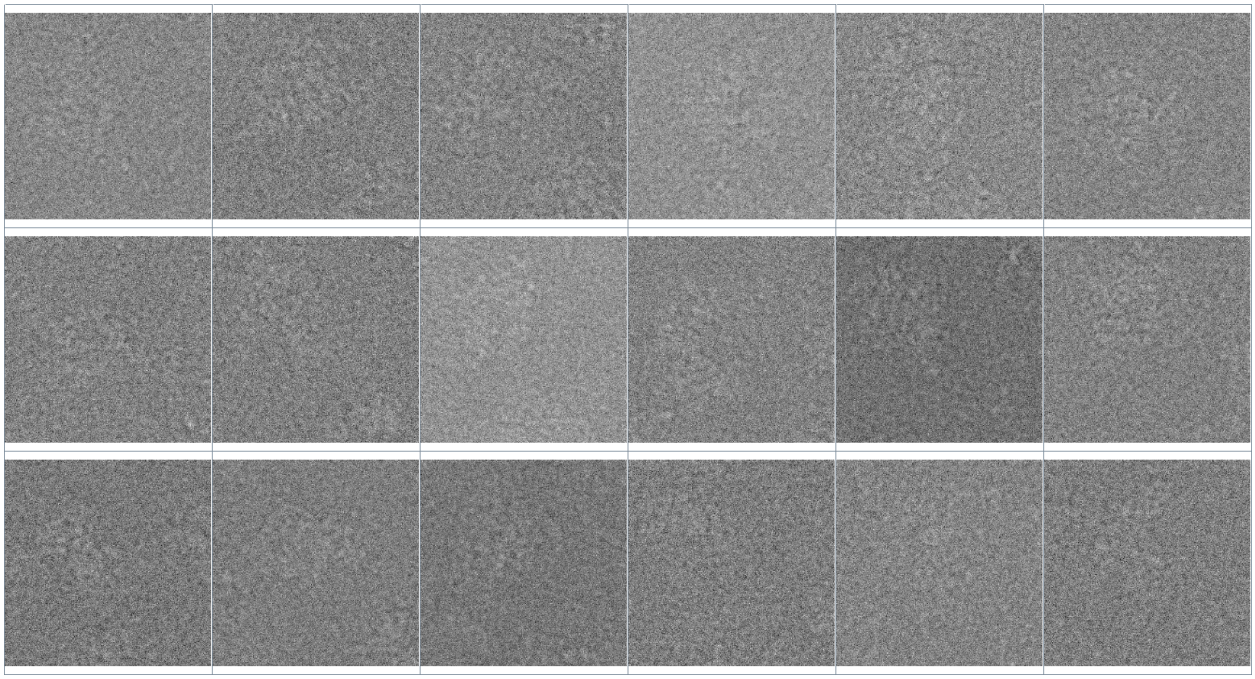


Figure 5.4: Sample of the spliceosome particles

HER-2

The HER-2 protein, also known as human epidermal growth factor receptor 2, is a crucial molecule in the context of cell growth, division, and differentiation. HER-2 is particularly remarkable for its involvement in cancer biology, as its overexpression or amplification has been identified in a variety of malignancies, most notably breast cancer. When HER-2 is overexpressed, it can lead to uncontrolled cell proliferation, increased survival, and enhanced invasive properties, contributing to the aggressive nature of certain cancer types.

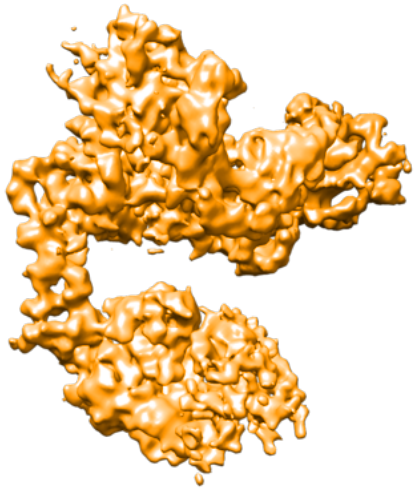


Figure 5.5: HER-2 reconstruction

In the results detailed on this chapter, we have experimented with a in-house dataset which was previously processed by Dr. Marcos Gragera Cabezudo. This dataset is comprised of $352,500$ aligned particles of size $200 \times 200\text{px}^2$ acquired with a sampling rate of 1.3Å/px. A sample of these particles is shown in Figure 5.6. Similarly to the pre-cathalytic spliceosome, HER-2 also exhibits flexibility. In particular, it is comprised of two sub-units that are flexibly connected. Thus, by performing a 3D classification on it, we aim to observe multiple states.

5.2 Experiments

Using the previously described datasets we have conducted an exhaustive evaluation of our 3D classification method. This evaluation involved a comprehensive comparison with widespread state-of-the-art solutions, specifically Relion[19] and CryoSPARC[21]. To ensure a fair comparison, we have conducted tests under conditions as similar as possible. Due to our limitation of only being able to categorize into two classes, we have enforced this condition on all tests.

