TEL AVIV UNIVERSITY

Tel-Aviv University

The Raymond and Beverly Sackler
Faculty of Exact Sciences
School of Mathematical Sciences

Cryo-EM Symmetry Detection Problem

Thesis submitted in partial fulfilment of the requirements
for the M.Sc. degree of Tel-Aviv University

by
Maayan Yesharim

This thesis was prepared under the supervision of
Prof. Nir Sharon and Prof. Yoel Shkolnisky

October 2024

Acknowledgement

I would like to express my sincere gratitude to my supervisors, Professor Nir Sharon and Professor Yoel Shkolnisky, for their invaluable guidance, support, and expertise throughout the preparation of this thesis. Their insights and encouragement have been instrumental in shaping this work.

I am particularly indebted to Ido Hadi, whose generous assistance and mentorship were crucial during challenging phases of my research. His help in interpreting my results and providing direction when I felt stuck was invaluable, and I am deeply grateful for his support.

I would like to express my appreciation to the faculty and staff of the Department of Applied Mathematics at Tel Aviv University for providing a stimulating academic environment and the resources necessary for my research.

Finally, I wish to thank my family and friends for their unwavering support and encouragement throughout my studies.

Contents

List of Figures

Introduction

Single-particle cryo-electron microscopy (cryo-EM) is a powerful imaging technique that enables the determination of high-resolution three-dimensional structures of biological macromolecules without the need for crystallization [4]. This method captures two-dimensional projections of molecules from multiple unknown viewing angles, effectively creating tomographic images. The cryo-EM reconstruction process then faces the challenge of determining the three-dimensional structure from these noisy two-dimensional projections. By overcoming this key hurdle, cryo-EM has emerged as a revolutionary tool capable of revealing near-atomic resolution details of complex biological structures [3].

Many biological molecules exhibit symmetry, which means certain structural features repeat themselves in regular patterns within the molecule. Under the assumption that the molecule has no symmetry, an *ab-initio* reconstruction algorithm has been introduced in [16], and there is a series of studies developing theories and algorithms for various cryo-EM-related tasks [6, 7, 17, 15, 20]. When a molecule with symmetry is present, another problem arises when attempting to reconstruct the three-dimensional structure from its two-dimensional images: the same two-dimensional projections of a single structure allow for the generation of different three-dimensional structures when imposing different symmetries [21]. This ambiguity compels us to accurately determine the underlying symmetry of the molecule during the reconstruction process.

The common practice today when addressing the presence of symmetries is to have the user provide the symmetry information when attempting a reconstruction [21, 8]. This prior knowledge is then used to guide the reconstruction process, as demonstrated in [21, 16]. While knowing the symmetry can improve the quality of the reconstructed structure, this reliance on user-supplied symmetry information can be a limiting factor, particularly when the symmetry is uncertain or unknown.

To address the limitations of user-supplied symmetry information, there is a growing need for data-driven approaches that can detect the symmetry group of molecules directly from cryo-EM projection images. This work, based on the foundational approach of [2], proposes such a method, aiming to determine molecular symmetry from the data itself, thus advancing the field of cryo-EM structure de-

termination. Our approach eliminates the dependence on prior symmetry assumptions, potentially leading to more accurate and unbiased reconstructions. Related work that suggests a solution for the symmetry detection problem is [14], based on the principle of self-consistency under symmetry operations. Their approach analyzes 3D density maps by applying various symmetry transformations and evaluating whether the transformed map coincides with its symmetrized self.

Our method involves estimating the auto-correlation function of the molecular volume from the projection data and analyzing its rank to determine the underlying symmetry group. The main contributions of this thesis are

1. Formulating the symmetry detection problem in cryo-EM.

2. Presenting an auto-correlation approach to symmetry detection.

3. Implementing a symmetry detection solver and supporting code.

4. Numerically studying the symmetry detection problem via simulations

The thesis is organized as follows. Chapter 1 mathematically defines the problem of symmetry detection and explains the idea for the solution. Chapter 2 presents the theoretical and analytical foundations for the proposed solution. Chapter 3 specifies the algorithm in detail and showcases experimental results. Chapter 4 is conclusions and future work.

Chapter 1

The Symmetry Detection Problem in Cryo-EM

1.1 The Cryo-EM Imaging Process

Recovering the three-dimensional structure of a molecule using cryo-EM involves several steps. Initially, a sample containing multiple copies of the molecule is rapidly frozen to preserve its natural state. Then, this frozen sample is imaged using an electron microscope, which results in two-dimensional projection images of the sample, taken from random orientations. Third, the two-dimensional images are combined into a three-dimensional model of the molecule.

Mathematically, the molecular structure is described by a scalar function $V(\vec{r})$, where $\vec{r} = (x, y, z)^T \in [-1/2, 1/2]^3$. This function V stands for *volume* and evaluates the electric potential induced by the molecule in the three-dimensional space. A molecule exhibits symmetry if there exists a subgroup $G \subseteq \mathrm{SO}(3)$ (for a list of all possible subgroups of $\mathrm{SO}(3)$ see Appendix A) such that

$$V(\vec{r}) = V(g\vec{r}) \quad \text{for any } g \in G. \tag{1.1}$$

In this case, we say that G is the symmetry group of V.

In the cryo-EM recovery process, we are given a large set of images $P_1, \ldots, P_N$, where each image is formed by projecting the volume $V(\vec{r})$ in some unknown direction $R_i \in \mathrm{SO}(3)$. Specifically, the pixel intensities in $P_i(x, y)$ are given by

$$P_i(x, y) = \int_{-\infty}^{\infty} V(R_i \vec{r}) \, dz, \tag{1.2}$$

where $\vec{r} = (x, y, z)^T$. In practice, each P_i is further corrupted by noise, which poses a significant challenge for the reconstruction process.

1.2 The Problem of Symmetry Detection

The problem of symmetry detection in cryo-EM is defined as follows: *Given N cryo-EM images $P_1, \ldots, P_N$, taken from an unknown volume V, infer the symmetry group $G \subseteq SO(3)$ of V.*

This problem has been hardly considered in the literature and the only relevant prior work seems to be [14] which proposes based on the principle of self-consistency under symmetry operations. This technique directly analyzes 3D density maps by applying symmetry transformations and checking if the transformed map coincides with its symmetrized version. This approach overcomes the limitations of 2D projection-based methods and can be applied at any resolution level. This work thus represents a significant advancement in symmetry detection for 3D reconstructions from cryo-EM data.

Our approach introduces a different strategy by utilizing the l-th order auto-correlation function of the cryo-EM projections. This function, defined later in Eq. (2.6), has a *rank* that encodes the symmetry group of the volume. The use of auto-correlation functions for structure reconstruction was initially suggested in [10], and we extend this idea to the setting of symmetry detection, following [2].

While the method in [14] employs a statistical approach, analyzing symmetry by transforming 3D density maps and checking for self-coincidence, our approach tries to investigate the feasibility of determining symmetry directly from the cryo-EM images.

