“DCEIPHERING THE HETEROGENEITY OF THREE-DIMENSIONAL VOLUMES USING CRYO-ELECTRON MICROSCOPY”

Thesis Report

Submitted to

Indraprastha Institute of Information Technology, Delhi

In partial fulfillment of the requirement

for the Degree of

MASTER’S OF TECHNOLOGY

IN

COMPUTATIONAL BIOLOGY

Submitted By:

Chandan Saini

Under the supervision

Dr. Vibhor Kumar

2023

CERTIFICATE

This is to certify that the thesis titled "**DECIPHERING THE HETEROGENEITY OF THREE-DIMENSIONAL VOLUMES USING CRYO-ELECTRON MICROSCOPY**" being submitted by Chandan Saini to the Indraprastha Institute of Information Technology Delhi, for the award of the **Master of Technology (Computational Biology)**, is an original research work carried out by him under my supervision. The thesis has reached the standards fulfilling the requirements of the regulations relating to the Degree.

The results in this thesis have not been submitted in part or whole to any other university or institute to award any degree/diploma.

20 May 2023

Dr. Vibhor Kumar

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi 110 020

Acknowledgments

I am writing to express my sincere gratitude to my supervisor, Dr. Vibhor Kumar, for his guidance, support, and encouragement throughout this research project. With his expertise and patience, this thesis was possible. I want to thank my family for their love and support. Their unwavering belief in me has constantly motivated and inspired me. I also thank all the faculty members and staff of the Department of Computational Biology and IIIT Delhi for always helping us throughout our college journey. I am deeply grateful to all of you for your help and support.

INDEX

Abbreviations

Abbreviations	Title
Apix	Angstrom per pixel
CCD	Charged coupled device
Cryo-EM	Cryo-Electron Microscopy
CTF	Contrast Transfer function
DED	Direct electron detectors
EMAN2	Electron Microscopic Analysis 2
FEG	Field emission Gun
NMR	Nuclear Magnetic resonance
PCA	Principal component analysis
PDB	Protein data bank
SPR	Single particle reconstruction
SVD	Singular value decomposition
TEM	Transmission electron microscopy

List of Figures

Abstract

The cryo-electron microscopy (cryo-EM) technique captures 2D projections of biological structures, revealing their inherent Heterogeneity arising from structural flexibility, conformational changes, and distinct functional states. Factors such as ligand binding, conformational rearrangements, and variations in subunit composition can influence this variability. Accurate identification and understanding of this Heterogeneity play a pivotal role in unraveling the intricate structure-function relationships that govern biological processes and facilitating the development of targeted therapeutic interventions.

Researchers have developed numerous computational approaches to address the heterogeneity challenge in cryo-EM. These methods aim to extract distinct conformations from heterogeneous datasets, enabling the resolution of underlying structural variability. These approaches unveil hidden details and provide deeper mechanistic insights by generating high-resolution 3-D reconstructions of individual states within a heterogeneous sample.

In this research project, our objective was to identify the flexible and fixed regions of the LDL protein. We employed principal component analysis based on the singular value decomposition (SVD) algorithm, which enhances the resolution of the 3-D volume and facilitates the classification of 2-D images using EMAN2. Our study successfully obtained different 3-D volumes with improved resolution through multiple refinement iterations. Further analysis identified two highly variable principal components, allowing us to distinguish between the fixed and variable parts of the protein structure.

CHAPTER 1

INTRODUCTION & BACKGROUND

1.1Structural Biology

Structural biology is a field of science that utilizes principles and techniques from molecular biology, biochemistry, and biophysics to investigate the molecular structure and dynamics of biologically relevant molecules. Recent advances in instrumentation have provided new opportunities for structural biology, allowing for the analysis of complex biological molecules with unprecedented ease and efficiency. By determining the three-dimensional structure of proteins and protein complexes, scientists can gain essential insights into the mechanisms of life and the development of diseases, which can aid in the design of new diagnostic and therapeutic agents.[1] Scientists use three main techniques in structural biology: single-crystal X-ray diffraction, nuclear magnetic resonance, and cryo-electron microscopy. However, each technique has its advantages and limitations, and no one-size-fits-all method can be used for all research purposes.[2]

Structural biology came to light during the 1950s. Figure 1 indicates the recent structural proportion present on RCSB-PDB till April 2023, based on the biological methods they obtained: X-ray, Nuclear Magnetic Resonance (NMR), or Electron Microscopy (EM). There is a linear growth in the structure submission on RCSB-PDB based on Electron Microscopy Method after 2000.[3]

1.2 Different Methods in Structural Biology

A crystal's structure can be determined using X-ray crystallography. Primarily, single-crystal X-ray diffraction is used in the case of X-ray crystallography, where scattered beams are generated from the X-ray beam bombarded on the crystal atom. This method has its

disadvantage, where scientists have to crystallize large molecules, which is very challenging because of its large size and poor solubilization.[2, 3]