To be more specific, Relion has been tested with its default execution parameters, which implies that 25 EM iterations will be performed by it. Similarly, CryoSPARC will be tested

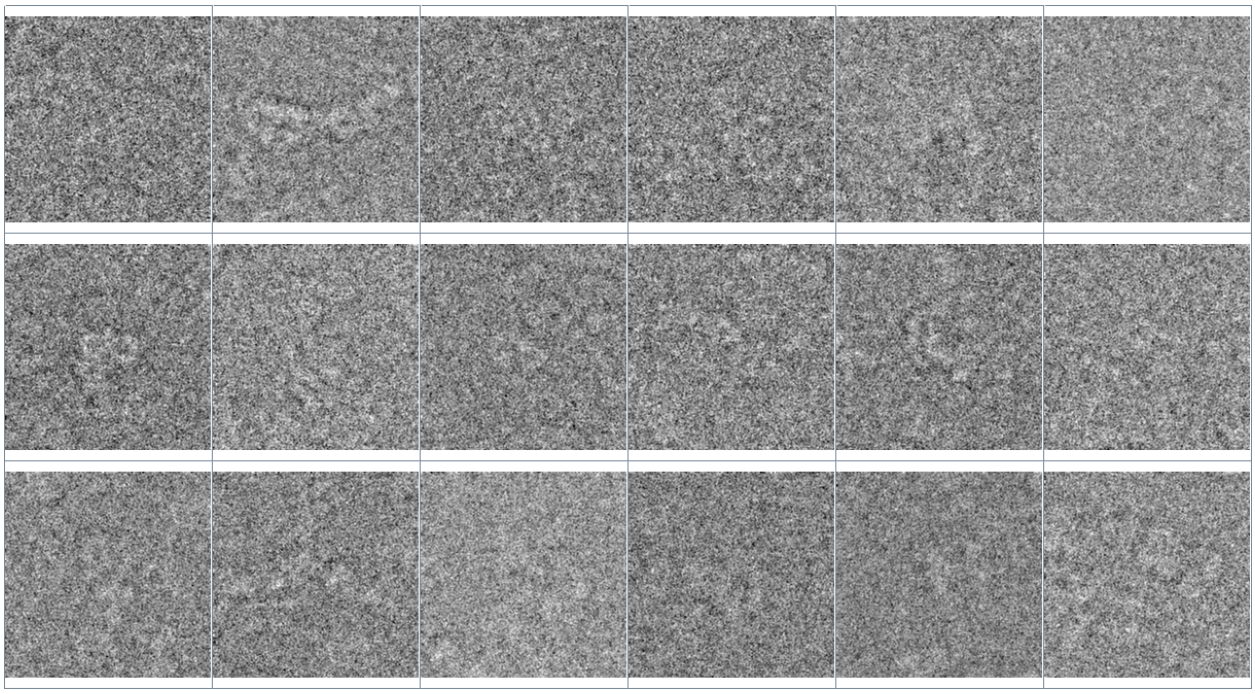


Figure 5.6: Sample of the HER-2 particles

both with “simple” (random) initialization and “PCA” initialization, leaving the rest of the parameters with their default values. In addition, our graph based approach will be executed with its default parameters. Lastly, a combination of our classification method with 2 Relion EM iterations will be assessed. The results are presented on a dataset basis, so that information needed for comparisons can be easily gathered. Nevertheless, performance results are presented jointly.

Due to a lack of ground truth, we will qualitatively evaluate the results by comparing the reconstructed classes. To do so, we will present representative slices from the reconstructions side by side, highlighting the difference. Similarly, resolution measurements are not suitable for assessing classification quality, as this depends greatly on the number of particles used for reconstruction.

Last but not least, computation time is measured, so that the performance alignment of each of the implementations can be added to the balance. The tests were conducted on a workstation with stable ambient conditions, so that performance measurements can be compared across executions. This workstation features dual Intel Xeon X5647 Central Processing Units (CPUs) and a NVIDIA Titan X GPU.

TRPV-5

As mentioned in the introduction of this dataset, it consists of a mixture of two separate experiments, one of which contains a mutation that disables calmodulin from binding at a particular spot. In this test we will try to identify this position using our own approach.

Unluckily, neither Relion nor CryoSPARC converged with this dataset, regardless of the configuration used in the execution. We suspect that they suffered attraction, as almost all particles ended in a single class. This attraction theory is supported by the evidence shown in Figure 5.7, which plots the class distribution across EM iterations of the Relion 3D classification. Therefore, we are not able to provide a comparison for this experiment. Nevertheless, we have measured their execution times for later comparison.

