AB INITIO TOMOGRAPHY WITH OBJECT HETEROGENEITY AND UNKNOWN VIEWING PARAMETERS

Arunabh Ghosh[1]† *Ritwick Chaudhry*[2]† *Ajit Rajwade*[3]*

[1] Dept. of EE, IIT Bombay
[2] Adobe Research
[3] Dept. of CSE, IIT Bombay
arunabhghosh@iitb.ac.in; rchaudhr@adobe.com; ajitvr@cse.iitb.ac.in

ABSTRACT

In this paper, we present an algorithm to automatically construct all the conformations of a heterogeneous planar object from their tomographic projections at random unknown view angles. Our statistically motivated approach can reveal and analyze the heterogeneity in the projection dataset and segregate the projections belonging to different structures without requiring prior structural information or templates, expert human intervention or even the knowledge of the number of conformations present in the sample. Even in the presence of high noise variance (low SNR) and a large number of conformations, our algorithm can estimate the structures of each conformation to a high degree of accuracy. We demonstrate the broad applicability of our algorithm by evaluating its performance on synthetic 2D datasets of well-known protein complexes such as Lipase under varying levels of noise and different number of conformations.

Index Terms— Cryo-electron microscopy, Heterogeneity, Tomography, Ab initio reconstruction

1. INTRODUCTION

Over the years, single particle cryo-electron-microscopy [1] (referred to hereafter as Cryo-EM) has seen the development of many sophisticated algorithms to analyze the structures of macromolecular complexes up to a resolution of 7-10Å [2, 3, 4, 1]. However, biological assemblies are dynamic machines that adopt a wide range of structures (called 'conformations') necessary for carrying out some of their vital functions. For example, a sample of a certain ribosome may have different sub-units as they have to synthesize various polypeptide chains [5], or a virus sample may consist of virions in different maturation stages [6]. Such conformational variability poses a challenge to standard Cryo-EM reconstruction algorithms which often assume that projections belong to identical structures. Not segregating the projections belonging to different conformations will limit the level of detail revealed in the estimated structures, as the information from the different conformations will be averaged out in the final reconstruction.

Prior work: Much of the prior research in this area belongs to one of the following categories: (1) The most general strategy used for separating heterogeneous projections is based on supervised classification [7, 8], which uses known reference structures to isolate the projections. The dependency on *a priori* knowledge severely restricts the widespread use of this approach. (2) Other methods use alternating schemes, in which orientation and class assignment are simultaneously refined. However, as this is a highly non-convex optimization problem, the effectiveness of this class of methods is generally difficult to predict. (3) The few algorithms that do not require a prior template require the user to provide the number of conformations K [9, 10, 11]. An incorrect input is detrimental, as essential structures can be overlooked and reconstructed structures may compose of an average of two or more conformations [12]. In [9] particularly, results are only shown for the case of no noise and an equal number of projections from each conformation.

Contributions: To address the heterogeneity issue we have developed an algorithm that makes it possible to discover all the conformations present in a sample without expert intervention, prior templates, or an input of the number of discrete sample conformations. Through a robust statistical analysis, we separate the noisy projections belonging to different conformations and then independently reconstruct all the different structures using a noise-resistant procedure. In this paper, we will describe our approach and present results on reconstructing 2D images (and hence simulated 1D parallel-beam projections), even though actual objects are 3D. This follows previous work in the image processing community which has studied the *2D variant* of this problem extensively [13, 14, 15]. Nonetheless, the underlying principles remain the same, and the computational problem remains very challenging even for reconstructing heterogeneous 2D images.

Algorithm Overview: In Sec. 2, we describe a pre-processing procedure to obtain a representative set of denoised projections, from noisy ones. We then present a method to cluster these projections according to their respective conformations in Sec. 3. After the projections are separated, we proceed to independently reconstruct each of the conformations using the approach described in Sec. 4. The performance of the algorithm is demonstrated in Sec. 5. Finally, we conclude in Sec. 6 with a discussion of future work.