Moreover, our approach tries to lay the groundwork for developing a more comprehensive statistical framework in future work. By first confirming the method's robustness under ideal conditions, we aim to extend it to more complex scenarios involving noise and real data. This progression could form a strong basis for integrating statistical techniques to enhance symmetry detection in challenging cryo-EM datasets.

Chapter 2

Solving The Symmetry Detection Problem: Theory

In this chapter, we explore the theoretical framework for detecting symmetry in molecular structures using projection images. Following the thesis of [2], we adapt the concept of *spatial correlation* originally proposed in [10] to the context of symmetry detection.

2.1 The l-th Order Auto-Correlation Function

We introduce the auto-correlation function based on the theory in [10], and the notation of the spherical Fourier transform as in [18].

Let $\hat{V}(r, \theta, \phi)$ be the three-dimensional Fourier transform of the function V, defined on a spherical region where $r \leq a$. The function $\hat{V}(r, \theta, \phi)$ can be expanded in terms of the basis functions $\Psi_{lms}(r, \theta, \phi)$ [18]:

$$\hat{V}(r, \theta, \phi) = \sum_{s=1}^{\infty} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lms} \Psi_{lms}(r, \theta, \phi), \qquad (2.1)$$

where

$$\Psi_{lms}(r, \theta, \phi) = j_{ls}(r) Y_{lm}(\theta, \phi). \qquad (2.2)$$

Here, $Y_{lm}(\theta, \phi)$ represents the spherical harmonics, which describe the angular dependence of the function, and the radial component is given by

$$j_{ls}(r) = \frac{\sqrt{2}}{\sqrt{a^3} j_{l+1}(R_{ls})} j_l \left(\frac{R_{ls} r}{a} \right), \qquad (2.3)$$

where $j_l(r)$ is the spherical Bessel function of order l and R_{ls} is the s-th root of $j_l(r)$.

3

According to the Fourier Slice Theorem [13], the Fourier-transformed projection at orientation $\omega \in SO(3)$ is

$$\hat{P}(r,\phi)^\omega = \sum_{s=1}^\infty \sum_{l=0}^\infty \sum_{m=-l}^l \sum_{m'=-l}^l A_{lms} \Psi_{lms}(r,\theta,\phi) D_{m,m'}^l(\omega) \Big|_{\theta=\frac{\pi}{2}} \qquad (2.4)$$

where $D_{m,m'}^l(\omega)$ is the Wigner-D matrix of order l for rotation $\omega \in SO(3)$ [9]. The auto-correlation function is defined as

$$C(r_1, r_2, \phi_1, \phi_2) = \int_{SO(3)} \left\{ \hat{P}^\omega(r_1, \phi_1) \cdot \overline{\hat{P}^\omega(r_2, \phi_2)} \right\} d\omega$$

$$= \sum_{s_1,s_2} \sum_{l,m} \frac{1}{2l+1} A_{lms_1} \Psi_{lms_1}\left(r_1, \frac{\pi}{2}, \phi_1\right) \overline{A_{lms_2} \Psi_{lms_2}\left(r_2, \frac{\pi}{2}, \phi_2\right)}$$

$$= \sum_{s_1,s_2} \sum_{l,m} \frac{1}{4\pi} P_l(\cos\psi) A_{lms_1} j_{ls_1}(r_1) \overline{A_{lms_2} j_{ls_2}(r_2)},$$

$$(2.5)$$

where $d\omega$ is the Haar measure, $\psi = \phi_1 - \phi_2$, and P_l is the Legendre polynomial of degree l. Using the orthogonality of $P_l(\cos\psi)$, we define the l-th order auto-correlation as follows:

$$C_l(r_1, r_2) = 2\pi(2l+1) \int_0^\pi C(r_1, r_2, \phi_1, \phi_2) P_l(\cos\psi) \sin\psi \ d\psi$$

$$(2.6)$$

$$= \sum_{s_1,s_2} \sum_{m=-l}^l A_{lms_1} A_{lms_2}^* j_{ls_1}(r_1) j_{ls_2}(r_2).$$

Finally, we expand $C_l(r_1, r_2)$ using the functions $\{j_{ls}(r) : s = 1, 2, ...\}$ defined in Eq. (2.3), which form an orthonormal basis of the space $L_2([0,a], r^2 dr)$. Due to their orthogonality, we obtain:

$$(C_l)_{s_1,s_2} = \int_0^a \int_0^a C_l(r_1, r_2) \cdot j_{ls_2}(r_2) \ r_2^2 \ dr_2 \cdot j_{ls_1}(r_1) \ r_1^2 dr_1$$

$$= \sum_{s_1',s_2'} \sum_{m=l}^l A_{lms_1'} A_{lms_2'}^* \int_0^a j_{ls_1'}(r_1) j_{ls_1}(r_1) \ r_1^2 dr_1 \int_0^a j_{ls_2'}(r_2) j_{ls_2}(r_2) \ r_2^2 dr_2$$

$$= \sum_{m=l}^l A_{lms_1} A_{lms_2}^*.$$

$$(2.7)$$

Table 2.1: The rank of C_l according to [5]. The first two members of the cyclic groups and dihedral groups are shown.

G/l	0	1	2	3	4	5	6	7	8	9	10
No Symmetry	1	3	5	7	9	11	13	15	17	19	21
C_2	1	1	3	3	5	5	7	7	9	9	11
C_4	1	1	1	1	3	3	3	3	5	5	5
D_2	1	0	2	1	3	2	4	3	5	4	6
D_4	1	0	1	0	2	1	2	1	3	2	3
T_{12}	1	0	0	1	1	0	2	1	1	2	2
O_{24}	1	0	0	0	1	0	1	0	1	1	1
I_{60}	1	0	0	0	0	0	1	0	0	0	1

Eq. (2.7) allows us to express the l-th order correlation of a volume in terms of its spherical Fourier-Bessel expansion coefficients.

2.2 Classification of Symmetry via l-th Order Auto-Correlation

As stated above, the problem is to identify, from finitely many projections, the symmetry group of the molecule as defined in Eq. (1.1). This symmetry group is a subgroup of $SO(3)$. Our approach is to estimate $(C_l)_{s_1,s_2}$, the l-th order auto-correlation function of the volume, (as defined in Eq. (2.7)) from the projections. Following [2] and [5], we know that the rank of C_l encodes the type of symmetry group that the volume possesses. This classification "rank map" is provided in Table 2.1, which is based on the theory in [5], and page 89 in [2].

In the following section, we present the theory for approximating C_l from the projection images. Using this approximation, we propose the following classification algorithm: estimate $(C_l)_{s_1,s_2}$ for a sufficient range of l values and then classify the symmetry using the rank map provided in Table 2.1.