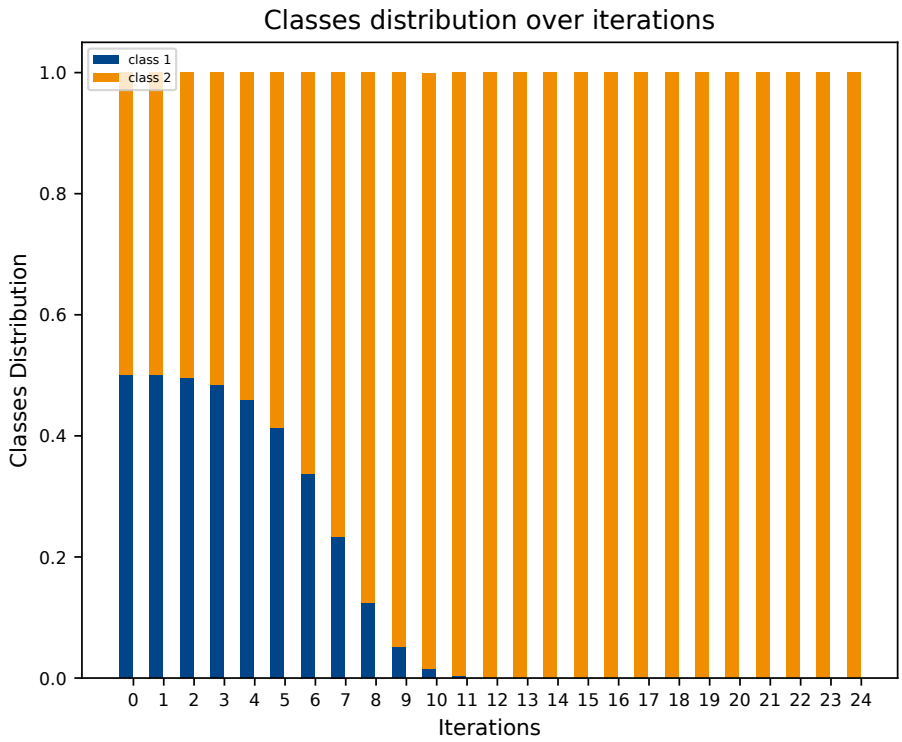


Figure 5.7: Class distribution across Relion's EM iterations with the TRPV-5 dataset

In spite of this, our algorithm successfully provided the correct solution. This is showcased in Figure 5.8 which represents a XY slice centered around the C lobe. In the binding site, there is a noticeable density difference between classes 1 and 2. What is more, when 2 Relion 3D classification iterations are applied after our 3D classification, this difference in density is further amplified. After these two iterations we have observed that the separation stops improving.

The results obtained with this dataset are particularly interesting because Relion by itself was not able to converge. However, when provided with our initial solution, it managed to converge in just 2 iterations.

Pre-cathalytic spliceosome

The next analyzed dataset will be the Pre-cathalytic spliceosome. As introduced later, this protein is comprised of multiple flexible areas, which will be explored separately. To do so,

we will use the provided masks to focus the classification on the region we are interested in.

When focusing our classification on the helicase part of the spliceosome, we observed inferior results when using our graph based algorithm standalone. However, when combined with Relion, it managed to obtain qualitatively similar results to CryoSPARC or standalone Relion. In addition, as discussed later, this combined approach run faster than the rest.

Similarly, when focusing the classification on the SF3b part of the protein, the separation with our approach is not as clear as in the rest of the cases. However, once again, two additional iterations of Relion are sufficient to achieve qualitatively similar results while remaining faster than the other approaches.

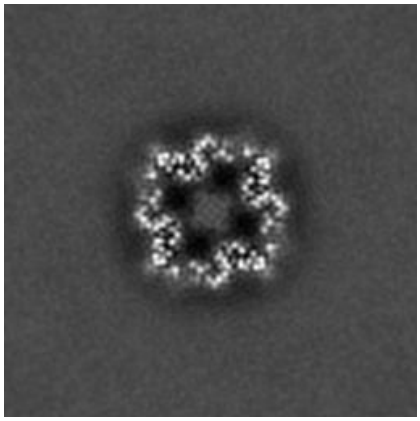
HER-2

The last assessed dataset is HER-2. As shown in Figure 5.11, all of the algorithms were able to find the same conformational variations, which relate to the upper sub-unit flexing side by side. Qualitatively it is not easy to judge which one provides better results. In fact, approximately 78.72% of the images were categorized stably across all classifications (including ours), indicating that classifiers mostly agree on their classifications. Possibly, the other 21.27% belongs to intermediate states, so that they can not be easily attributed to one of the extreme classes.

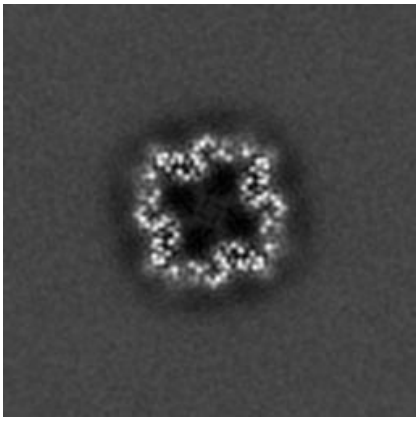
Performance

Regarding the performance of our algorithm, Figure 5.12 shows that it is consistently faster than other solutions. Even when factoring the additional Relion iterations, it can keep pace with CryoSPARC, even surpassing it with certain datasets. The difference is specially notable with the standalone Relion executions, as these are not GPU accelerated.

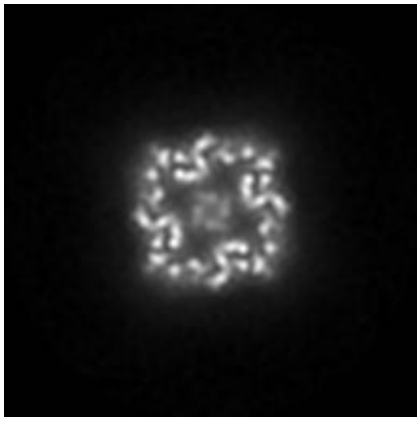
Regardless of the algorithm used, it is remarkable that the execution time varies greatly across datasets. Indeed, we have used two different time scales when representing the execution times in Figure 5.12, as processing the spliceosome was much slower. This is probably related to the fact that particles of this dataset are the largest ones. In addition, its amount of particles is also substantial, only being surpassed by the HER-2 dataset.