2. PRE-PROCESSING OF THE PROJECTION DATASET

2.1. Robust class-based clustering

Typically in the case of Cryo-EM, we seek to obtain a representative set of less noisy projections by clustering them into a small number of classes, based on orientation and structural similarity [16, 17, 18]. In the case of multi-particle reconstruction, however, we have observed that standard algorithms like K-means generate clusters containing projections belonging to different conformations, making them ineffective in reconstructing either of the conformations. To generate 'pure' clusters, we turn to other clustering algorithms,

*AR thanks IIT-B Seed Grant 14IRCCSG012

† These authors contributed equally to this work.

such as the Single-linkage clustering algorithm [19]. This is a variant of hierarchical clustering [20] which treats each data point as a singleton cluster, and then successively merges clusters until a user-set condition is satisfied. Here, we merge in each step, the two clusters with the smallest minimum pairwise Euclidean distance d_{min} until $d_{min} < \epsilon$. The threshold ϵ determines our estimate of the distance between projections $R_{\theta_1} f$ and $R_{\theta_2} f$ of the same conformation of image $f(x, y)$ at angles θ_1, θ_2 respectively. We choose ϵ to be a small value such that only clusters with projections belonging to the same conformation are merged at each step. However, ϵ should also be large enough, such that there are a sufficient number of projections assigned to each cluster to perform the averaging step effectively. In our experiments, we have found that a small value of ϵ (empirically determined as 1.15, see Sec. 5) works for a broad range of biological complexes, that is, in each case, the clusters produced have a sufficient number of projections belonging predominantly to one conformation. Henceforth, we will denote the number of clusters generated as K_c. Within each pure cluster, we define the processed representative projection $\tilde{p}_j$ (for cluster index j), where $1 \leq j \leq K_c$, to be the average of all the projections assigned to that cluster. This averaging induces a basic form of filtering to remove the noise.

2.2. Patch-Based Denoising

The processed cluster centers $\{\tilde{p}_j\}_{j=1}^{K_c}$ as obtained in the previous step are significantly less noisy. The residual noise is removed by passing the cluster centers ($\{\tilde{p}_j\}_{j=1}^{K_c}$) through a patch-based PCA denoising algorithm adapted from [21] (see supplementary material for more details). Hereafter, we use the symbol $\tilde{q}_i$ to refer to the denoised version of the cluster center $\tilde{p}_i$.

3. CLASSIFICATION

We consider that heterogeneity is intrinsically a clustering/classification problem. Here we present a feature-based method to help segregate the projections belonging to distinct conformations.

3.1. Preliminary Classification using a moment-based approach

In a majority of cases, heterogeneity is the result of the addition/removal of certain subunits. For example, in the case of proteins, heterogeneity is often caused by addition of carbohydrates or lipids into the protein polypeptide chain [22]. This implies a change in the average electron density of the object. In other words, different conformations will likely have a different zero frequency/average value. As it turns out, using the Helgason Ludwig Consistency Conditions (HLCC) [23], we transform this empirical observation into an equivalent metric for classifying projections belonging to different conformations. The HLCC [23] give the following relationship between the geometric moments of the underlying image $f(x, y)$ and those of its projections at any angle: $m_\theta^{(n)} = \sum_{j=0}^{n} \binom{n}{j} (\cos\theta)^{n-j} (\sin\theta)^j v_{n-j,j}$, where $v_{p,q} \triangleq \int_{\infty}^{-\infty} \int_{\infty}^{-\infty} f(x, y)\ x^p y^q\ dx\ dy$ denotes the order (p, q) image moment and $m_\theta^{(n)} \triangleq \int_{\infty}^{-\infty} g(\rho, \theta) \rho^n\ d\rho$ denotes the order n projection moment. When $n = 0$, we obtain $m_\theta^{(0)} = v_{0,0}$, i.e. the zeroth-order moment of a projection (LHS) is equal to the sum-total of the (density) values of the underlying image f (RHS). This implies that all projections from the same conformation have the same zeroth-order moment and its value differs across conformations having a different average value. Therefore, it can be used as a statistic for classification as demonstrated by Fig. 1. In the practical scenario, the classification provided by the zeroth-order moment, although good, needs to be refined by a robust scheme which takes into account the characteristics of the entire set of projections. For this, we introduce a graph Laplacian-based algorithm, the details of which are described next.