2.3 Approximating $(C_l)_{s_1,s_2}$ From Projections

Given N projections $P_1, \ldots, P_N$, our goal is to approximate $(C_\ell)_{s_1,s_2}$ using only the images. Let the Fourier transformed projections be $\hat{P}_i$, $i = 1, \ldots, N$. Following

the approach in [19], we expend each $\hat{P}_i$ in the Fourier-Bessel basis as

$$\hat{P}_i(r, \phi) = \sum_{k=-\infty}^{\infty} \sum_{q=1}^{\infty} \frac{1}{\sqrt{2\pi}} a_{kq}^i J_{kq}(r) e^{ik\phi}, \tag{2.8}$$

where

$$J_{kq}(r) = \frac{\sqrt{2}}{\sqrt{a} J_{k+1}\left(R_{kq}\right)} J_k\left(\frac{R_{kq} r}{a}\right). \tag{2.9}$$

Here, $J_k(r)$ is the Bessel function of order k, and R_{kq} denotes the q-th root of $J_k(r)$. The expansion coefficients a_{kq}^i are given by

$$a_{kq}^i = \int_0^a \int_0^{2\pi} P_i(r, \phi) J_{kq}(r) e^{-ik\phi} \, r \, d\phi \, dr. \tag{2.10}$$

We define the empirical auto-correlation function, which is invariant under rotation and mirror reflection, as:

$$C^{(proj)}\left(r_1, r_2, \phi_1, \phi_2\right) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \int_0^{2\pi} \frac{1}{2\pi} \left\{ \hat{P}_i(r_1, \phi_1 + t) \cdot \overline{\hat{P}_i(r_2, \phi_2 + t)} \right.$$
$$\left. + \hat{P}_i^r(r_1, \phi_1 + t) \cdot \overline{\hat{P}_i^r(r_2, \phi_2 + t)} \right\} dt, \tag{2.11}$$

where $\hat{P}^r$ is the mirror-reflected projection, i.e. $\hat{P}^r(r, \phi) := \hat{P}(r, 2\pi - \phi)$. Expanding $\hat{P}_i$ using the Fourier-Bessel basis of Eq. (2.8), and substituting this expansion into the definition of $C^{(proj)}$ in Eq. (2.11), we obtain

$$C^{(proj)}\left(r_1, r_2, \phi_1, \phi_2\right) =$$
$$\sum_k \sum_{q_1, q_2} \left(\frac{1}{N} \sum_i a_{kq_1}^i \overline{a_{kq_2}^i} \right) \cdot J_{kq_1}(r_1) J_{kq_2}(r_2) \cdot \frac{1}{2\pi} \cos\left(k(\phi_1 - \phi_2)\right). \tag{2.12}$$

Next, analogously to Eq. (2.6), we define the empirical ℓ-th order auto-correlation function as

$$C_l^{(proj)}\left(r_1, r_2\right) = 2\pi \left(2l + 1\right) \int_0^{\pi} C^{proj}\left(r_1, r_2, \phi_1, \phi_2\right) P_l\left(\cos\psi\right) \sin\psi \, d\psi$$
$$= \sum_{k=-\infty}^{\infty} \sum_{q_1, q_2} \left(\frac{1}{N} \sum_i a_{kq_1}^i \overline{a_{kq_2}^i} \right) \cdot J_{kq_1}(r_1) J_{kq_2}(r_2) \cdot \alpha_{kl}, \tag{2.13}$$

6

where
$$\alpha_{kl} = \int_0^\pi \cos\left(k\psi\right)(2l+1)P_l(\cos\psi)\sin\psi d\psi. \tag{2.14}$$

Notice that for $\ell > 0$, $\alpha_{0l} = 0$. Therefore, we can write

$$
C_\ell^{(proj)}\left(r_1, r_2\right) = \sum_{k=1}^\infty \alpha_{kl} \sum_{q_1, q_2} \left(\frac{1}{N} \sum_i a_{kq_1}^i \overline{a_{kq_2}^i} + \frac{1}{N} \sum_i a_{-kq_1}^i \overline{a_{-kq_2}^i} \right) J_{kq_1}(r_1) J_{kq_2}(r_2)
$$
$$
= \sum_{k=1}^\infty \alpha_{kl} \sum_{q_1, q_2} 2\Re\mathfrak{e} \left(\frac{1}{N} \sum_i a_{kq_1}^i \overline{a_{kq_2}^i} \right) J_{kq_1}(r_1) J_{kq_2}(r_2). \tag{2.15}
$$

This equality is achieved thanks to the identity $\overline{a_{-kq}} = a_{kq}$, which holds because P_i is a real-value function and $J_{-kq}(r) = (-1)^k J_{kq}(r)$.

Following [19], we define

$$\left(C^{FB} \right)_{(k_1, q_1),(k_2, q_2)} = \delta_{k_1, k_2} \frac{1}{N} \sum_i 2\Re\mathfrak{e} \left(a_{k_1 q_1}^i \overline{a_{k_2 q_2}^i} \right). \tag{2.16}$$

C^{FB} is the rotationally invariant covariance matrix built from the expansion coefficients in the Fourier-Bessel basis. The expansion coefficients a_{kq}^i can be computed from P_i using the fast method presented in [19]. With the above definition (Eq. (2.16)):

$$C_l^{(proj)}\left(r_1, r_2\right) = \sum_{k=1}^\infty \alpha_{kl} \sum_{q_1, q_2} \left(C^{FB}\left(k\right) \right)_{q_1, q_2} \cdot J_{kq_1}(r_1) J_{kq_2}(r_2). \tag{2.17}$$

Finally, same as Eq. (2.7), we expand $C_l^{(proj)}\left(r_1, r_2\right)$ using the spherical Bessel series $\left\{ j_{ls}(r) : s = 1, 2, ... \right\}$ and obtain

$$\left(C_l^{(proj)} \right)_{s_1, s_2} = \sum_{k=1}^\infty \alpha_{kl} \sum_{q_1, q_2} \left(C^{FB} \right)_{(k, q_1),(k, q_2)} \cdot \beta_{(k, q_1),(l, s_1)} \beta_{(k, q_2),(l, s_2)}, \tag{2.18}$$

where
$$\beta_{(k, q),(l, s)} := \int_0^a J_{kq}(r) \cdot j_{ls}(r) r^2 \, dr. \tag{2.19}$$

$C_l^{(proj)}$ in Eq. (2.19) is the empirical l-th order auto-correlation expanded using spherical Bessel functions. It approximates C_l as defined in Eq. (2.7), using projection images of the volume instead of the volume itself.