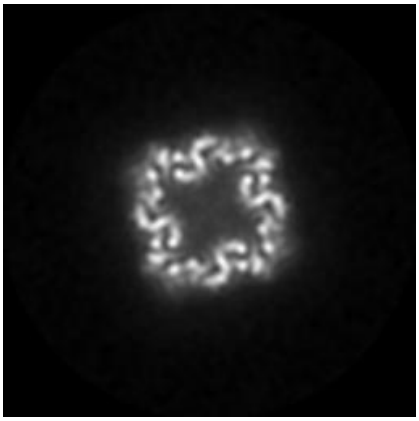
(a) Class 1 reconstruction of our classification algorithm



(b) Class 2 reconstruction of our classification algorithm

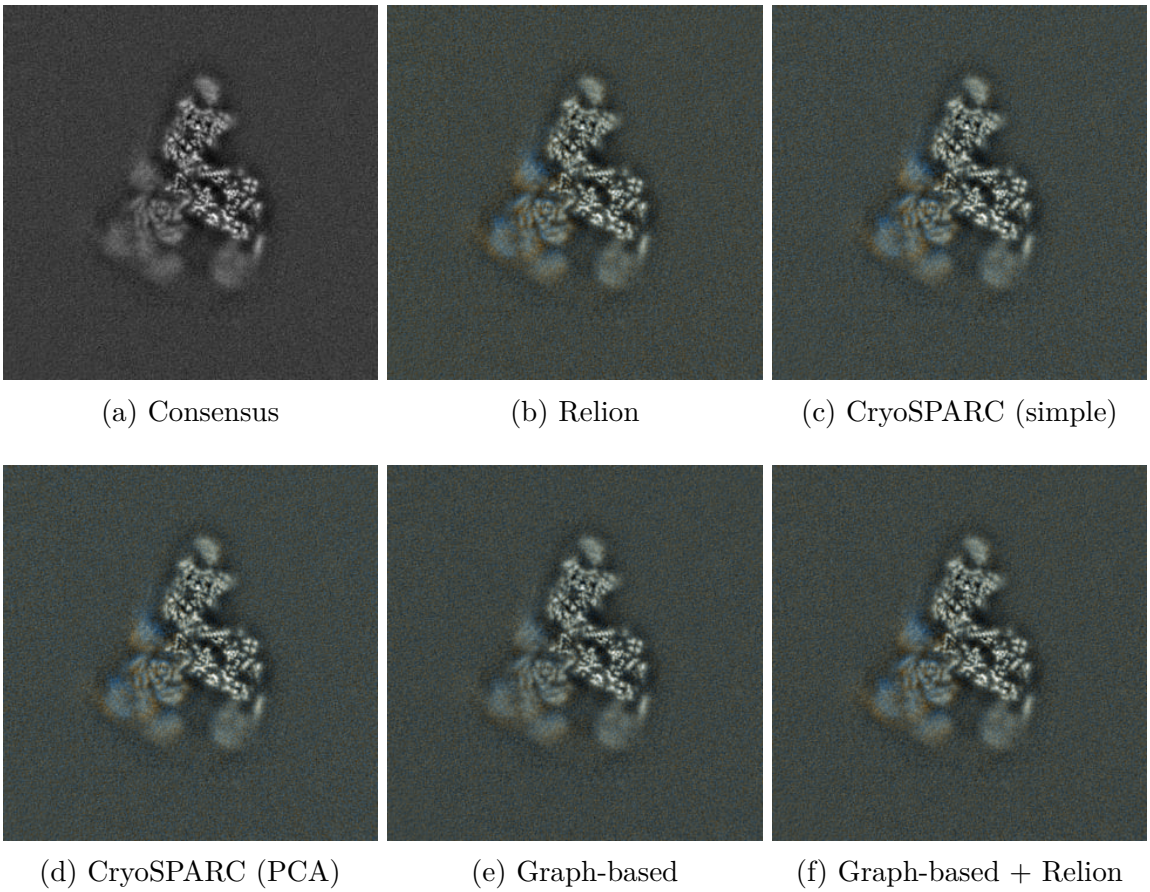


(c) Class 1 of reconstruction of our classification algorithm and 2 subsequent Relion iterations



(d) Class 2 reconstruction of our classification algorithm and 2 subsequent Relion iterations

