

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Master's Theses, 2022-2026

Computer Science, 2023

2024

Tools for Analysis of Large-Scale Cryo-EM and Cryo-ET Data

Raghu, Rishwanth

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01ft848v014>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

TOOLS FOR ANALYSIS OF LARGE-SCALE CRYO-EM AND CRYO-ET DATA

RISHWANTH RAGHU

A MASTER'S THESIS
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF MASTER OF SCIENCE

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
COMPUTER SCIENCE

ADVISOR: ELLEN D ZHONG

MAY 2024

© Copyright by Rishwanth Raghu, 2024. All rights reserved.

Abstract

Cryo-electron microscopy (cryo-EM) is a powerful technique for visualizing the structure and dynamics of macromolecules in near-physiological conditions. Owing to cryo-EM’s ability to image biomolecules in varying conformations, methods are increasingly being developed to reconstruct heterogeneous ensembles of 3D biomolecular structures from 2D cryo-EM image datasets. However, the field lacks standardized benchmarks for quantitative evaluation of these methods. We present CryoBench, a set of challenging synthetic cryo-EM datasets and novel metrics for the evaluation of heterogeneous reconstruction methods, along with benchmarking analysis of existing state-of-the-art methods. Next, we explore the semantic segmentation task for cryo-electron tomography (cryo-ET), a related biophysical technique for visualizing macromolecules within their cellular environments. We quantitatively benchmark prevailing supervised training approaches on two datasets, while exploring work towards more generalizable segmentation of large-scale cryo-ET data given only sparse manual annotations.

Acknowledgements

Thank you to my advisor, Professor Ellen Zhong, for her dedicated guidance and support in helping me grow as a researcher. Thank you to the members of the E.Z. lab for all their helpful feedback and contributions to this work. Through my thesis, I've found a deep passion for research in machine learning for structural biology.

Thank you to Mazdak Abulnaga, Professor Gira Bhabha, Professor Damian Ekiert, Mahrukh Usmani, and Nicolas Coudray for their guidance on the cryo-ET segmentation project.

Thank you to my second reader, Professor Ben Raphael, for his feedback on this manuscript.

Thank you to my friends for providing endless encouragement and laughter throughout my years at Princeton.

Thank you to my parents and brother for their unwavering support of my personal journey.

Contents

Abstract	3
Acknowledgements	4
1 Introduction	7
1.1 Cryo-electron Microscopy	7
1.2 The Cryo-EM Reconstruction Problem	8
1.3 Cryo-electron Tomography	9
2 Synthetic Benchmarks for Heterogeneous Cryo-EM Reconstruction	11
2.1 Motivation	11
2.2 Related Work	13
2.3 Synthetic Dataset Generation	13
2.3.1 Simulation of Heterogeneity	13
2.3.2 Simulation of Images	15
2.4 Evaluation Metrics	16
2.4.1 Latent Space Metrics	16
2.4.2 Reconstruction Quality Metrics	17
2.5 Benchmarking	18
2.5.1 Methods	18
2.5.2 Results	20
2.6 Summary and Future Work	21
3 Automated Segmentation of Cryo-electron Tomograms	24
3.1 Motivation	24

3.2	Related Work	25
3.2.1	Cryo-ET Segmentation	25
3.2.2	Diffusion Models for Segmentation	25
3.3	Methods	26
3.3.1	Supervised Learning	26
3.3.2	Unsupervised Learning	27
3.3.3	Metrics	29
3.4	Results	30
3.4.1	Datasets	30
3.4.2	Single Tomogram Supervised Learning	31
3.4.3	Generalization	32
3.5	Summary and Future Work	35
4	Summary	36
	Appendix A: Supplemental Information for Chapter 2	37
	Appendix B: Supplemental Information for Chapter 3	39
	Bibliography	50

Chapter 1

Introduction

Cryo-electron microscopy (cryo-EM) and cryo-electron tomography (cryo-ET) have emerged as powerful techniques for visualizing biomolecules in near-physiological conditions [1]. They have proven to be key tools in advancing structural biology research, as well as downstream applications such as drug discovery [2]. As cryo-EM and cryo-ET data grows in both abundance and quality, new computational tools are required uncover the breadth of biological insights present in the data. In this chapter, we review the basics of the cryo-EM imaging process, an important class of computational tasks called the cryo-EM reconstruction problem, and the cryo-ET imaging process. In Chapter 2, we present CryoBench, a novel set of benchmark datasets and evaluation pipelines designed to accelerate methods develop in cryo-EM reconstruction. In Chapter 3, we present ongoing work in novel algorithms for the semantic segmentation of cryo-electron tomograms, which contain rich structural information about biomolecules and organelles *in situ*. In Chapter 4, we summarize our contributions and look ahead at future work in the field.