Chapter 3

Algorithm and Numerical Experiments

In practice, we follow [19] for the truncation of the Fourier-Bessel expansion in Eq. (2.8). Suppose that the projection $P_i(x, y)$, where $(x, y) \in [-1/2, 1/2]^2$, is sampled on a Cartesian grid of size $2K \times 2K$. For any $1 \leq k \leq K$, let $q(k)$ denote the number of integers q such that R_{kq}, the q-th root of the Bessel function $J_k(r)$, satisfies $R_{kq} \leq \pi K$. Likewise, for the expansion of $C_l(r_1, r_2)$ and $C_l^{(proj)}(r_1, r_2)$ in Eq. (2.6) and Eq. (2.17) respectively, we use $\left\{ j_{ls}(r) \right\}_{s=1}^{s(l)}$ with the choice of $s(l)$ being the number of integers s such that $\tilde{R}_{ls}$, the s-th root of the spherical Bessel function $j_l(r)$, satisfies $\tilde{R}_{ls} \leq \pi K$. We consider l in the range $1 \leq l \leq L$, where L is a positive integer. Here, we set $L = 10$, which is sufficient for making classification inferences based on Table 2.1.

After applying the truncation to Eq. (2.16), we define, for any $1 \leq k \leq K$ and $1 \leq q_1, q_2 \leq q(k)$

$$\left(C^{FB}(k) \right)_{q_1, q_2} = \frac{1}{N} \sum_i 2 \mathfrak{Re} \left(a_{kq_1}^i \overline{a_{kq_2}^i} \right). \tag{3.1}$$

$C^{FB}(k)$ is the k-th diagonal block of C^{FB} (Eq. (2.16)) which is a sample covariance matrix of size $q(k)$-by-$q(k)$ [19]. Using this, Eq. (2.18) can be rewritten as

$$C_l^{(proj)} = \sum_{k=1}^K \alpha_{kl} B_{kl}^T C^{FB}(k) B_{kl}, \tag{3.2}$$

where B_{kl} is a matrix of size $q(k)$-by-$s(l)$ defined by

$$(B_{kl})_{q,s} := \beta_{(k,q),(l,s)}, \ 1 \leq q \leq q(k), \ \ 1 \leq s \leq s(l). \tag{3.3}$$

Then, $C_l^{(proj)}$ is a $s(l)$-by-$s(l)$ matrix. We compute the numbers $\beta_{(k,q),(l,s)}$ and α_{kl} explicitly using Gauss–Legendre quadrature with $2K$ and K points, respectively.

The symmetry detection algorithm is summarized as follows: Given N projections P_i, $i = 1, \ldots, N$, which are sampled on a $2K \times 2K$ Cartesian grid,

1. Choose a positive integer L. For any $1 \leq k \leq K$ and $1 \leq l \leq L$, compute the matrices $B_{k,l}$ as defined in Eq. (3.3) and the values α_{kl} as in Eq. (2.14).

2. Compute the Fourier-Bessel coefficients of the projection images as shown in Eq. (2.10). This can be efficiently done using the fast method proposed in [19].

3. For each $1 \leq k \leq K$, compute the matrix $C^{FB}(k)$ as in Eq. (3.1), which is of size $q(k) \times q(k)$.

4. For $1 \leq l \leq L$, compute the empirical l-th order auto-correlation function $C_l^{(proj)}$ as in Eq. (3.2), which is of size $s(l) \times s(l)$.

5. Compute the eigenvalues of the matrices $C_l^{(proj)}$ and estimate their rank for $l = 1, \ldots, L$ based on the nullity of the eigenvalues. Compare the estimated rank with Table 2.1 to determine the symmetry group.

The algorithm was implemented in Python and tested on both simulated data and projection images of a real volume, corresponding to all symmetry types presented in Table 2.1. To generate the simulated projections, we used Gaussian mixture volumes that exhibit the desired symmetries and are constructed to be compactly supported within the unit ball. For real volumes, we utilized data from `https://rcsb.org`. Each volume V was sampled on a three-dimensional grid of size $300 \times 300 \times 300$ pixels. We then computed the spherical Fourier-transformed coefficients A_{lms} of the volume V for $1 \leq l \leq K$, $1 \leq s \leq s(l)$ using the SPR toolbox ASPIRE.

The values of A_{lms} were used to compute the auto-correlation function of order l according to Eq. (2.7), resulting in an $s(l) \times s(l)$ matrix denoted as $C_l^{(vol)}$. We treated $C_l^{(vol)}$ for $l = 1, \ldots, 10$ as the ground truth and compared it with the matrices $C_l^{(proj)}$, which were computed from the projections. From these volumes, we generated $N = 1000$ projections of size 300×300. We then compared the matrices $C_l^{(vol)}$ and $C_l^{(proj)}$ and analyzed the spectrum of each. The implementation code is available at `https://github.com/MaayanYesharim/SymmetryDetection`.

3.1 Simulations and Experiments

In this section, we demonstrate our algorithm on simulated and experimental data. We present three examples using three different volumes. For each volume, we include the corresponding projection images, correlations of order l, and spectrum analysis. The complete experiments are available in the project's GitHub repository. Additionally, we provide the *TestComplement* suite, which contains copies of these experiments for volumes from all symmetry types listed in Table 2.1.

3.1.1 Simulated Volume

Example 1: Asymmetric volume

For the asymmetric volume, we first visualize the projection images in Figure 3.1. Here, we see four projection images, $P_1, \ldots, P_4$, taken from the asymmetric simulated volume, constructed of random three-dimensional Gaussian blobs with compact support in the unit sphere. These projection images serve as the input for the algorithm.

Figure 3.1: Projection images from four different orientations for an asymmetric synthetic volume.
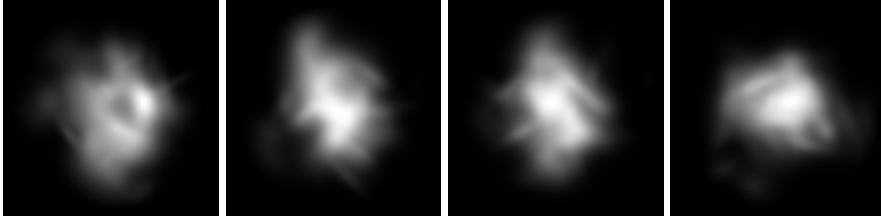

Figure 3.2 visualizes the matrices $C_l^{(vol)}$ and $C_l^{(proj)}$ for $l = 1, 2, 3$, computed from $N = 1000$ projection images, along with their *log-relative difference*, defined as

$$\log \left| \frac{C_l^{(vol)} - C_l^{(proj)}}{C_l^{(vol)}} \right|.$$

The first column displays $C_l^{(vol)}$, computed directly from the volume, serving as the ground truth. The second column shows $C_l^{(proj)}$, the approximation derived from the projection images. The third column depicts the log-relative difference. We expect the first two columns to be similar, with a small difference between them. The log-relative difference is plotted to provide a detailed view of the entry-wise error.

From Figure 3.2, we observe that $C_l^{(proj)}$ closely resembles $C_l^{(vol)}$, confirming that the calculations were performed correctly and that $C_l^{(proj)}$ approximates $C_l^{(vol)}$. Additionally, the log-relative difference shows a majority of the area with uniformly low values, reinforcing the strong match between $C_l^{(vol)}$ and $C_l^{(proj)}$. However, we also observe light and dark regions indicating significant differences, often corresponding to contour lines in the original matrices, where notable changes in correlation matrix values occur. We suspect that these discrepancies result from the smoothing effect of averaging from calculating C^{FB} (Eq. (2.16)) or the multiple numerical steps involved in computing $C_l^{(proj)}$, including the calculation of B_{lk} (Eq. (3.3)), and averaging in Eq. (3.2).