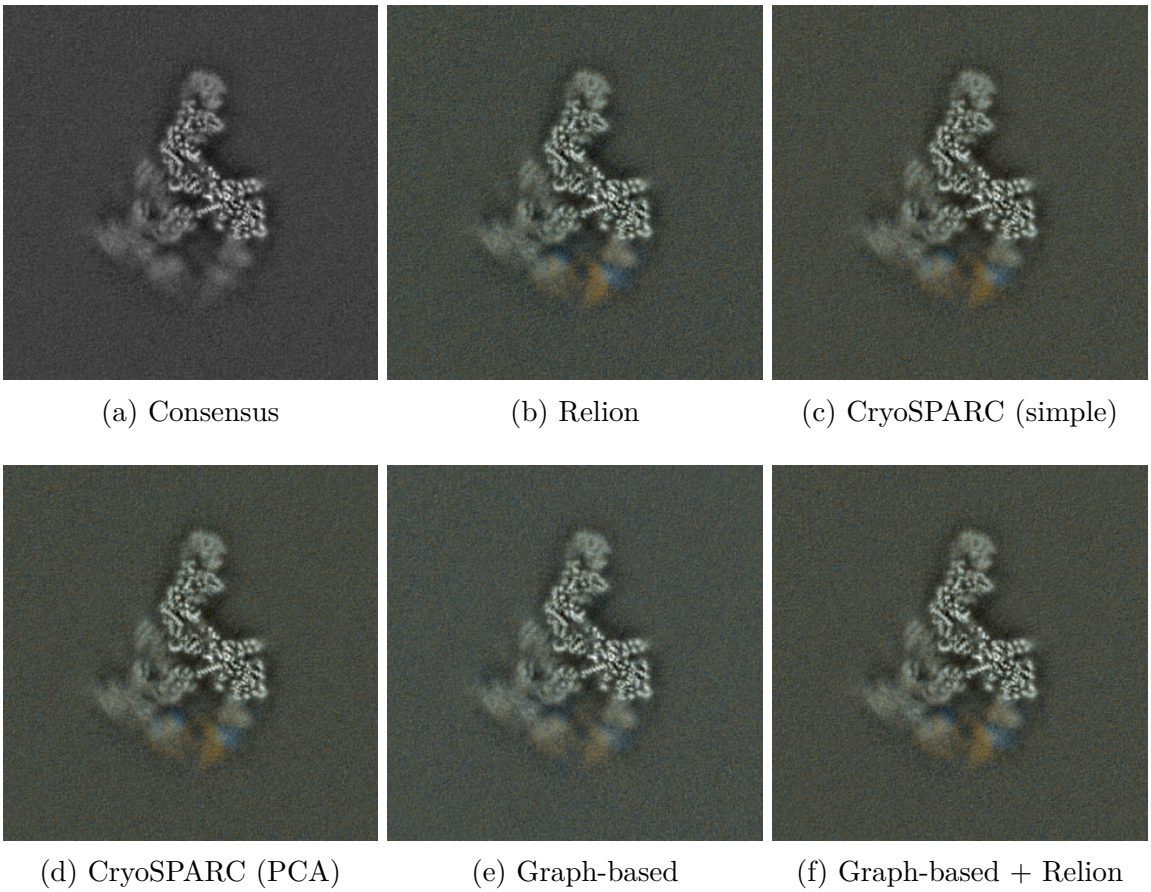
Figure 5.8: Slice 127 of the reconstructed volumes of TRPV-5 after classification



(a) Consensus (b) Relion (c) CryoSPARC (simple)

(d) CryoSPARC (PCA) (e) Graph-based (f) Graph-based + Relion

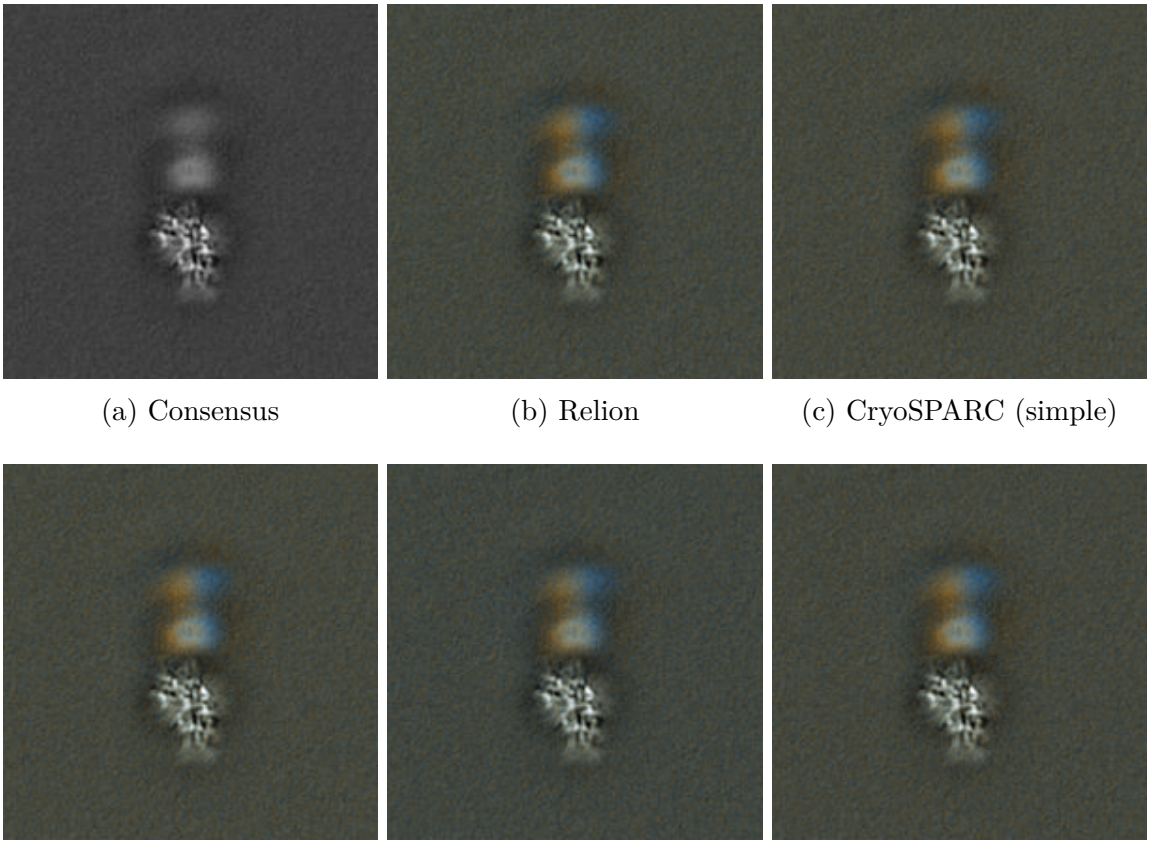
Figure 5.9: Classification experiments with the spliceosome dataset focusing on the helicase subunit