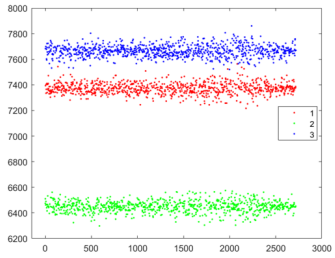


Fig. 1: Scatter plot of zeroth moment of the projections corresponding to three different conformations (shown in Fig. 4).

3.2. Graph Laplacian-based clustering

Like in any many areas of information retrieval, we observe here that although the sampled projections $\tilde{q}_i$ are points in a high dimensional space, they are intrinsically restricted to a manifold/curve in a low-dimensional space (3D) as shown in [14]. This curve may have a complicated non-linear structure that may not be captured by linear methods such as PCA. Therefore, we employ a graph Laplacian-based non-linear dimensionality reduction algorithm [24] to compute a 3D representation that optimally preserves local neighborhood information. By trying to preserve local information, we keep projections belonging to the same conformation close even in the low-dimensional representation and thus implicitly emphasize the conformational information in the dataset. The locality-preserving character of this algorithm also makes it relatively insensitive to outliers and noise. In Fig. 2, we show two examples demonstrating how this approach highlights the conformations present in the sample. Along with Fig. 1 it can be used to identify the number of conformations through a simple visualization (hierarchical clustering can also be used to suggest the number of clusters).

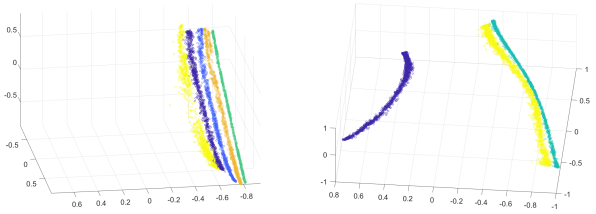


Fig. 2: Low-dimensional representation of the data set of projections (30% additive white noise): 5 different conformations of Lipase from Fig. 4 (left); 3 different conformations of holo-glyceraldehyde-3-phosphate dehydrogenase (holo-GAPDHase) from Fig. 4 (right). A different color used for points from different conformations.

3.3. Nearest-neighbor Based Clustering

As the graph Laplacian-based algorithm preserves local neighborhood information, the projections belonging to the same conforma-

tion lie on a smooth manifold in the 3D space. An important implication of this is that the nearest neighbors of a projection in the low-dimensional representation will in all likelihood belong to the same conformation. The original projection space does not ensure this, due to high amounts of noise and other non-relevant features. This immediately suggests an intuitive way of refining the initial separation (of different conformations) provided to us by the zeroth-order moment: apply a Nearest-neighbor based scheme in the low-dimensional space, i.e. examples are separated into different clusters based on the majority cluster label of their k nearest neighbors. Using this scheme, we correct the few incorrectly labeled cluster centroids. As mentioned above, because the graph-Laplacian algorithm is relatively insensitive to outliers and noise, combined with the k-NN scheme, this is a robust clustering procedure. In Fig. 3 we demonstrate the classification refinement performed by (k-NN) classification scheme. After the classification is refined and the cluster

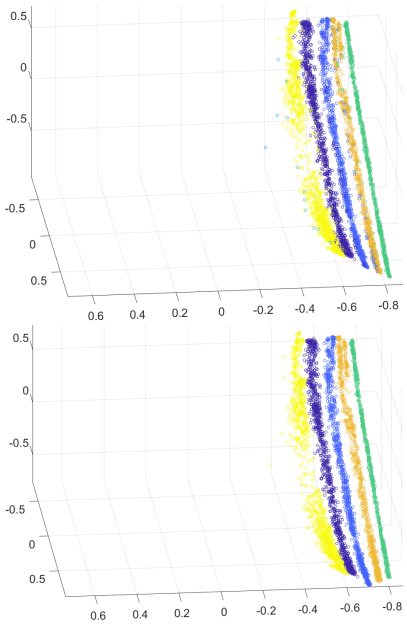


Fig. 3: Top: Classification provided by the moment-based estimation with a few incorrectly classified cluster centroids. Bottom: The few incorrectly classified centroids are corrected using the k-NN method.