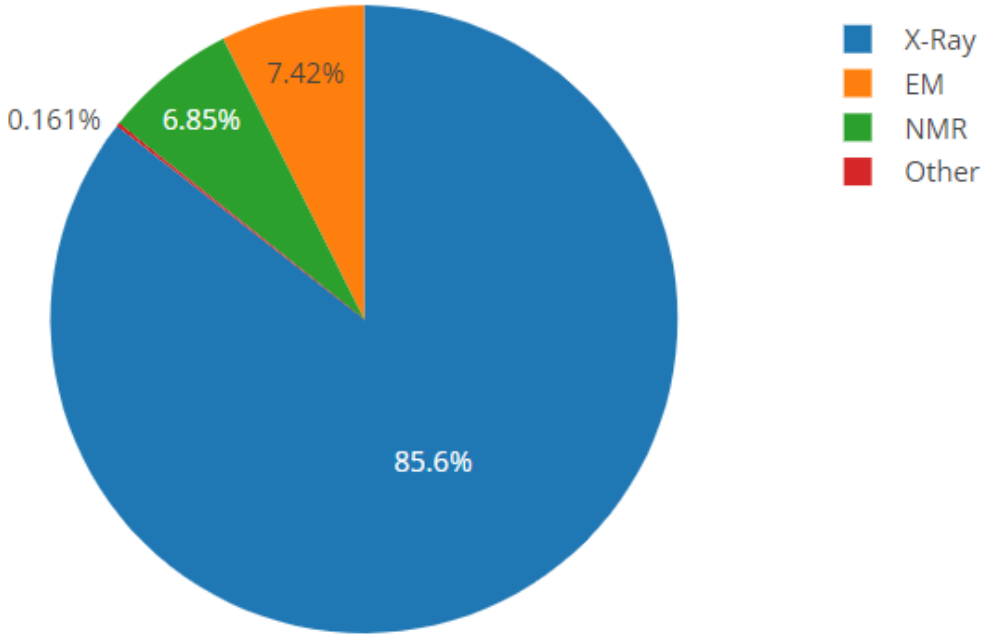

Figure 1. Proportion-wise Data of availability of Data on RCSB-PDB based on the biological Methods. The data was taken from RCSB-PDB. Blue indicates X-Ray, Orange indicates Electron Microscopy, Green Indicates Nuclear Magnetic Resonance, and Red as Other methods.

NMR is based on fats spinning and charged nuclei particles. The movement of the nuclei and their interaction with surrounding atoms gives the structural information of the given molecule. The NMR method also has its disadvantage as it requires a considerable amount of pure sample to avoid the noise level.[4]

The third and most crucial method of obtaining protein structure is Electron Microscopy (EM). Earlier, Transmission Electron microscopy (TEM) was used to obtain images from a stained sample, but the fixation method in TEM came out to cause damage to the samples and disturbed the final structure of the sample. To overcome this issue with TEM, scientists freeze samples at -196º C; the process is termed vitrification. The process involves rapidly freezing samples to avoid crystallization in water-based solutions. The complete process evolved as cryo-Electron Microscopy (cryo-EM), where the samples need not be fixed and stained. Also another advantage of vitrified sample is that it protects samples from beam damage during imaging.

The basic principle of the cryo-em includes the electron beam bombarded on the vitrified sample, which was captured by the detector as a magnified noisy image.[5]

There are mainly three sources of electron emission in the cryo-em process. Firstly and the best among all three is the Field emission gun (FEG) which generates the brightest and the most coherent electron beam due to its electron beam coming off from the little electrified tip. This is one of the most expensive sources of electrons in cryo-em. Secondly, the lanthanum hexaboride (LaB6) filament is cheaper than FEG but costlier than the tungsten filament. It gives a bright cohesive electron beam and can be used for a long time. At last, the Tungsten filament, used in most universities because of its comparatively cheap maintenance, provides good contrast for viewing.

In the case of cryo-em, Image capture technology plays a vital role as an electron emission source. There are only a few ways to capture and convert scattered electron signals to images coming from the sample. First, Charged Couple device (CCD) converts electrons to photons which, in

resultant, gives a noisy image with a loss of information. Due to this noise and loss of information, the resultant structure was of low resolution. Secondly, Direct electron detectors (DED) capture every individual electron when it passes through the thin membrane of the semi-conductor after striking the sample. This device collects data from electrons faster than CCD, resulting in a movie frame. The resultant movie frame helps scientists to track and correct the movement of particles affected by the electron beam. It is observed that DED, in combination with FEG, gives a high-resolution image and structure.

1.3 Single Particle Reconstruction

Single particle reconstruction using cryo-electron Microscopy (cryo-EM) is a technique that has revolutionized the field of structural biology. It involves the analysis of individual molecules or particles of different projections, which are imaged using an electron microscope after being frozen in vitreous ice. The resulting images contain much noise, which is then processed to produce three-dimensional (3D) reconstructions of the sample, providing detailed information on the structure and function of the molecule. To proceed with the SPR method, we should have thousands of images obtained from cryo-em with different projections.[6, 7]

Single-particle reconstruction involves several steps, each of which is critical in generating accurate and high-resolution 3D structures. Initiating from Motion correction and CTF correction to classification and alignment, all these steps are put together and result in a high-resolution 3D volume.[8] As shown in Figure 2.