(a) Consensus (b) Relion (c) CryoSPARC (simple)

(d) CryoSPARC (PCA) (e) Graph-based (f) Graph-based + Relion

Figure 5.10: Classification experiments with the spliceosome dataset focusing on the SF3b subunit



(a) Consensus (b) Relion (c) CryoSPARC (simple)

(d) CryoSPARC (PCA) (e) Graph-based (f) Graph-based + Relion

Figure 5.11: Classification experiments with the HER-2 dataset

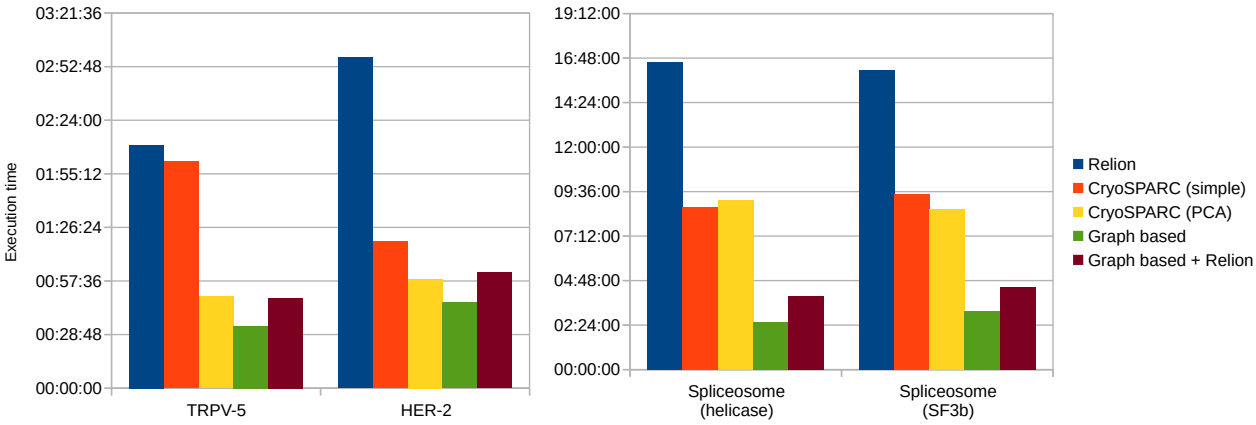


Figure 5.12: Execution time comparison

6. Conclusions

The tests described in the previous chapter show the importance of the initial solution in the 3D classification problem. This is not surprising, as there are countless examples in the scientific literature that describe this issue. The method presented here proves to be a very relevant approach for providing a reliable and high quality initial solution.

In fact the TRPV-5 experiment was of particular interest, as neither Relion nor Cryosparc were able to converge to the two distinct underlying classes. However, our approach managed to provide a correct solution. In addition, when its outcome was fed into Relion for further refinement, this was able to converge in a few iterations. This reinforces the previous statement, as Relion on its own did not converge, but when provided with our initial solution it completed successfully. Indeed, the experiments have shown that similar or better results were obtained when combining our algorithm with a couple of Relion EM iterations. In addition, this pathway not only provides superior results but also demonstrates to be one of the most performing ways to achieve 3D classification, only being beaten by our standalone initial solution with slightly worse results.

Even though discrete 3D classification is deemed obsolete for elucidating continuous movements of macromolecules, the continuous flexibility analysis tools are computationally very expensive, requiring hours or days to complete. Thus, the high throughput of the algorithm described in this work proves to be a viable option for preliminarily testing of conformational heterogeneity in a dataset. What is more, continuous flexibility analysis is not suited for compositional heterogeneity experiments, which are better modeled by the classical 3D classification methods. As a consequence, 3D classification remains as a necessary and crucial step in CryoEM image processing.

All in all, our method has proven to be a very effective approach to 3D classification, a necessary step in many CryoEM image processing scenarios. We have empirically demonstrated the improvements obtained by using it, both in terms of the quality of the results and the computational time required to obtain them. Therefore, we hope that structural biologists will take advantage of it for elucidating the behavioral insights of critical proteins involved in diseases.

7. Future work

7.1 Generalization to multiple classes

One of the largest limitations of our algorithm is that it is fixed to providing two classes. At the same time, many problems in biology require more that these two classes. As the classification algorithm is integrated in Scipion, the user may run it repeatedly in a hierarchical manner to obtain a greater amount of classes. However, this is tedious and inefficient. Thus, our next step is to automatize this hierarchical classification process.

Several of such approaches of this hierarchical classification already exist. For instance, as described by J. Gomez-Blanco et al., their approach subdivides each class until no resolution improvement can be obtained from this partitioning. To do so, they employ the ResLog plot criteria, which relates the reconstruction resolution with the amount of particles used at it[52]. A separation necessarily involves that less particles will be used for each reconstruction, so the overall resolution is expected to decrease. In their approach, the ResLog plot is used to test if the resolution improves in relative terms to the particle count. At the end, similar classes can be merged to gain back resolution[29].