centroids are segregated according to their respective conformations, all the projections belonging to the cluster are also assigned the same conformation/class. Since the clusters were highly pure, as ensured by the single-linkage clustering algorithm, we end up with a highly accurate classification of individual projections. Our problem has thus reduced to the case of *independently* reconstructing all of the *individual* conformations from their projections using a robust, noise-resistant single particle reconstruction algorithm, described next.

4. SINGLE-PARTICLE RECONSTRUCTION ALGORITHM

We now follow a three-step algorithm for reconstruction from projections of identical structures, albeit under unknown viewing parameters. First, we cluster similar projections together and denoise the projection centroids. We then use a moments based approach to obtain an initial estimate for the orientations and finally optimize for the structure of the unknown object along with a refinement of the orientations.

4.1. Robust Clustering and Denoising

Clustering: We use the K-means algorithm to cluster the large number of projections into a much smaller number of clusters for the same reasons as described in Sec. 2.1. Let K_s denote the number of clusters thus generated. *Averaging:* Within each cluster, we compute the average of all the projections assigned to that cluster. *Denoising:* The processed projections are then denoised using patch-based PCA denoising algorithm described in Sec. 2.2. The final representative projections are denoted as $\tilde{q}_{i,s}, 1 \leq i \leq K_s$.

4.2. Initialization of the orientations using Helgason Ludwig Consistency Conditions (HLCC)

We harness the information available in the image moments and projection moments (via the HLCC from Sec. 3.1), to estimate the orientations of the projections: $m_\theta^{(n)} = \sum_{j=0}^{n} \binom{n}{j} (\cos\theta)^{n-j} (\sin\theta)^j v_{n-j,j}$. For each order n, we can write the constraints in matrix form, $\mathbf{m}^{(n)} = \mathbf{A}^{(n)} \mathbf{v}^{(n)}$. Here, for a total of K_s projections and for the n^{th} order equation, $\mathbf{A}^{(n)}$ is the $K_s \times (n+1)$ matrix defined by $A_{ij}^{(n)} \triangleq \binom{n}{j} (\cos\theta_i)^{n-j} (\sin\theta_i)^j$, and $v^{(n)} \triangleq \{v_{p,q} | (p+q) = n, p, q \in \mathbb{Z}_{\geq 0}\}$. Since, in practice, the projections are noisy, the above equations will not be satisfied exactly. Instead, we define an energy function as follows

$$E(\{\theta_i\}, \mathbf{v}) = \sum_{n=0}^{N_{max}} \sum_{i=1}^{K_s} \left(m_{\theta_i}^{(n)} - \sum_{j=0}^{n} A_{i,j}^{(n)} v_{n-j,j} \right)^2. \quad (1)$$

where N_{max} denotes the highest order moment to be considered. In practice, a value of $N_{max} = 7$ suffices for all cases. A greater value significantly increases computational time without leading to a discernible increase in gain. By minimizing this energy function, we derive an initial estimate of the angles using an iterative coordinate descent strategy as implemented in [25].

4.3. Optimization strategy to obtain the structure of the object

Starting from the initial estimate from the previous section, we further refine the orientations along with the object structure by minimizing the following energy function in an alternating fashion:

$$\mathcal{M}(\{\theta_i\}, f) = \sum_{i=1}^{K_s} \|\tilde{q}_{i,s} - \mathcal{R}_{\theta_i}(f)\|_2^2. \quad (2)$$

Given an estimate of $\{\theta_i\}_{i=1}^{K_s}$, we solve for f using filtered back-projection (FBP). Given an estimate of the structure f, the orientation of each projection is estimated by independent 1D brute-force search.