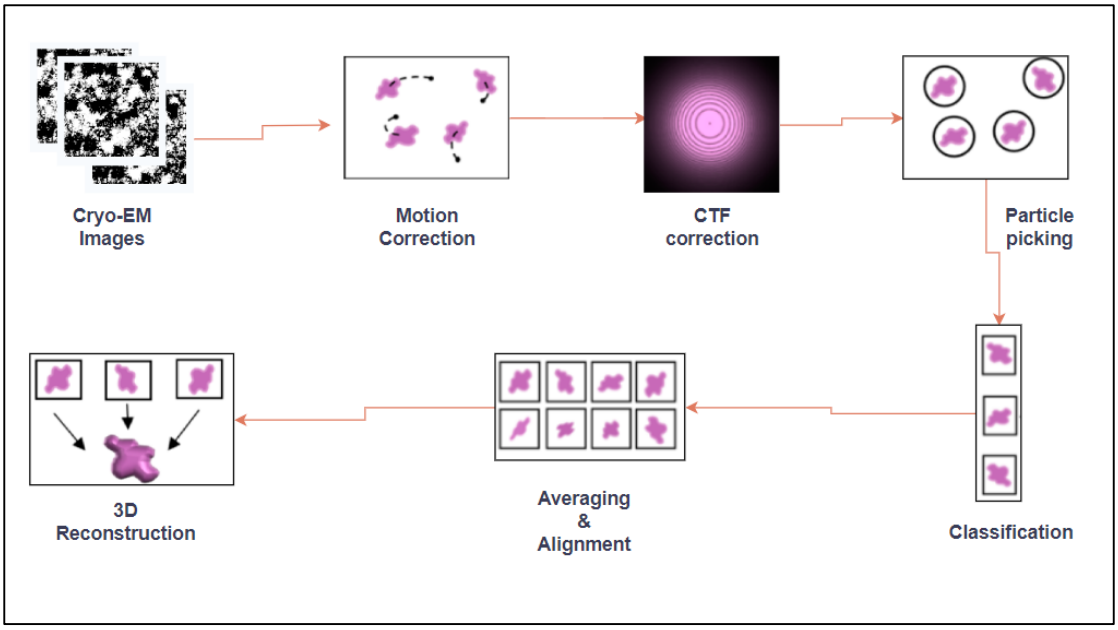

Figure 2: Conventional workflow for single particle reconstruction

1.3.1 Particle Picking

From thousands of raw images, selecting or picking particles from non-particles like frost, deformed particles, is essential. Particle picking can be done in two ways, either manual or automatic picking. Automatic picking involves thousands of particles from several micrographs, which becomes a laborious job. Automatic particle picking selects particles based on the correlation peak finding algorithm. The correlation value of the model and the micrograph was calculated, and a comparison between these indicates the region that contains particles; further, the region with a high correlation value is selected and extracts small square boxes, which are then further stacked into particle images. Some tools used a different approach than this, like, selecting particles based on the relative orientation of micrographs from the same samples; the only requirement is that the relative orientation of the two micrographs should be known.

1.3.2 CTF Determination and Motion Correction

While collecting images from a Cryo-em specimen, two main types of noise can be detected, which could be either due to contrast transfer or motion of the specimen. To overcome this noise, CTF correction, and Motion correction are performed. The CTF could be because of the microscope's optics on imaging the specimen. There could be various reasons for contrast transfer noise, like defocus, spherical aberration, and astigmatism, which further affect the resolution and contrast of the final image.

Similarly, heavy bombardment of the electrons could disturb the specimen, which can result in blurred images and inaccurate reconstruction. Motion correction algorithms are used to resolve this kind of noise, which align the images and remove any blur present. Motion correction is important to correctly align the images, which could lead to an inaccurate interpretation of the final 3D structure.[9]

1.3.3 2D Classification

Moving forward with the processing of cryo-EM images, classifications help a lot in eliminating poorly aligned and low-quality images. The 2D classification is based on the principle of reference-based classification. The complete image dataset is classified by comparing with the reference 2D projection, known as the cross-correlation algorithm. In order to represent each class, 2D class averages are generated, which further helps to guide through the 3D structure.

As there were multiple challenges in the cross-correlation algorithm, like noise or Heterogeneity in images, 2D classification evolved and started using a multi-reference algorithm and supervised classification. In multi-

reference algorithms, multiple sets of 2D projections are used as a reference and divide the images into a more accurate set of classes. As the Heterogeneity in the reference set increases, this algorithm will become hard to simplify. Supervised classification could be another promising approach where the user provides information about the images, like size, shape, or composition of the specimen, and based on this information, images will be more accurately classified.

While working with a standard set of images or high Heterogeneity, the maximum-likelihood algorithm works well where the probability distribution of the data is calculated, and a higher 2D average resolution is generated, which on iterative refinement provides better classification parameters.

1.3.4 Alignment

Alignment is needed to identify the orientation of each image concerning the typical reference frame. In general, particles are automatically aligned in the center of the square boxes after picking up particles, but this could not be reliable in the case of asymmetrical particles. Earlier, the projections were aligned based on two models, either by calculating the weighted average of pixel value at spatial locations or by translating the image so that there would be high-intensity peaks at the center. Nowadays, alignment is done by comparing the images from a reference model. The projection of images is compared to different projections of the reference model. This is called projection matching. This is an iterative process where the reference model gives different projections at every iteration and compares them with the set of projections. The alignment parameters are

refined after every iteration, and the process is repeated till convergence is reached. The final iteration will identify the best alignment parameters based on the cross-correlation, resulting in high-resolution 3D reconstruction.

1.3.5 Generation of initial 3D volume (3D reconstruction)

Three-dimensional reconstruction is the final phase of the single particle reconstruction, which started after getting 2D averages in the alignment step. This step involves the merging of multiple stacks of images at different projections. Various methods can be used to perform this task, like the Fourier-based and model-based techniques.