1.1 Cryo-electron Microscopy

Cryo-electron microscopy (cryo-EM) is a biophysical technique that allows for imaging biomolecules in their near-native distributions. As depicted in Figure 1.1a, a sample of a protein of interest is purified, vitrified, and imaged under an electron microscopy [7]. Due to the flash freezing process, molecules within the sample may assumed to be in

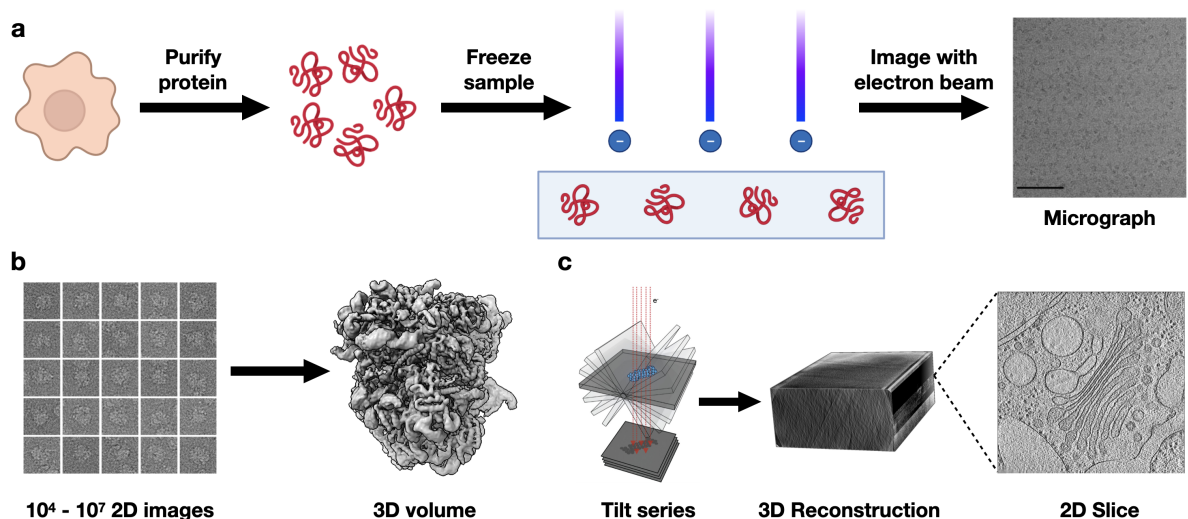


Figure 1.1: **Overview of cryo-EM and cryo-ET.** (a) In cryo-electron microscopy (cryo-EM), a sample of a protein of interest is purified from the cell, vitrified in ice, and imaged under an electron beam. Created with BioRender. Micrograph image adapted from Walls et al. [3]. (b) The cryo-EM reconstruction problem involves constructing one or more 3D density volumes from 2D particle images. Shown is the *plasmodium falciparum* 80s ribosome [4]. (c) Cryo-electron tomography is a technique in which tomographic projections of a thin cell section are acquired at different angles and reconstructed into a 3D volume, allowing for visualization of biomolecules within their cellular context. Adapted from Turk and Baumeister [5] and Bykov et al. [6].)

conformations consistent with their free energy landscape under biological conditions. Owing to this preservation of aqueous state, as well as the lack of an upper limit on the size of particles that can be imaged, cryo-EM has emerged as a powerful technique with unique advantages over prior methods such as x-ray crystallography. From the 2D images produced by cryo-EM, computational pipelines can be used to reconstruct the 3D structure and dynamics of a protein, delivering rich biological insights about structure, function, and interactions of macromolecules.

1.2 The Cryo-EM Reconstruction Problem

The cryo-EM acquisition process produces “movies” of different areas of the sample. Motion correction algorithms can be applied to each movie to produce a corresponding 2D micrograph. A micrograph represents the projected electron scattering potential through an area of the sample. Particle picking algorithms produce individual particle

images by extracting crops of the micrographs, typically on the order of tens of thousands to over a million.