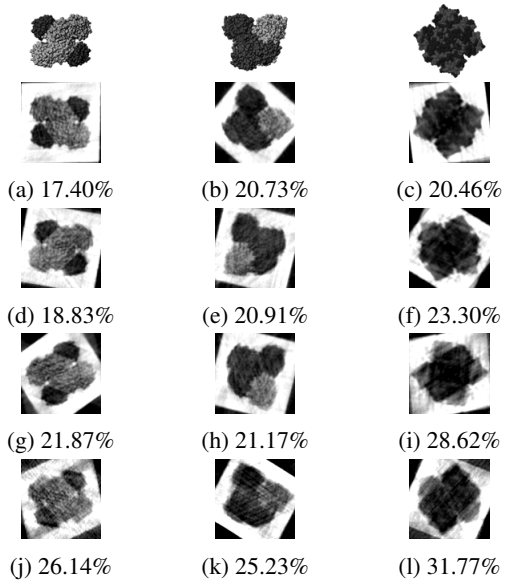
5. RESULTS

In this section, we present a comprehensive set of results demonstrating the ability of our algorithm to perform ab initio classification of projections without the use of any prior knowledge and then proceed to reconstruct each conformation independently. The images used for our experiments were taken from the Database of Macromolecular Movements [26] and had size 100×100. The value of ϵ is chosen to be 1.15 and remains the same across all our experiments. The value is chosen based on empirical observation of the difference between projections of the same conformation at similar angles. We analyze the algorithm with respect to (1) *noise tolerance* and (2) *number of conformations*. The error metric used to

assess the quality of reconstruction is the Relative Mean Squared Error (RMSE) between the registered reconstruction (reconstruction aligned with the test image) and the test image. The RMSE is defined as $\text{RMSE}(f, \hat{f}) = \|f - \hat{f}\|_2 / \|f\|_2$, where $\hat{f}$ is the registered reconstructed estimate for f.

5.1. Noise Tolerance

The effect of additive noise on the reconstruction is depicted in Fig. 4. A total of $Q = 3 \times 10^4$ projections are simulated, where each projection may belong to either of the conformations. All projections were subjected to additive i.i.d. noise from $\mathcal{N}(0, \sigma^2)$. Here we assume $\sigma \triangleq \beta a$ to be known in advance, where $\beta \in [0, 1]$ is a fraction, and a is the average noiseless projection value. As shown, even in the presence of high noise variance, we can reconstruct all the conformations of the object successfully.



(a) 17.40% (b) 20.73% (c) 20.46%

(d) 18.83% (e) 20.91% (f) 23.30%

(g) 21.87% (h) 21.17% (i) 28.62%

(j) 26.14% (k) 25.23% (l) 31.77%

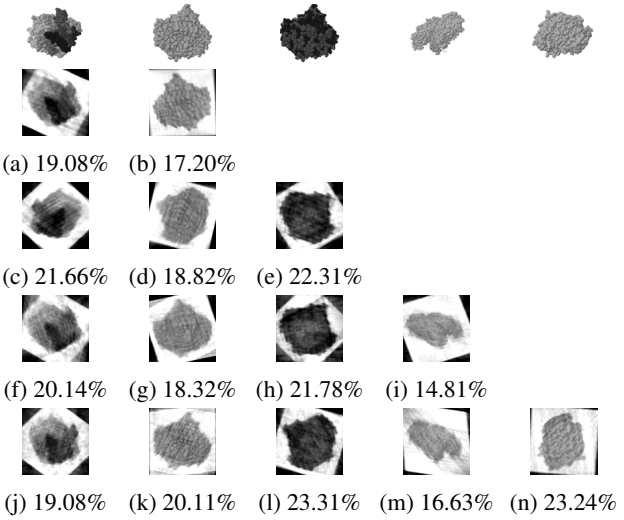
Fig. 4: Top: The three original conformations of holo-GAPDHase, a macromolecular complex. Simultaneous reconstruction results for different σ, along with RMSE: $\sigma = 0.0a$ (second row), $\sigma = 0.1a$ (third row), $\sigma = 0.2a$ (fourth row), and $\sigma = 0.3a$ (fifth row).