The Fourier-based technique work on the Fourier theorem, which states that any continuous function can be represented as a sum of sine and cosine functions of different frequencies. This principle leads to the amplitude and phase of the electron density in the specimen, which can further result in three-dimensional reconstruction.

Model-based methods involve the use of a model of the specimen to generate a three-dimensional reconstruction. The reference model can be from prior information like the size and shape of the specimen's structure. This way of processing requires multiple iterations to refine the model, taking images as a reference and resulting in a better three-dimensional reconstruction.

The final three-dimension structure needs to be validated if that precisely represents the biological specimen. Fourier shell correlation can be used to validate by comparing two independently reconstructed volumes and to approximate the final resolution of the reconstruction.

1.4 Application of Cryo-EM

Cryo-EM has numerous applications, like generating 3D structures, drug discovery, molecular imaging, etc.

1. Cryo-EM is used to generate three-dimensional structures of proteins. Cryo-EM is the most practiced method to generate the 3D structure of protein without noise instead of traditional X-ray crystallography.[10]
2. Cryo-EM is very effective in generating high-resolution structures, which can be further used for validating the drug-protein interaction. It will be helpful for the development of new effective drugs.[11]
3. Cryo-EM can be used to study the structure of viruses and other pathogens, which will be helpful in identifying foreign antigens' acts against immune response. This will help design effective vaccines against deadly infections.[12]

1.5 Challenges of Cryo-EM

1. Sample preparation: Biological samples should be purified and applied to a thin carbon film on the grid. Another essential task in cryo-em is to perform vitrification of samples at specific cryogenic temperatures to prevent ice crystals formation.[13, 14]
2. Image processing and analysis: Cryo-em images contain noise and artifacts. Removing that noise is tedious and should be performed sequentially, like Motion correction, CTF correction, particle picking, 2D classification, and so on, to obtain high-quality three-

dimensional structure. The images obtained from Cryo-em are of a heavy size and need heavy computational resources to visualize and analyze.

3. Three-dimensional model building and refinement: Model building and refinement needs multiple iterations to get the best parameter and refine the 3D model to its best. This could be time-consuming.[10]
4. Heterogeneity and size of biomolecules: Some biomolecules are hard to analyze and build 3D models because of their size and flexibility, as it sometimes becomes hard to obtain high-resolution 3D structures.[15]

1.6 Heterogeneity in the 3D structure of a biomolecule

Biomolecules, such as proteins and nucleic acids, function and interact with one another within biological systems in large part due to their three-dimensional form. Although we frequently think of biomolecules as permanent objects, their structural configurations exhibit great variability. Flexibility, interactions with ligands, post-translational modifications, environmental effects, genetic variability, and limits imposed by experimental methods are some causes of this variation. Understanding the causes and effects of this structural variability is essential for developing a deeper comprehension of biomolecules' biological significance and functional traits.[16, 17]

In order to determine the Heterogeneity of the biomolecule's fixed and variable part needs to be identified. This project aims to identify Heterogeneity in the LDL protein. Due to the considerable molecular weight of the protein, it is reported that LDL possesses different

conformations. Principal component analysis can be performed on the three-dimensional volumes to identify this.

To obtain the biomolecule's three-dimensional volumes, the cryo-em's raw images can be processed; multiple tools can process the images, like RELION, EMAN2, cisTEM. EMAN2 was used to perform all the processing steps in this project. Before processing the 3D volumes to identify the fixed and variable regions of the biomolecule, the 3D structure needs to be refined, which can also be done with the help of EMAN2. Multiple 3D volumes are generated with the help of reference models and then can be refined using EMAN2.

Chapter 2

Material & Methodology

2.1 Data Collection

Data were obtained from the thesis work of Dr. Vibhor Kumar at the Helsinki University of Technology, where they collected human plasma from healthy human donors with their consent. Human LDL particles were isolated from the samples. Samples were vitrified at 37 ºC and underwent cryo-electron Microscopy at 200 Kv and x 50000 magnification on an FEI Technical F20 field emission gun transmission electron microscope (Electron Microscopy Unit, Institute of Biotechnology, University of Helsinki, Finland).[18] Particles were picked from the micrographs and further processed for motion correction and astigmatism using EMAN's Boxer program. At particle picking, 29,844 images were obtained and used for further processing as part of this thesis.

2.2 Computational Requirements

Complete Data processing and visualization were done at IIITD using the open-source Linux-based operating system. The processing was done on the remote cell server of RegGen Lab. The server has the architecture of x86_64 and the processor of Intel Xeon Silver.

Python was used for the processing of multiple three-dimensional reconstructions. Python is an easy-to-use programming language with the latest version, 3.9. Multiple integrated development environments are freely available that support Python, like PyCharm, VS code, and Google Colab. All the necessary Python libraries, like pandas, numpy, and profile, were installed to import and process the data.

2.3 Tools Used for processing Cryo-EM Images

EMAN2 (Electron Microscopy Analysis 2)

EMAN2 is a powerful software package based on C\C++ programming language for the processing and analyzing cryo-EM data. It is an open-source project developed by a team of researchers at the National Centre for Macromolecular Imaging (NCMI) at Baylor College of Medicine. EMAN2 can easily be downloaded and installed. It works on the Linux operating system. EMAN2 provides a comprehensive set of tools for image processing, 3D reconstruction, and visualization of cryo-EM data. EMAN2 has a user-friendly interface that helps users perform a wide range of image-processing tasks with just preliminary programming knowledge.