These observations lead us to conclude that while the overall approximation was accurate, there are identifiable regions where the approximation deviates from the expected results. These deviations suggest localized inaccuracies in the approximation process, which might affect the precision of subsequent analyses. With this in mind, we proceed to compare and analyze the differences in the spectra.

Figure 3.2: From left to right by column: $C_l^{(vol)}$, $C_l^{(proj)}$, the log-relative difference of these two for $l = 1, 2, 3$.
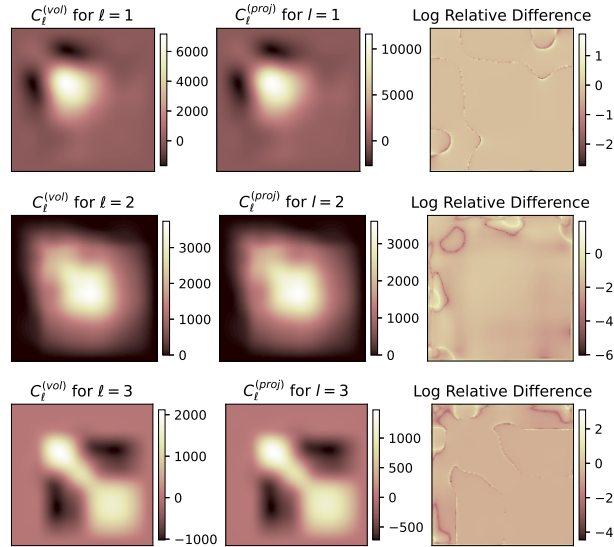
The eigenvalues of $C_l^{(vol)}$ and $C_l^{(proj)}$ for $l = 1, 2, 3$ are shown in Fig. 3.3. For $l = 1$, $C_l^{(vol)}$ has three non-zero eigenvalues, as predicted by theory (Table 2.1). Similarly, the eigenvalues for $l = 2$ and $l = 3$ are as expected. These results confirm the accuracy of $C_l^{(vol)}$, our ground truth, and indicate the expected ranks of matrices derived from an asymmetric volume. The accuracy of $C_l^{(vol)}$ provides high confidence that the method can be effective, assuming a good approximation

Figure 3.3: Bar plot of eigenvalues of $C_\ell^{(vol)}$ and $C_\ell^{(proj)}$ for a simulated asymmetric volume
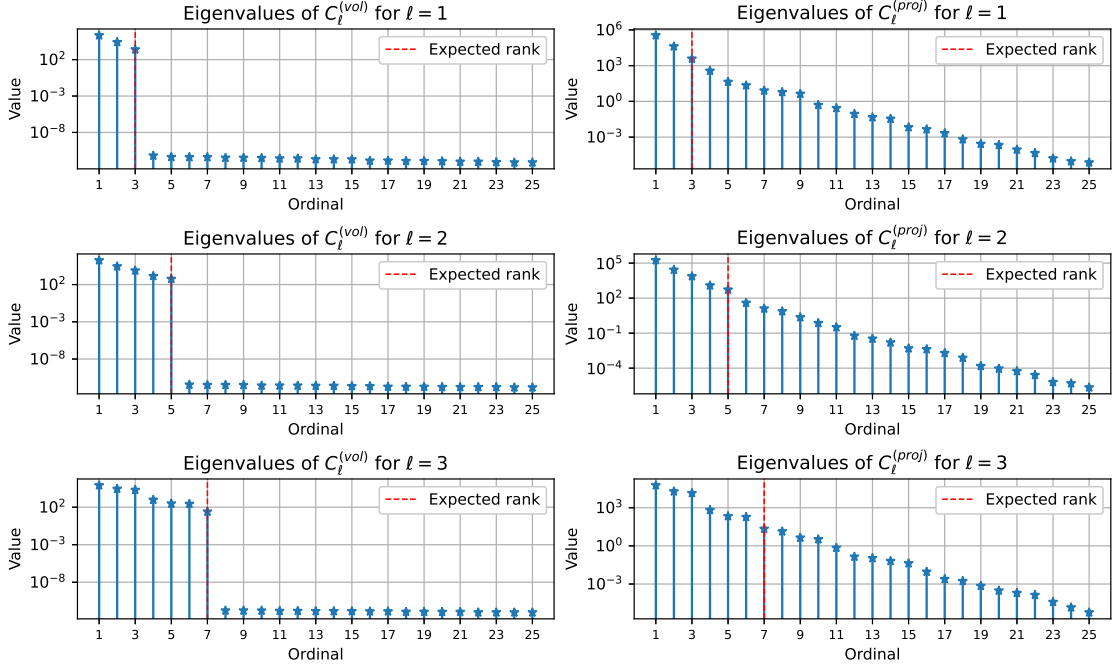
is achieved. We aim to observe similar rank indications in the spectrum of $C_l^{(proj)}$ as well. However, this example does not show a significant difference between the magnitudes of the first set of eigenvalues (expected to be non-zero) and those expected to be zero.

In summary, this example demonstrates the computation of $C_l^{(vol)}$ and $C_l^{(proj)}$ using projection images from an asymmetric volume, establishing $C_l^{(vol)}$ as the ground truth. The visualizations and eigenvalue comparisons indicate that while the approximation process captures the general structure, there are significant deviations that highlight areas needing refinement. These findings underscore the importance of further investigation into the sources of these discrepancies, particularly in how the projection images and numerical steps impact the results.

We begin with this example because it effectively demonstrates the ground truth, providing a clear baseline for comparison. Despite its limitations, this example is crucial for highlighting where discrepancies arise and guiding subsequent improvements in our methodology.
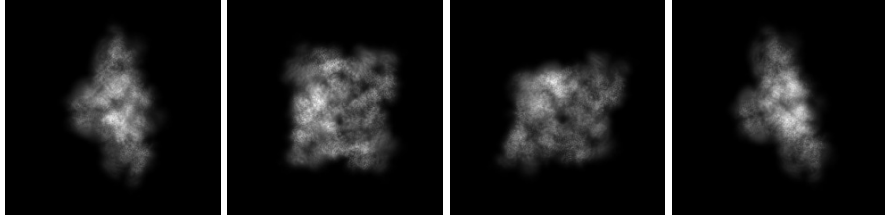
3.1.2   Real Volume

In practice, we know the molecular structures have a semi-symmetrical structure but do not exhibit perfect symmetry. We use the informal term semi-symmetrical

to describe structures that have an overall appearance of symmetry but contain subtle differences or imperfections that break that symmetry. For example, identical subunits may adopt different conformations (quasi-symmetry), different subunits may form approximately symmetrical complexes (pseudo-symmetry), or oligomers with different symmetries may interact along their respective symmetry axes (symmetry mismatch). [14]

Example 2: Cyclic Volume of Order 4**,** C_4

In this example, we explore the rank of correlations of order l to assess their alignment with theoretical expectations for C_4. We also examine how the process performs with images generated from a real volume. Here, we use the structure of RyR1 [12]. As in the previous example, Figures 3.4 ,3.5 and 3.6 present the projection images, the correlations $C_l^{(vol)}$ and $C_l^{(proj)}$ for $l = 1, 2, 3$, and the spectrum analysis, respectively.