7.2 Performance improvements

Even though our algorithm has demonstrated superior performance, its implementation can be improved in many ways to maximize computational resource usage. During testing, we have observed that the majority of time is devoted to the 2D group classifications. At the same time, this process was barely using the GPU of the system.

We suspect that this is due to a disk Input/Output (I/O) bottleneck, as many images need to be loaded onto the GPU, which can incur a higher computation time than the actual image processing. Currently, the classification program is invoked once per group. We believe that if the classification program is modified to consider a set of groups, loading only once duplicate particles can help to reduce the overall loading times.

7.3 3D classification refinement

In the results shown in this work, we have observed cases where subsequent EM iterations help to improve the initial partition in certain cases. In those results we have employed Relion to carry out those EM iterations. However, we intend to develop our own tools to refine the initial partition. A prototype of such a program has been already implemented in Xmipp, but extensive testing is yet needed.

A potential variant of this refinement program could also consider slight variations in the angular assignment from the consensus volume, as these were estimated from a partially incorrect map.

7.4 Classification consensus

Another moral obtained from the experiments presented in this project is that 3D classification algorithms are highly unstable and there is no single solution that is robust in all cases. This is not unique to the 3D classification problem, indeed, it is a common topic in all steps of the CryoEM image processing pipeline. One of the supporting pillars of Scipion is the ability to to contrast results from multiple executions of the same step, even allowing to compare distinct implementations. Such programs are named as consensus and their primary intention is to automatically combine the results in the best way possible.

Currently, we are also working on a 3D classification consensus protocol which provides confidence scores to each of the classes and automatically selects the optimal number of classes based on this score.

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

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

[5] O. L. Zarrabeitia, "Development of a fast image alignment algorithm for cryo-electron microscopy", M.Sc. Thesis, Universidad Politécnica de Madrid, 2023.

[6] 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.

[7] S. Jonić, "Cryo-electron microscopy analysis of structurally heterogeneous macromolecular complexes", *Computational and Structural Biotechnology Journal*, vol. 14, pp. 385–390, 2016, ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2016.10.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037016300551.

[8] A. K. Sinop, *Introduction to mathematical physics*, Sep. 2008. [Online]. Available: https://www.math.arizona.edu/~tgk/541/chap1.pdf (visited on 12/29/2023).

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] X.-L. Meng and D. Van Dyk, "The EM Algorithm—an Old Folk-song Sung to a Fast New Tune", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 59, no. 3, pp. 511–567, Jan. 2002, ISSN: 0035-9246. DOI: 10.1111/1467-9868.00082. eprint: https://academic.oup.com/jrsssb/article-pdf/59/3/511/49588939/jrsssb_59_3_511.pdf. [Online]. Available: https://doi.org/10.1111/1467-9868.00082.

[28] S. Scheres, "Chapter six - processing of structurally heterogeneous cryo-em data in relion", in *The Resolution Revolution: Recent Advances In cryoEM*, ser. Methods in Enzymology, R. Crowther, Ed., vol. 579, Academic Press, 2016, pp. 125–157. DOI: https://doi.org/10.1016/bs.mie.2016.04.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0076687916300301.

[29] J. Gomez-Blanco, S. Kaur, M. Strauss, and J. Vargas, "Hierarchical autoclassification of cryo-em samples and macromolecular energy landscape determination", *Computer Methods and Programs in Biomedicine*, vol. 216, p. 106 673, 2022, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.106673. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016926072200058X.

[30] C. Sorzano, J. Bilbao-Castro, Y. Shkolnisky, *et al.*, "A clustering approach to multireference alignment of single-particle projections in electron microscopy", *Journal of Structural Biology*, vol. 171, no. 2, pp. 197–206, 2010, ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2010.03.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047847710000882.

[31] C. Sorzano, D. Semchonok, S.-C. Lin, *et al.*, "Algorithmic robustness to preferred orientations in single particle analysis by cryoem", *Journal of Structural Biology*, vol. 213, no. 1, p. 107 695, 2021, ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2020.107695. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047847720302689.

[32] Y. Zhou, A. Moscovich, and A. Bartesaghi, "Data-driven determination of number of discrete conformations in single-particle cryo-em", *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106 892, 2022, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.106892. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722002747.

[33] P. A. Penczek, M. Kimmel, and C. M. Spahn, "Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-em images", *Structure*, vol. 19, no. 11, pp. 1582–1590, 2011, ISSN: 0969-2126. DOI: https://doi.org/10.1016/j.str.2011.10.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969212611003571.

[34] D. Herreros, R. R. Lederman, J. M. Krieger, *et al.*, "Estimating conformational landscapes from cryo-em particles by 3d zernike polynomials", *Nature Communications*, vol. 14, no. 1, Jan. 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-35791-y. [Online]. Available: http://dx.doi.org/10.1038/s41467-023-35791-y.

[35] J. Schwab, D. Kimanius, A. Burt, T. Dendooven, and S. H. Scheres, "Dynamight: Estimating molecular motions with improved reconstruction from cryo-em images", *bioRxiv*, 2023. DOI: 10.1101/2023.10.18.562877. eprint: https://www.biorxiv.org/content/early/2023/10/19/2023.10.18.562877.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2023/10/19/2023.10.18.562877.