EMAN2 has multiple algorithms which assist in single-particle reconstruction. These algorithms can be used to pre-process images, normalize contrast and defocus, and apply various types of filters to improve image quality. One of the commands used in the initial phase of single particle reconstruction is e2proc2d, which is used for pre-processing and normalizing 2D images. This command can be used for rescaling and filtering and have various other arguments which could be added.

Another key feature helpful in SPR is its alignment feature, which can be done with the e2refine2D command. The e2refine2D takes images as input and models for reference, which on multiple iterations, gives the best alignment results based on the correlation between images of the same particle.

EMAN2 offers several other features that can be used on its user interface and command line for particle picking, including e2boxer, e2make3dpar, and e2particlepicker.py. e2boxer is used to selection of particle images from micrographs usvariousy of methods, including manual selection and

automated threshold-based selection. e2make3dpar can be used to get parameter files of particle files which can guide a user for 3D reconstruction. e2particlepicker can be used for picking particles automatically from micrographs based on different arguments and thresholds.

After pre-processing and picking suitable particles from micrographs, EMAN2 can be further used for 3D reconstruction. Some of the commands available on EMAN2 for 3D reconstruction are e2initialmodel, e2refine, and e2refine_easy. e2initialmodel is used to generate an initial 3D model from aligned particle images. e2refine is used for the refinement of 3D volume over every iteration.

In conclusion, EMAN2 offers various powerful sets of features and commands for single particle reconstruction, which can be used at both small-scale and large-scale datasets with its dual way to use, either by user-friendly interface or command line.[19]

UCSF Chimera

UCSF chimera is a versatile open-source tool that offers multiple features for the visualization and analysis of 3D molecular structures. Its key feature includes a user-friendly interface. Chimera can handle large molecular structures and complex datasets. It can be used to import large molecular structures of different extensions. It can be used to view structures from different angles from inside and outside with different visualization options like solid, mesh, or surface. Chimera can be used for different purposes like model building and refinement, which can further improve the accuracy of 3D volumes. Chimera can also be used to analyze and compare multiple models at a time, like superimposition, grouping, segmentation,

and alignment. UCSF chimera includes features for calculating electrostatic potential and electrostatic surface potential maps. These features can be used for identifying protein-ligand interactions and different active sites present in 3D molecular structures.

Chimera can also be used to automate different tasks using its command line tool in case of long tasks. Along with other features, chimera also allows integration processing with other software packages like EMAN2 and RELION.

Its multiple features, tools, and user-friendly interface allow users to perform visualization, modeling, and refinement on 3D volumes.[20]

2.4 Data Preparation

2.4.1 Visualisation of Cryo-EM Images

EMAN2 can visualize cryo-em images of extensions like ".tiff" or ".mrc." Thousands of raw images obtained from cryo-em were in “.tiff” format, as shown in Fig 3. Two-dimensional images of LDL protein were visualized in EMAN2 software. The images contain noise which can be removed further with EMAN2.

Figure 3: One of the raw image

1.4.2 Pre-processing of Cryo-EM Images

The original cryo-em was in ".tiff" format, which was converted to ".mrc." The MRC format works best to process with the EMAN2 features. The cryo-EM images were of random size and dimensions, which were clipped to uniform dimensions of 98*98*98 using clip argument in e2proc2d.py. The angstrom per pixel of the images was set at 1.4 Aº. Dimensions of the images can also be changed using mean shrink or media shrink.

2.4 Data Processing

2.5.1 Three-dimensional models from raw 2D images

Cryo-em MRC images generated an initial three-dimensional model after multiple refinement iterations. In the case of e2initialmodel.py, the input is a set of a few class averages instead of full-size images. The output of e2initialmodel.py could be of low resolution, but after visualizing it, the model can be used further for the refinement feature of EMAN2 for higher resolution.

Multiple initial models were generated with e2initialmodel with EMAN2. Those models were used as reference models to generate better-resolution models. This process can be done with "e2refine_easy.py" of EMAN2. E2refine_easy was processed with five iterations and generated multiple three-dimensional volumes with different reference models.

2.5.2 Identifying Heterogeneity in the three-dimensional volume

Refined multiple models were used further to obtain the two best principal components using a Python package named 'profile' as shown in the workflow, fig 4.