The *cryo-EM reconstruction problem* is the task of computing one or more 3D volumes of a protein from the set of 2D particle images, as depicted in Figure 1.1b. There are various aspects that make this task challenging: the images have low signal-to-noise ratios due to radiation damage from the electron beam, some images may contain “junk” particles owing to suboptimal sample conditions, the orientation of a particle within each image is unknown, and particles may exhibit varying conformations across images. Formally, given 2D particle images $X_i, i \in \{1, \dots, N\}$, the traditional cryo-EM reconstruction problem can be formulated as jointly solving for the volume V and image poses $\phi_i = (R_i, t_i)$, where R_i is a 3D rotation and t_i is an in-plane translation, typically finding a maximum a posteriori estimate of V while marginalizing over ϕ [8]:

$$V^{\text{MAP}} = \underset{V}{\operatorname{argmax}} \sum_{i=1}^N \log p(X_i | \phi, V) p(\phi) d\phi + \log V$$

However, as mentioned before, particles in a sample may be heterogeneous, in which case we can model images as arising from one of K independent volumes, with marginalization over the probability π of belonging to a class [8]:

$$V_1, \dots, V_K \sum_{i=1}^N \log \sum_{j=1}^K (\pi_j \int p(X_i | \phi, V_j) p(\phi) d\phi) + \sum_{j=1}^K \log p(V_j)$$

It is worth noting that while such a formulation may suffice computationally for a small number of discrete protein conformations, K may be arbitrarily large for proteins that undergo continuous motions.

1.3 Cryo-electron Tomography

While cryo-EM is powerful in its ability to image biomolecules in their near-native conformations, the purification process has the downside of removing the biomolecules from their cellular context. Thus, cryo-electron tomography (cryo-ET) is emerging as a related

technique for imaging biomolecules *in situ* [5]. In this method, depicted in Figure 1.1c, tomographic projections of a thinned, vitrified section of a cell are acquired at different angles, termed a *tilt series*. The 2D projections from the tilt series can be aligned, electron dose-corrected, and backprojected to form a 3D volume displaying biomolecules and coarser cellular structures. Common computational pipelines include subtomogram averaging, in which extracted instances of a particle are averaged to obtain a high resolution structure, and segmentation, in which voxels of the tomogram are semantically annotated.

Chapter 2

Synthetic Benchmarks for Heterogeneous Cryo-EM Reconstruction

This work was done jointly with Minkyu Jeon, Alkin Kaz and Ryan Feathers.

2.1 Motivation

A single species of protein can have a highly dynamic structure. This heterogeneity can broadly be characterized as *compositional heterogeneity*, where the protein displays variations in subunit composition, or *conformational heterogeneity*, where the protein undergoes flexible motions. A key benefit of the cryo-EM acquisition process is that it freezes particles in their near-native states, thus producing images that capture the heterogeneity of the protein. The goal of the 3D reconstruction problem discussed in Section 1.2 is to not only achieve high resolution reconstruction, but to also recover the heterogeneity present in the data.

There has been a recent growth in computational methods for heterogeneous cryo-EM reconstruction, employing varying parameterizations of the space of volumes, such as through neural fields [8], as a contributions of PCA components [9], or as a learned deformation of a canonical volume [10]. However, the field lacks standardized bench-

	Biomolecule	# of images	Image size	Data Type	Noise
CryoDRGN ²³	Ribosome 80S (EMPIAR-10028)	50,000	128 x 128	Synthetic	No/Yes
		105,247	90 x 90	Experimental	Yes
	Ribosome 50S (EMPIAR-10076)	131,899	90 x 90	Experimental	Yes
	Protein complex	50,000	64 x 64	Synthetic	N/A
CryoFold ³⁵	Haemoglobin (PDB 5NI1)	50,000	128 x 128	Synthetic	Yes
3DFlex ³²	Tri-snRNP spliceosome (EMPIAR-10073)	102,500	180 x 180	Experimental	Yes
	TRPV1 ion-channel (EMPIAR-10059)	200,000	128 x 128	Experimental	Yes
CryoPoseNet ³⁴	E.coli adenylate kinase (PDB 4AKE1)	9,000	128 x 128	Synthetic	No/Yes
E2GMM ³³	Ribosome 50S (EMPIAR-10076)	124,900	N/A	Experimental	Yes
	Precatalytic spliceosome (EMPIAR-10180)	327,490	N/A	Experimental	Yes
	SARS-CoV-2 spike protein (EMPIAR-10492)	55,159	N/A	Experimental	Yes
FSTdiff ³⁹	GroEL-GroES protein	40,000	128 x 128	Synthetic	Yes
CryoVAEGAN ⁴²	Ribosome 80S (EMPIAR-10028)	2,544	128 x 128	Synthetic	No
		5,119 - 8,278 - 4,917	128 x 128	Experimental	Yes
AtomVAE ¹⁷	Aurora A Kinase (Simulated)	63,000	64 x 64	Synthetic	Yes
CryoAI ¹⁸	80S ribosome (PDB 3J79 and 3J7A)	1,000,000	128 x 128	Synthetic	No/Yes
	SARS-CoV-2 spike protein (PDB 6VYB)	1,000,000	128 x 128	Synthetic	No/Yes
	Spliceosome (PDB 5NRL)	1,000,000	128 x 128	Synthetic	No/Yes
	80S ribosome (EMPIAR-10028)	105,247	256 x 256	Experimental	Yes
CryoGAN ³⁸	β -galactosidase	41,000	180 x 180	Synthetic	Yes
	β -galactosidase (EMPIAR-10061)	41,123	192 x 192	Experimental	Yes
Multi-CryoGAN ⁴⁴	Heat-shock protein Hsp90	100,000	32 x 32	Synthetic	Yes