Figure 3.4: Projection images from four different orientations for a C_4 volume.

To highlight the characteristics of images from a real volume, we compare Figures 3.1 and 3.4. Four projection images are exhibited in Figure 3.4, which display sharper features, enhanced edge definition, and finer structural details compared to synthetic data. These differences are evident to the eye and demonstrate the distinct qualities of real versus synthetic data.

The spectrum analysis in Figure 3.6 shows slightly different results compared to Example 1. We observe that the magnitude of the eigenvalues of the ground truth, $C_l^{(vol)}$, exhibits three distinct groups: one set with the expected rank, a residual group that aligns with the ranks of an asymmetric volume, and one group that is null. We hypothesize that the "residual" is a result of the semi-symmetrical structure of the volume. Given the nature of real structures and our goal to identify the closest symmetry, our approach focuses on detecting the first significant drop in eigenvalue magnitude and determining semi-symmetries accordingly. For $C_{l=1,2,3}^{(vol)}$, we observe a drop of at least 10^4 after the first eigenvalue, indicating that these correlations correspond to rank 1.

For $C_l^{(proj)}$ in the right column of Figure 3.6, we observe that the expected rank is noticeable, as indicated by a sufficient difference between the first eigenvalue and the subsequent ones. However, this drop is not as pronounced as in the ground truth case; the first eigenvalue shows only a gap of 10^1 from the following eigenvalues. Using these results, we can now infer the symmetry. Since Figure 3.6 implies that $\text{rank}(C_{l=1}^{proj}) = 1$, $\text{rank}(C_{l=2}^{proj}) = 1$, and $\text{rank}(C_{l=3}^{proj}) = 1$, we can detect that these image projections originate from a volume with C_4 symmetry.

On the one hand, this result demonstrates that the experiment for $C_l^{(proj)}$ partially succeeds in reflecting the rank we aim to detect. On the other hand, the inference drawn from this approximation is not as strong or accurate as we would expect when compared to $C_l^{(vol)}$. To improve this, future work should focus on refining the approximation to make $C_l^{(proj)}$ closer to $C_l^{(vol)}$ in all possible parameters, such as normalization and the magnitude of eigenvalues.

Example 3: Tetrahedral Volume, T_{12}

Here, we repeat the above experiments and plots for a real volume with T_{12} symmetry. We use the structure of 6KPA [1]. The results are shown in Figures 3.7, 3.8, and 3.9. Using the spectrum analysis of $C_l^{(proj)}$ for $l = 3, 4, 6$ in Figure 3.9, and the encoding given in Table 2.1, we aim to infer that the volume has T_{12} symmetry.

Indeed, the rank of $C_{l=3}^{(proj)}$ and $C_{l=4}^{(proj)}$ is relatively noticeable, as the first eigenvalue has a gap of at least 10^2 from the others. Focusing on the spectral analysis of $C_{l=6}^{(proj)}$, we can still observe a gap between the first two eigenvalues and the next one, but now it is only by 10^1. This noticeable discrepancy highlights the challenges in achieving a pure solution for eigenvalue analysis. The observed gaps in eigenvalues suggest that definitive certainty in symmetry detection is difficult to achieve and emphasize the need for a carefully defined threshold for identifying non-zero eigenvalues.

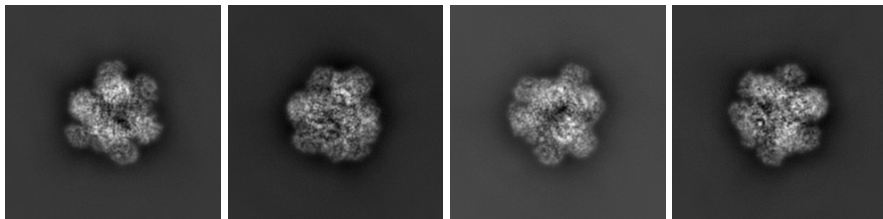Figure 3.7: Projection images from four different orientations for a T_{12} volume.

Figure 3.5: From left to right by column: $C_l^{(vol)}$, $C_l^{(proj)}$, the log-relative difference of these two for $l = 1, 2, 3$.
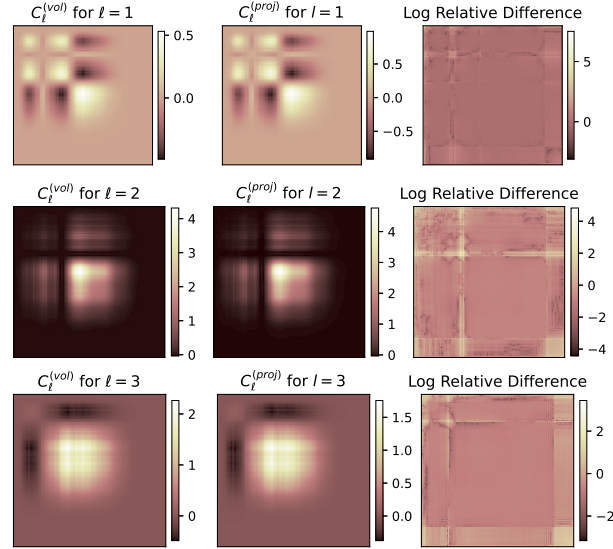

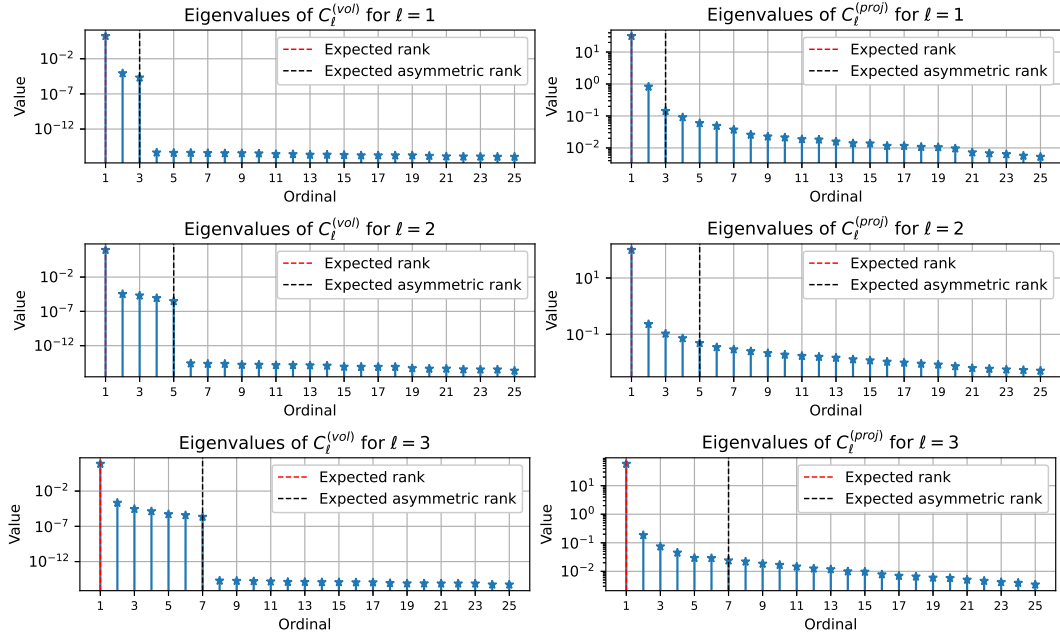Figure 3.6: Bar plot of eigenvalues of $C_\ell^{(vol)}$ and $C_\ell^{(proj)}$ for real C4 volume