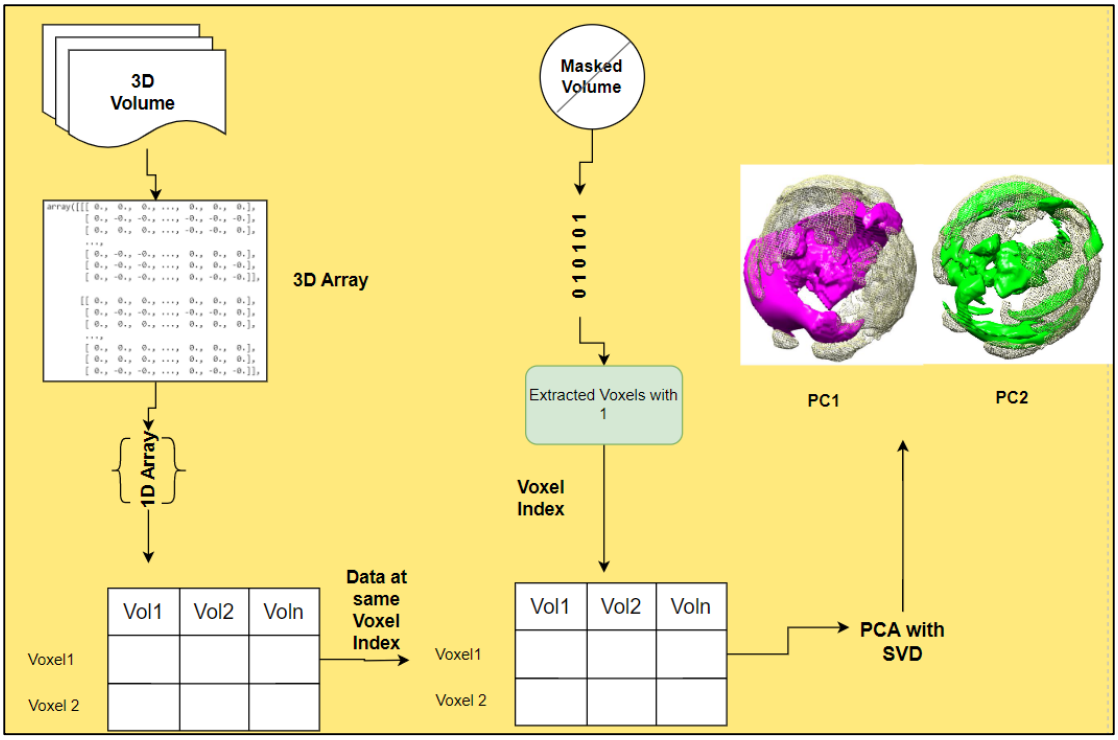

Figure 4: Workflow for identification of Heterogeneity in a biomolecule

All the 3d volumes were imported using "profile.read," which was read as a 3D numpy array. All the 3D array of different volumes was reshaped into a 1D array and merged as a data frame. The data frame generated represents

volumes as columns and rows as voxels for each volume. Secondly, one of the refined models with better structure is used to mask the lipid core, and the masked part's voxel was labeled as 1, the rest as 0. The masked model was also imported and converted to a one-dimensional numpy array. From the masked model, the masked part, which was labeled as 1, was extracted, and the voxels where the mask was generated were extracted from the merged data frame of refined volumes. The extracted voxel was used further to obtain principal component 1 and principal component 2 using Singular Value Decomposition (SVD) algorithm.

SVD is a matrix decomposition technique that divides the matrix into three matrices U, Σ, and VT. 'U' represents the left singular matrix, which indicates the orthogonal basis for the rows of the original matrix. It is a 1D array corresponding to the columns of the original matrix. 'Σ' or 's' represents the diagonal matrix. The singular values indicate the importance of each singular vector. 'VT' represents the transposition of the right singular matrix. It contains the singular vectors corresponding to the rows of the original matrix. After computing this matrix, principal components were calculated after multiplying the first and second columns of 'U' by the first and second singular values to obtain the first two principal components.

$$PC1 = U[:, 0] * s[0]$$

$$PC2 = U[:, 1] * s[1]$$

The first two principal components were taken to reduce the dimension of the original data matrix, which contains the highest Degree of variability or information.

These principal components were used to generate 3D volumes and visualized in the UCSF chimera to identify the rigid and flexible part of the volume.

Chapter 3

Results & Discussion

3.1 Initial model

To obtain multiple initial models, e2initialmodel was used, and different reference models were generated of uniform shape, size, apex, and dimension. Apix (Angstrom per pixel) was set as 2.8, and the dimensions of the images were kept as 98*98*98. A few reference models are shown below in Figure 5 (a,b,c,d).

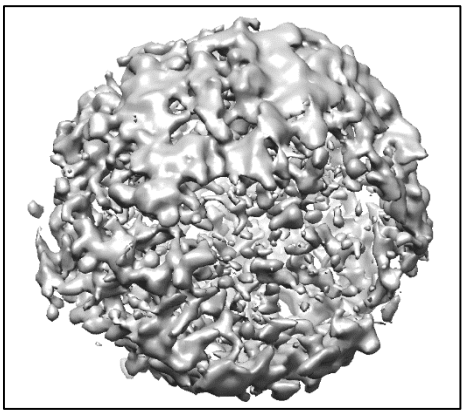

Figure 5(a)

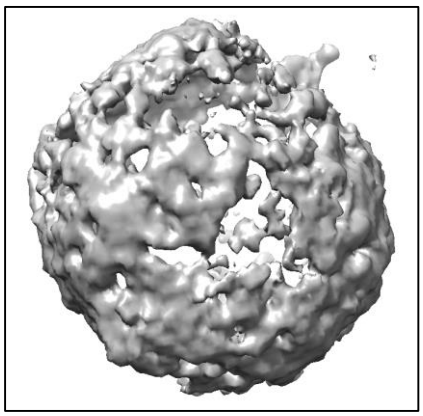

Figure 5(b)