Table 2.1: **List of common cryo-EM reconstruction methods and their evaluation datasets.** Adapted from Donnat et al. [11].

marks on which methods can be quantitatively evaluated, owing to the lack of ground truth volumes in experimentally collected datasets. Currently, methods papers employ a combination of “benchmarking-by-eye”, wherein the reconstructions of a few canonical experimental datasets are visually inspected for quality, and synthetic datasets, which tend to have simplistic motions or blob-like particles that are not representative of natural protein structures. Even with these approaches, there is a large amount of variation in the evaluation datasets used, making it difficult to compare methods against one another, as shown in Table 2.1.

Thus the aim of this project is to construct the first standardized benchmark dataset for heterogenous cryo-EM reconstruction, which we term *CryoBench*. The contributions are threefold:

1. New synthetic datasets with challenging forms of heterogeneity to motivate methods development
2. Novel metrics for quantitative evaluation of these methods
3. Benchmarking of existing state-of-the-art methods

Armed with these datasets and metrics, we hope that methods developers are inspired to tackle the cryo-EM reconstruction task and rigorously evaluate their methods. Likewise,

we hope that the evaluation of methods on a standardized benchmark can aid methods users in choosing the right tools for their dataset.

2.2 Related Work

The Cryo-EM Heterogeneity Challenge is an ongoing public competition in which users can submit 80 reconstructed volumes representative of each of two datasets, one synthetic and one experimental [12]. The challenge focuses on recovering the underlying distribution of particles, as these datasets have non-uniform distributions over the conformations. Another related work simulates non-uniform datasets from continuous rotations of an atomic model chain [13]. The focus here is on encoder metrics and building priors into model latent spaces. We aim to supplement these works with a more expansive suite of benchmarks involving more forms of complex heterogeneity, and a focus on evaluating reconstruction quality.

2.3 Synthetic Dataset Generation

2.3.1 Simulation of Heterogeneity

Heterogeneity can be generated by introducing compositional and conformational manipulations to atomic protein structures. As shown in Figure 2.1a-c, the full benchmark consists of three forms of heterogeneity: (1) two rotational motions of a Fab domain of an antibody, simulating conformational heterogeneity; (2) stitching of different antigens to the Fab domain of the same antibody, simulating compositional heterogeneity; and (3) frames from molecular dynamics simulations of the SARS-CoV-2 spike protein in different discrete states, also simulating conformational and compositional heterogeneity, albeit with motions that are more fine-grained and realistic. For the remainder of this chapter, we focus on the molecular dynamics dataset construction and benchmarking.

Molecular dynamics (MD) is a computational technique which predicts how the atoms of a molecular system move over time based on a model of interatomic interactions [22].