5.2. Number of conformations

In this section, we assess the algorithm with respect to its performance as the number of discrete conformations increases. The results are shown in Fig. 5. A total of $Q = K \times 10^4$ projections, where K is the number of conformations, are considered. Each projection is uniformly randomly simulated from either of the conformations. In each case, the projections were subjected to noise from $\mathcal{N}(0, \sigma^2)$ where $\sigma = 0.3a$. Fig. 6 presents the reconstructions in the case of 8 distinct structures composed of a mixture of conformations of Lipase (Fig. 5) and holo-GAPDHase (Fig. 4).

6. DISCUSSION AND CONCLUSION

From the results presented here, we conclude that the algorithm described in this paper successfully tackles one of the most important problems in Cryo-EM - heterogeneity - albeit in 2D. For example,



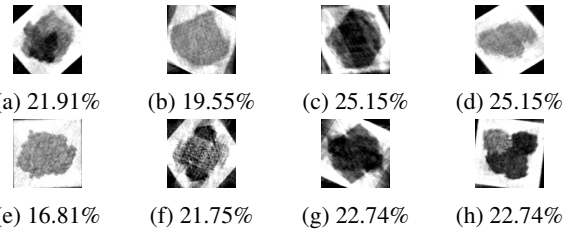
(a) 19.08% (b) 17.20%

(c) 21.66% (d) 18.82% (e) 22.31%

(f) 20.14% (g) 18.32% (h) 21.78% (i) 14.81%

(j) 19.08% (k) 20.11% (l) 23.31% (m) 16.63% (n) 23.24%

Fig. 5: Top row: Five different conformations of Lipase, a protein complex. Bottom: Reconstruction results for different number of conformations. Second row: 2 conformations. Third row: 3 conformations. Fourth row: 4 conformations. Fifth row: 5 conformations



(a) 21.91% (b) 19.55% (c) 25.15% (d) 25.15%

(e) 16.81% (f) 21.75% (g) 22.74% (h) 22.74%

Fig. 6: The reconstruction results of all 8 structures. The original eight structures correspond to the 5 conformations of Lipase (Fig. 5) and 3 conformations of holo-GAPDHase (Fig. 4).

the importance of 'the ability to obtain an entire inventory of coexisting states of a macromolecule from a single sample' has been prominently emphasized in the field [27]. It is further stressed that much room for improvement remains and that current methods have many drawbacks such as the inability to automatically identify the number of conformational states. In contrast, the algorithm described in this paper is capable of estimating the original underlying conformations to a high degree of accuracy, without any prior knowledge about the number of conformations or any prior structural information. Further, our method applies to a broad range of proteins, with the number of conformations ranging from two to eight, even under high amounts of noise. In future works, our method will be extended to the 3D case and tested on Cryo-EM datasets. Moreover, searching for other invariant features over and above the zeroth order moments is also an important avenue of investigation.

Supplemental material: For the authors implementation and additional results refer to the supplemental material.

Note: The authors have submitted a different paper [28] which exclusively deals with reconstruction of a *single* conformation and focuses on different types of *outliers* in the projections. The problem of heterogeneity cannot be solved by simply considering projections of other conformations as outliers, during the reconstruction of one conformation. This is because the algorithm in [28] deals with a limited percentage of outliers.

7. REFERENCES

[1] J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, Oxford University Press, 3 2006.

[2] Y. Cheng et al., "Single Particle Reconstructions of the TransferrinTransferrin Receptor Complex Obtained with Different Specimen Preparation Techniques," *Journal of Molecular Biology*, vol. 355, no. 5, pp. 1048–1065, 2 2006.

[3] N. Elad et al., "Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections," *Journal of Structural Biology*, vol. 162, no. 1, pp. 108–120, 4 2008.

[4] N. A. Ranson et al., "Allosteric signaling of ATP hydrolysis in GroELGroES complexes," *Nature Structural & Molecular Biology*, vol. 13, no. 2, pp. 147–152, 2 2006.

[5] S. C. Blanchard et al., "tRNA dynamics on the ribosome during translation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 35, pp. 12893–8, 2004.

[6] W. R. Wikoff et al., "Time-resolved molecular dynamics of bacteriophage HK97 capsid maturation interpreted by electron cryo-microscopy and X-ray crystallography," *Journal of Structural Biology*, vol. 153, no. 3, pp. 300–306, 3 2006.