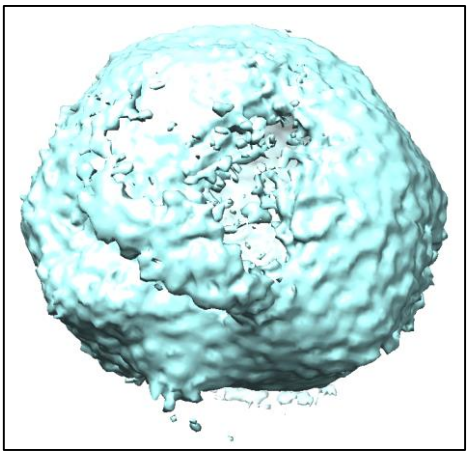

Figure 5(c)

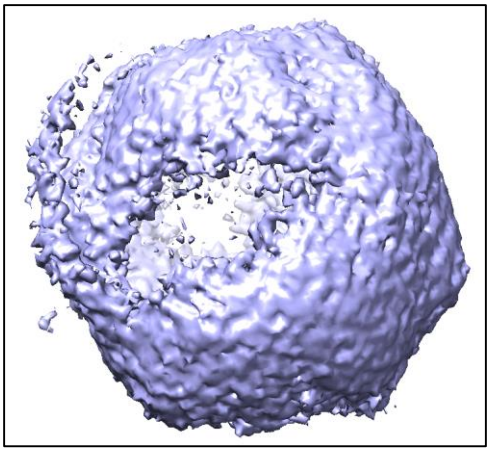

Figure 5(d)

Figure 5: Represents a few of the initial models which were used as a reference for refinement

3.2 Refined model

Initial models of uniform dimensions and apix were used as input for 'e2refine_easy.py'. e2refine_easy was processed for five iterations at 2.8 angstroms per pixel. In this project, we generated 39 initial models and refined models. Some of the refined models are shown below in Figure 6 (a,b).

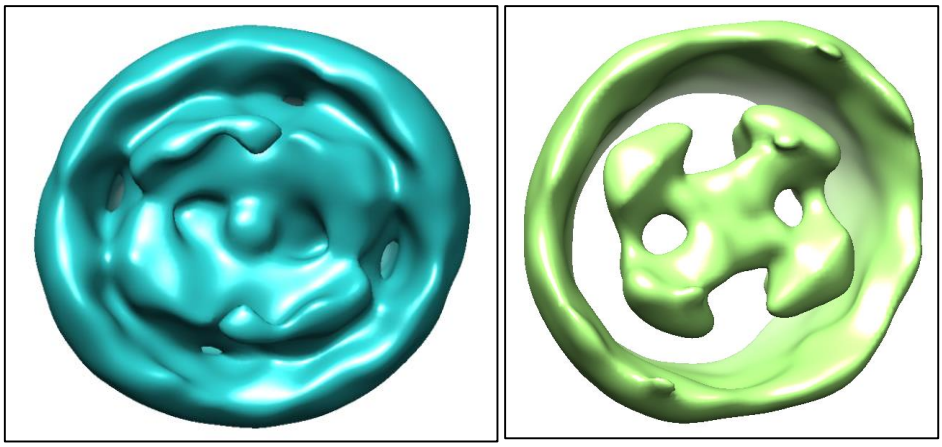

Fig 6a Fig 6b

Figure 6: Represents 3D volumes after refinement

3.3 Eigenvector decomposition

Generated 3D volumes were used to perform principal component analysis from the extracted voxels with the help of a masked model. The principal component was obtained from the SVD algorithm. The principal components were converted back to 3D volume by keeping the remaining voxels as 0. The 3D volumes were visualized in a chimera. The obtained 3D volumes of principal components were superimposed with the three-

dimensional reference model and identified that structure of the outer layer of LDL and inner lipid core are dependent, which was easily visualized. In principal component 1 (Fig 7a), the inner lipid core was attached to the outer phospholipid layer, due to which the outer part of the LDL is present in a more secure manner. In contrast, in the case of principal component 2 (fig 7b), the inner lipid core is free from outer phospholipid layer due to which the outer part is present in free form.

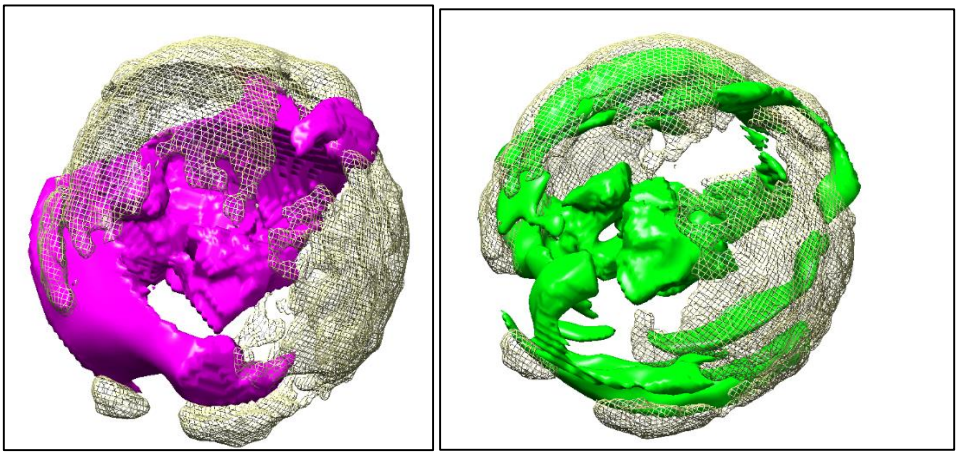

Figure 7a Figure 7b

Figure 7: Represents 3D volume after principal component analysis, PC1 (pink), PC2 (green).