[36] "Game math: Swing-twist interpolation". (), [Online]. Available: https://allenchou.net/2018/05/game-math-swing-twist-interpolation-sterp/ (visited on 12/29/2023).

[37] T. Kurita, "Principal component analysis (pca)", *Computer Vision: A Reference Guide*, pp. 1–4, 2019.

[38] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis", *Journal of Signal and Information Processing*, vol. 4, no. 3B, p. 173, 2013.

[39] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library", in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[40] N. ALON, M. KRIVELEVICH, and B. SUDAKOV, "Maxcut in h-free graphs", *Combinatorics, Probability and Computing*, vol. 14, no. 5–6, pp. 629–647, 2005. DOI: 10.1017/S0963548305007017.

[41] H. Chan, *Lecture notes in max-cut, hardness of approximations*, Sep. 2014. [Online]. Available: https://www.cs.cmu.edu/afs/cs/academic/class/15854-f05/www/scribe/lec02.pdf (visited on 12/29/2023).

[42] A. K. Sinop, *Lecture notes in semidefinite programming and max-cut*, Feb. 2008. [Online]. Available: https://www.cs.cmu.edu/~anupamg/adv-approx/lecture14.pdf (visited on 12/29/2023).

[43] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems", *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.

[44] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization", *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[45] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[46] 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.

[47] 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.

[48] S. Dang, M. K. van Goor, D. Asarnow, *et al.*, *Cryo-EM structure of trpv5 full length in nanodisc*, Feb. 2019. DOI: 10.6019/empiar-10253. [Online]. Available: https://doi.org/10.6019/empiar-10253.

[49] 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.

[50] T. Nakane and S. H. W. Scheres, "Multi-body refinement of cryo-emcryo-electron microscopy (cryo-em) images in relionrelion", in *cryoEM: Methods and Protocols*, T. Gonen and B. L. Nannenga, Eds. New York, NY: Springer US, 2021, pp. 145–160, ISBN: 978-1-0716-0966-8. DOI: 10.1007/978-1-0716-0966-8_7. [Online]. Available: https://doi.org/10.1007/978-1-0716-0966-8_7.

[51] T. Nakane and S. H. W. Scheres, *Cryo-EM structure of a pre-catalytic spliceosome*, May 2017. DOI: 10.6019/empiar-10180. [Online]. Available: https://doi.org/10.6019/empiar-10180.

[52] S. M. Stagg, A. J. Noble, M. Spilman, and M. S. Chapman, "Reslog plots as an empirical metric of the quality of cryo-em reconstructions", *Journal of Structural Biology*, vol. 185, no. 3, pp. 418–426, 2014, ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2013.12.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047847713003377.

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

A.1 Introduction

This study is dedicated to introducing a novel approach to a complex issue in CryoEM, as is the case of 3D classification. Indeed, this is a very powerful tool employed by structural biologists for investigating proteins and their interactions. Such insights play a crucial role in the development of new drugs and vaccines, ultimately contributing to the improvement of citizens' quality of life.

Given the direct relation between the project and the pharmaceutical and biotechnology sectors, a comprehensive examination of the potential ethical, social, economic, and environmental impacts of the project is necessary. The objective of this section is to meticulously evaluate each of these facets and describe the potential impact of this algorithm.

A.2 Description of impacts related to the project

Social impacts

The social impacts of this project are primarily related to science. Research groups in the field of structural biology may be positively affected by the advances proposed on this work. Firstly, the results presented here show that the algorithm leads to superior results when it is used in conjunction with current solutions. In addition, this same tandem performs better

in terms of computational time. As a consequence, the project has the potential to increase the productivity of structural biologists by providing then with better results faster. This increased productivity of researchers leads to faster vaccine and drug developments, which has implications in the pharmaceutical and healthcare industry.

In addition, the project has been developed on an academic environment under a Free and Open Source Software (FOSS) licensing terms. This means that it can be used as a basis for further improvements or new fields of applications.

Environmental impacts

As mentioned earlier, one of the advances of the algorithm relate to the lower computational cost associated to the 3D classification process. This directly relates into a reduction of the energy consumption of the compute infrastructure.

Economic impacts

As a consequence of the previous arguments, the project poses a notable economic impact on research facilities. Firstly, the fact that this algorithm is distributed under a FOSS lices means that its utilization comes at no cost. Secondly, the project reduction in power consumption directly translates into lower power bills. Consequently, these two factors combine to decrease the operational expenses for various research groups.

A.3 Conclusions

In conclusion, this project offers numerous advantages to the scientific community and research institutes. Firstly, it shows potential aiding the development of drugs and vaccines. In addition, the project could yield positive outcomes for both the environment and the economy by minimizing the power consumption linked to the image processing pipeline in CryoEM.

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 THREE HUNDRED THIRTY-EIGHT AND ONE TENTH EUROS** ($39,338.10$€)

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 €
Logitech MX Ergo Mouse	125.00 €	6 months	36 months	20.83 €
Logitech MX Ergo Keys	99.00 €	6 months	36 months	16.50 €
Thinkpad T480 Laptop	875.70 €	6 months	36 months	145.95 €
Github Pro monthly subscription	4.00 €	6 months	1 months	24.00 €
Total				1,270.28 €

Total direct costs		28,270.28 €
Indirect costs	15 %	4,240.54 €
Budget subtotal		32,510.82 €

VAT	21 %	6,827.27 €

Total budget	39,338.10 €

Table B.1: Budget