[7] JB. Heymann, JF. Conway, and AC. Steven, "Molecular dynamics of protein complexes from four-dimensional cryo-electron microscopy," *Journal of Structural Biology*, vol. 147, no. 3, pp. 291 – 301, 2004, Time-Resolved Imaging of Macromolecular Processes and Interactions.

[8] H. Gao, M. Valle, M. Ehrenberg, and J. Frank, "Dynamics of EF-G interaction with the ribosome explored by classification of a heterogeneous cryo-EM dataset," *Journal of Structural Biology*, vol. 147, no. 3, pp. 283–290, 9 2004.

[9] BB. Cheikh, E. Baudrier, and G. Frey, "A tomographical reconstruction method from unknown direction projections for 2D gray-level images," *Pattern Recognition Letters*, vol. 86, pp. 49–55, 1 2017.

[10] S. H. W. Scheres, "A Bayesian View on Cryo-EM Structure Determination," *Journal of Molecular Biology*, vol. 415, no. 2, pp. 406–418, 1 2012.

[11] N. Grigorieff, "FREALIGN: High-resolution refinement of single particle structures," *Journal of Structural Biology*, vol. 157, no. 1, pp. 117 – 125, 2007, Software tools for macromolecular microscopy.

[12] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination," *Nature Methods*, vol. 14, no. 3, pp. 290–296, 3 2017.

[13] S. Basu and Y. Bresler, "Feasibility of tomography with unknown view angles," *IEEE Transactions on Image Processing*, vol. 9, no. 6, pp. 1107–1122, 6 2000.

[14] A. Singer and H.-T. Wu, "Two-Dimensional Tomography from Noisy Projections Taken at Unknown Random Directions," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, pp. 136–175, 1 2013.

[15] Y. Michels and E. Baudrier, "Retrieving the parameters of cryo electron microscopy dataset in the heterogeneous ab-initio case," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3189–3193.

[16] P. Huber, *Robust Statistics*, John Wiley & Sons, 2009.

[17] C.O.S Sorzano et al., "A clustering approach to multireference alignment of single-particle projections in electron microscopy," *Journal of Structural Biology*, vol. 171, pp. 197–206, 2010.

[18] J. Vargas et al., "Particle alignment reliability in single particle electron cryomicroscopy: a general approach," *Scientific Reports*, vol. 6, no. 1, pp. 21626, 4 2016.

[19] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *Comput. J.*, vol. 26, pp. 354–359, 1983.

[20] SC. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep 1967.

[21] D.D. Muresan and T.W. Parks, "Adaptive principal components and image denoising," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*. vol. 1, pp. 101–4, IEEE.

[22] M. Baumann and S. Meri, "Techniques for studying protein heterogeneity and post-translational modifications," *Expert review of proteomics*, vol. 1, no. 2, pp. 207–217, 2004.

[23] F. Natterer, *The Mathematics of Computerized Tomography*, Society for Industrial and Applied Mathematics, 1 2001.

[24] RR. Coifman et al., "Graph Laplacian tomography from unknown random projections," *IEEE Transactions on Image Processing*, 2008.

[25] E. Malhotra and A. Rajwade, "Tomographic reconstruction from projections with unknown view angles exploiting moment-based relationships," in *2016 IEEE International Conference on Image Processing (ICIP)*. 9 2016, pp. 1759–1763, IEEE.

[26] M. Gerstein and W. Krebs, "A database of macromolecular motions.," *Nucleic acids research*, vol. 26 18, pp. 4280–90, 1998.

[27] J. Frank, "Exploring the dynamics of supramolecular machines with cryo-electron microscopy," in *New Chemistry and New Opportunities from the Expanding Protein Universe*. 12 2014, pp. 313–317, World Scientific.

[28] R. Chaudhry, A. Ghosh, and A. Rajwade, "Noise-and Outlier-Resistant Tomographic Reconstruction under Unknown Viewing Parameters," *Submitted to 2019 IEEE International Conference on Image Processing (ICIP)*.