Principal components 3 and 4 were also calculated and visualized in chimera, where the lipid core was found free from the outer phospholipid layer.

Chapter 4

Conclusion & Future Scope

The thesis "Deciphering the heterogeneity of three-dimensional volumes using cryo-electron microscopy" delved into the impact of the Heterogeneity of 3D volumes, which could often be observed in the heavy biomolecule. To achieve this purpose, 2D raw cryo-em images of LDL protein were taken and processed using the EMAN2 tool to remove noises and generate multiple reference models; those reference models were further used to generate three-dimensional refined 3D volumes.

The refined 3D volumes were analyzed and processed as a numpy array in Python to identify the principal component 1 and 2. The obtained principal component array was converted to 3D volume and visualized in a chimera. When comparing these volumes with reference models, it was concluded that the inner core of the LDL is variable, which alters its structure of it. Due to this, there could be multiple conformations of LDL, and the approach followed in this project could be followed to identify variability in the 3D structure.

Future Scope

1. Investigate the potential of PC1 and PC2 for aligning larger datasets and diverse image classes, which could further lead to better resolution of the 3D volume.

2. Evaluate the impact of PC1 and PC2 alignment on downstream image analysis and interpretation.

References

[1] X.-C. Bai, G. McMullan, and S. H. Scheres, "How cryo-EM is revolutionizing structural biology," *Trends in biochemical sciences,* vol. 40, no. 1, pp. 49-57, 2015.

[2] A. A. Kermani, "A guide to membrane protein X-ray crystallography," *The FEBS journal,* vol. 288, no. 20, pp. 5788-5804, 2021.

[3] M. Smyth and J. Martin, "x Ray crystallography," *Molecular Pathology,* vol. 53, no. 1, p. 8, 2000.

[4] P. R. Markwick, T. Malliavin, and M. Nilges, "Structural biology by NMR: structure, dynamics, and interactions," *PLoS computational biology,* vol. 4, no. 9, p. e1000168, 2008.

[5] E. Nogales, "The development of cryo-EM into a mainstream structural biology technique," *Nature Methods,* vol. 13, no. 1, pp. 24-27, 2016.

[6] J. L. Milne *et al.*, "Cryo-electron microscopy–a primer for the non-microscopist," *The FEBS journal,* vol. 280, no. 1, pp. 28-45, 2013.

[7] J. Frank, "Single-particle reconstruction of biological macromolecules in electron microscopy–30 years," *Quarterly reviews of Biophysics,* vol. 42, no. 3, pp. 139-158, 2009.

[8] S. J. Ludtke, "Single-particle refinement and variability analysis in EMAN2. 1," *Methods in enzymology,* vol. 579, pp. 159-189, 2016.

[9] M. Czarnocki-Cieciura and M. Nowotny, "Introduction to high-resolution cryo-electron microscopy," *Postępy Biochemii,* vol. 62, no. 3, pp. 383-394, 2016.

[10] D. Agard, Y. Cheng, R. M. Glaeser, and S. Subramaniam, "Single-particle cryo-electron microscopy (cryo-EM): Progress, challenges,

and perspectives for further improvement," *Advances in imaging and electron physics,* vol. 185, pp. 113-137, 2014.

[11] S. Subramaniam, L. A. Earl, V. Falconieri, J. L. Milne, and E. H. Egelman, "Resolution advances in cryo-EM enable application to drug discovery," *Current opinion in structural biology,* vol. 41, pp. 194-202, 2016.

[12] J. A. Lees, J. M. Dias, and S. Han, "Applications of Cryo-EM in small molecule and biologics drug design," *Biochemical Society Transactions,* vol. 49, no. 6, pp. 2627-2638, 2021.

[13] M. Zielinski, C. Röder, and G. F. Schröder, "Challenges in sample preparation and structure determination of amyloids by cryo-EM," *Journal of Biological Chemistry,* vol. 297, no. 2, 2021.

[14] H.-W. Wang and X. Fan, "Challenges and opportunities in cryo-EM with phase plate," *Current Opinion in Structural Biology,* vol. 58, pp. 175-182, 2019.

[15] D. Lyumkis, "Challenges and opportunities in cryo-EM single-particle analysis," *Journal of Biological Chemistry,* vol. 294, no. 13, pp. 5181-5197, 2019.

[16] J.-C. Hus and R. Brüschweiler, "Principal component method for assessing structural heterogeneity across multiple alignment media," *Journal of Biomolecular NMR,* vol. 24, pp. 123-132, 2002.

[17] E. D. Zhong, T. Bepler, J. H. Davis, and B. Berger, "Reconstructing continuous distributions of 3D protein structure from cryo-EM images," *arXiv preprint arXiv:1909.05215,* 2019.

[18] V. Kumar *et al.*, "Three-dimensional cryoEM reconstruction of native LDL particles to 16Å resolution at physiological body temperature," *PloS one,* vol. 6, no. 5, p. e18841, 2011.

[19] G. Tang *et al.*, "EMAN2: an extensible image processing suite for electron microscopy," *Journal of structural biology,* vol. 157, no. 1, pp. 38-46, 2007.

[20] E. F. Pettersen *et al.*, "UCSF Chimera—a visualization system for exploratory research and analysis," *Journal of computational chemistry,* vol. 25, no. 13, pp. 1605-1612, 2004.