Figure 3.8: From left to right by column: $C_l^{(vol)}$, $C_l^{(proj)}$, the log-relative difference of these two for $l = 3, 4, 6$.
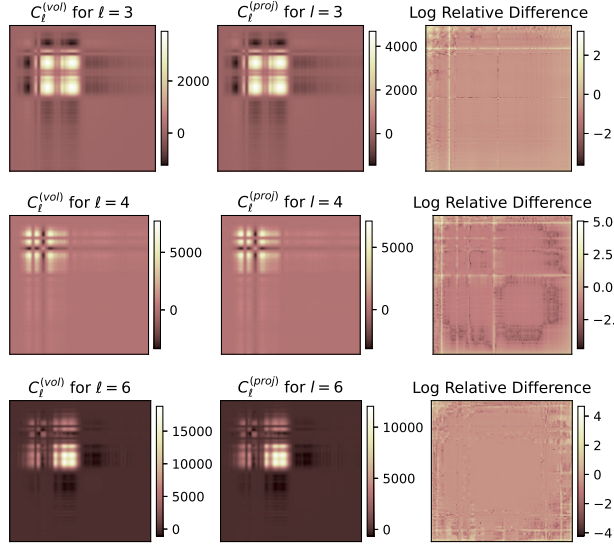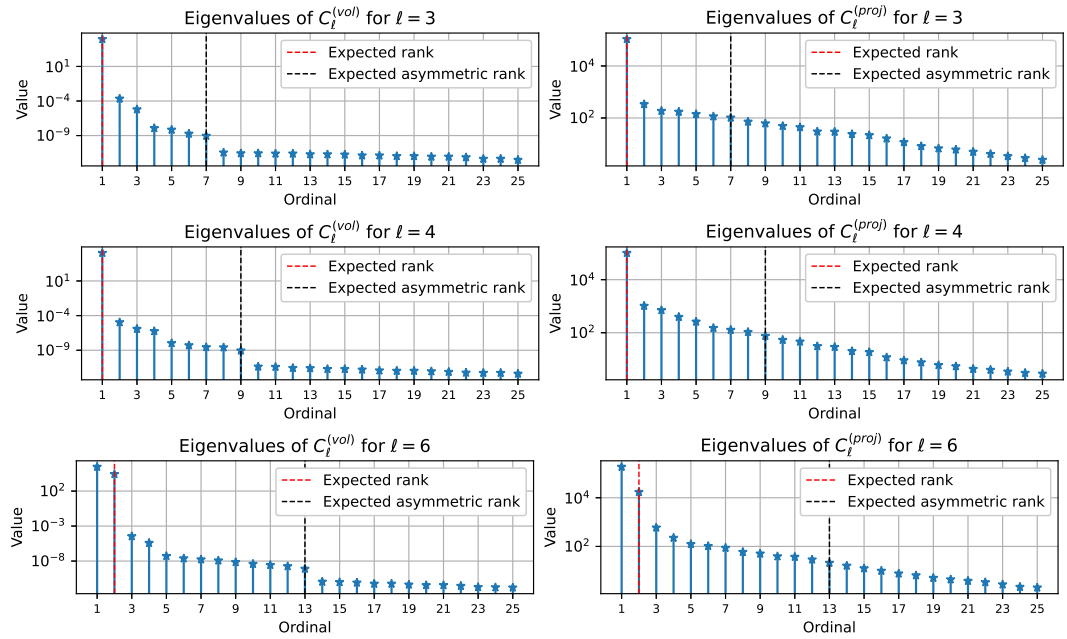

Figure 3.9: Bar plot of eigenvalues of $C_\ell^{(vol)}$ and $C_\ell^{(proj)}$ for real T_{12} volume

Chapter 4

Conclusions and Future Work

In this thesis, we introduced a symmetry detection algorithm based on the spectral analysis of spatial correlations. This concept, first introduced in [10] and later endorsed by [2], forms the foundation for our work on symmetry recovery. With a solid analytical foundation, our approach provides a well-defined starting point for a practical numerical solution to the symmetry detection problem for cryo-EM data. This method offers a significant improvement over the limitations of approaches based on the analysis of 2D projection images.

In addition to the development of the algorithm, this work includes a theoretical formulation that frames the symmetry detection problem and provides a mathematical solution. This formulation connects projection images with the underlying symmetries of molecular structures, facilitating a systematic exploration of these structures. By defining the problem in terms of rank determination, this framework offers a robust basis for further theoretical advancements and practical algorithmic improvements.

We validated the proposed algorithm through a series of experiments with both simulated and real volumes. These experiments demonstrated the algorithm's ability to approximate l-th order correlations from projection images and showed promising agreement between the correlations and their spectra when directly computed from volume data. This agreement reinforces the theoretical foundations of the algorithm. We also explored how to approximate these correlations from images and analyze their spectra, with improved performance observed in real-volume cases. Additionally, we addressed the challenges of near-symmetries or *semi-symmetrical* structures in real data and provided a strategy for handling them. However, some discrepancies were noted, particularly in approximating the eigenvalues of correlations from projection images. These discrepancies underscore the need for refinement, such as improving the robustness of spectral gap determination and reducing numerical errors in matrix computations.

Overall, our work contributes to advancing symmetry detection methods in

cryo-EM, laying the groundwork for more accurate and efficient algorithms that can be applied to a broader range of molecular structures. By addressing the current limitations and exploring future improvements, we aim to further enhance the robustness and applicability of our approach in practical settings.

Appendix

Appendix A

Symmetry Groups of SO(3)

In applied mathematics, particularly in the study of symmetry and group theory, the group SO(3) plays a fundamental role. SO(3), the special orthogonal group in three dimensions, consists of all 3×3 orthogonal matrices with determinant 1. These matrices represent rotations in three-dimensional space, preserving the origin and maintaining the coordinate system's orientation.

Formally, SO(3) is defined as

$$\text{SO}(3) = \{R \in \mathbb{R}^{3\times 3} \mid R^T R = I \text{ and } \det(R) = 1\},$$

where R^T denotes the transpose of R, and I is the identity matrix.

In group theory, the symmetry group of a geometric object is the group of all transformations under which the object is invariant, endowed with the group operation of composition. Mathematically, with our example SO(3), a subgroup G of SO(3) is a symmetry group of a three-dimensional object X if $gX = X$ for any $g \in G$.

Subgroups of SO(3) represent different types of symmetries. According to a classical theorem of Klein [11], all the discrete symmetry groups G, when not being the trivial group I_3, are isomorphic to one of the following:

1. **Cyclic groups** C_n: These groups consist of rotations about a single axis by angles that are multiples of $360°/n$. The group C_n has n elements.

2. **Dihedral groups** D_n: The dihedral group includes rotations by $360°/n$ around a central axis and rotations by $180°$ around perpendicular axes, which leave an n-sided polygon unchanged. D_n has $2n$ elements.

3. **Tetrahedral group** T_{12}: This group is the symmetry group of a regular tetrahedron. It has 12 elements.

4. **Octahedral group** O_{24}: The symmetry group of a cube or an octahedron. O_{24} has 24 elements.

5. **Icosahedral group** I_{60}: This group is the symmetry group of an icosahedron or a dodecahedron. It has 60 elements.

These subgroups provide a comprehensive understanding of the rotational symmetries within SO(3). By studying these groups, one can analyze the structure and properties of symmetries in various physical and mathematical contexts, ranging from Cryo-EM and crystallography to the solutions of differential equations.

Bibliography

[1] C.Y. Chen, Y.C. Chang, B.L. Lin, C.H. Huang, and M.D. Tsai. Structure of 6KPA, a T_{12} symmetry real structure. `https://www.rcsb.org/structure/6KPA`.

[2] Xiuyuan Cheng. *Random Matrices in High-dimensional Data Analysis*. diplomathesis, Princeton University, 2013.

[3] Edward H Egelman. The current revolution in cryo-em. *Biophysical journal*, 110(5):1008–1012, 2016.

[4] Joachim Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford university press, 2006.

[5] Shamgar Gurevich, Ronny Hadani, and Amit Singer. Representation theoretic patterns in three dimensional cryo-electron microscopy iii-presence of point symmetries, 2009.

[6] Ronny Hadani and Amit Singer. Representation theoretic patterns in three dimensional cryo-electron microscopy i: The intrinsic reconstitution algorithm. *Annals of mathematics*, 174(2):1219, 2011.

[7] Ronny Hadani and Amit Singer. Representation theoretic patterns in three-dimensional cryo-electron microscopy ii—the class averaging problem. *Foundations of computational mathematics*, 11:589–616, 2011.

[8] Shaoda He and Sjors HW Scheres. Helical reconstruction in relion. *Journal of structural biology*, 198(3):163–176, 2017.

[9] Zvi Kam. Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules*, 10(5):927–934, 1977.

[10] Zvi Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, 82(1):15–39, 1980.

[11] F. Klein and G.G. Morrice. *Lectures on the Icosahedron and the Solution of the Fifth Degree.* Cosimo classics / science. Cosimo Classics, 2007.

[12] R. Ma, O. Haji-Ghassemi, D. Ma, L. Lin, A. Samurkas, F. Van Petegem, and Z. Yuchi. Structure of RyR1, a C_4 symmetry real structure. `https://www.rcsb.org/structure/6M2W`.

[13] Frank Natterer. *The mathematics of computerized tomography.* SIAM, 2001.

[14] Cyril F Reboul, Simon Kiesewetter, Dominika Elmlund, and Hans Elmlund. Point-group symmetry detection in three-dimensional charge density of biomolecules. *Bioinformatics*, 36(7):2237–2243, 2020.

[15] Yoel Shkolnisky and Amit Singer. Viewing direction estimation in cryo-EM using synchronization. *SIAM journal on imaging sciences*, 5(3):1088–1110, 2012.

[16] Amit Singer and Yoel Shkolnisky. Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming. *SIAM journal on imaging sciences*, 4(2):543–572, 2011.

[17] Amit Singer, Zhizhen Zhao, Yoel Shkolnisky, and Ronny Hadani. Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM Journal on Imaging Sciences*, 4(2):723–759, 2011.

[18] Qing Wang, Olaf Ronneberger, and Hans Burkhardt. Fourier analysis in polar and spherical coordinates. Technical Report Internal Report 1/08, IIF-LMB, Computer Science Department, University of Freiburg, 2008.

[19] Zhizhen Zhao, Yoel Shkolnisky, and Amit Singer. Fast steerable principal component analysis. *IEEE transactions on computational imaging*, 2(1):1–12, 2016.

[20] Zhizhen Zhao and Amit Singer. Fourier–bessel rotational invariant eigenimages. *JOSA A*, 30(5):871–877, 2013.

[21] Jasenko Zivanov, Takanori Nakane, Björn O Forsberg, Dari Kimanius, Wim JH Hagen, Erik Lindahl, and Sjors HW Scheres. New tools for automated high-resolution cryo-EM structure determination in relion-3. *elife*, 7:e42166, 2018.

תקציר

שיטת המיקרוסקופ אלקטרונים הקריוגני (Croy-EM) מאפשרת קבלת מבנים תלת־ממדיים ברזולוציה גבוהה של מולקולות ביולוגיות ללא צורך בגיבוש. אחד האתגרים המרכזיים בתהליך השחזור הוא קביעת המבנה התלת־ממדי מתוך הטלות דו-ממדיות רועשות שנאספו מזוויות צפייה לא ידועות. בשנים האחרונות, התגלתה חשיבות רבה לזיהוי סימטריה מולקולרית כחלק מתהליך השחזור, שכן לסימטריה יש השפעה על המבנה שמתקבל מתוך ההקרנות.

העבודה מציעה גישה לזיהוי סימטריה ישירות מנתוני המיקרוסקופ, ללא הסתמכות על מידע מוקדם מהמשתמש. הגישה מבוססת על חישוב פונקציות אוטו-קורלציה של נפח המולקולה וניתוח הדרגה שלה לצורך קביעת קבוצת הסימטריה של המולקולה.

אוניברסיטת תל אביב
הפקולטה למדעים מדוייקים
ע"ש ריימונד ובברלי סאקלר
בית הספר למדעי המתימטיקה

בעית מציאת סימטריה במיקרוסקופ אלקטרונים קריוגני

חיבור זה הוגש כחלק מהדרישות לקבלת תואר

מוסמך במדעים (**M.Sc**.) באוניברסיטת תל אביב

על ידי

מעיין ישרים

העבודה הוכנה בהנחייתו של

פרופ' ניר שרון ופרופ' יואל שקולניצקי

תשרי, תשפ"ה