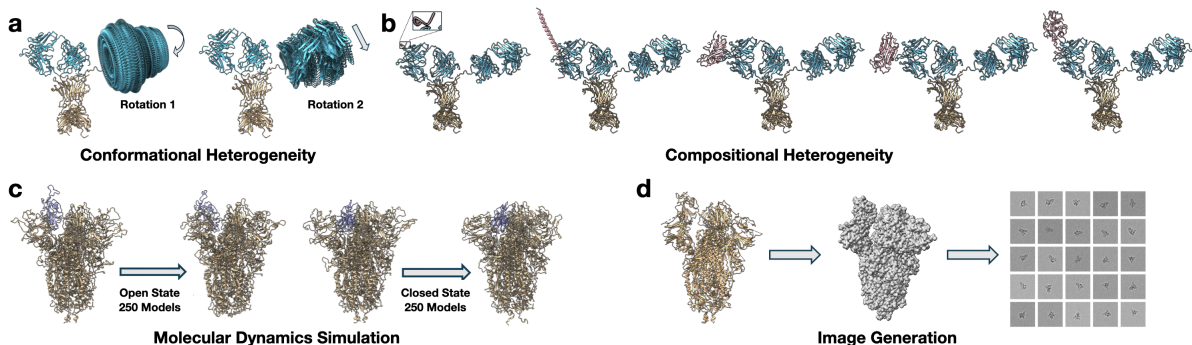


Figure 2.1: **Synthetic Dataset Construction.** (a) Conformational heterogeneity produced by rotating domains of the an antibody (PDB 1HZH [14]). (b) Compositional heterogeneity produced by stitching the antibody with different antigens (PDBs 1TJG [15], 4KHT [16], 7CHE [17], 7M6D [18], 8I5I [19]) (c) Fine-grained motions obtained from molecular dynamics simulation of the SARS-CoV-2 spike protein [20]. (d) Simulation of 2D cryo-EM images from atomic models. Adapted from Jeon et al. [21].

Since the evolution of the system is governed by a physics-based model, the states of the system are approximately biophysically realistic. MD simulations of a protein thus provide us with a natural set of heterogeneous, biophysically plausible conformations.

We use two publicly available MD simulations of the SARS-CoV-2 spike protein from D.E. Shaw Research [20]. The first is initiated in the “closed” state of the protein, in which all three receptor-binding domains (RBDs) of the trimeric protein are in their inactive position. The second is initiated in the “open” state, in which one RBD protrudes out in its active position. While the protein can naturally transition between the open and closed state in nature, this process occurs on the timescale of a few seconds [23], so we do not observe it in the simulations, which are 10 microseconds long. Thus, the conformational space can be seen as a discrete union of continuous manifolds — there are two “macrostates” corresponding to the open and closed states, and within each macrostate, there are thousands of “microstates” corresponding to fine-grained motions of the atoms. In practice, we take the first 250 frames of the open state simulation and first 250 states of the closed state simulation, for a total of 500 conformations. Additionally, we strip the atomic models of glycan chains coating the protein surface. We discuss the sampling of transition states between the two macrostates as future work in Section 2.6.

2.3.2 Simulation of Images

Simulating an image dataset from the MD simulation frames described in Section 2.3.1 involves two steps: simulating electron scattering potential maps from the atomic models corresponding to the simulation frames, then simulating images from the maps through a known forward imaging model.

In order to have a common frame of reference for the 500 atomic model frames, we align the first frame of the closed state simulation to the closed state experimental density map (via the ChimeraX *fitmap* command) [3], and then align the first frame of the open state simulation to the first frame of the closed state simulation (via the ChimeraX *align* command)[24]. Within each simulation, the frames are already aligned. Using the ChimeraX *molmap* command, the electron scattering density is simulated for each frame with a resolution of 2.1 Å/pix on a grid of dimension 400^3 with pixel size 1.05 Å/pix, matching the parameters of the experimental density map. For more efficient analysis, the maps are downsampled to dimension 256^3 (corresponding pixel size 1.64 Å/pix), though the original box size may be used in future work for higher resolution reconstruction.

For each of the 500 maps, 100 images are simulated using adapted scripts from *cryosim*, a private tool courtesy of the E.Z. Lab, for a total of 50,000 images of size 256^2 . The image simulation process follows a commonly used forward model involving (1) integrating the map along a random projection direction, (2) corrupting the resultant 2D image with the microscope’s contrast transfer function (CTF), and (3) corrupting the image with Gaussian noise. An image I can be formed from a volume V as follows:

$$I_i = C_i * P_{\phi_i} V + \eta_i$$

where P_{ϕ} is a projection along an axis followed by an in-plane translation, C is the CTF and η is Gaussian noise. We randomly sample poses $\phi \in SO(3) \times \mathbb{R}^2$, where projection directions are uniformly drawn from $SO(3)$ and in-plane translations are uniformly drawn from $[-20, 20]^2$ pixels. CTF values are taken from the experimental values of Veesler et al. with defoci clipped at 2.5 μm [3], and noise is added such that the signal-to-noise

ratio (SNR) is 0.1. The image simulation process is shown in Figure 2.1d.

2.4 Evaluation Metrics

2.4.1 Latent Space Metrics

Many cryo-EM reconstructions produce a latent space, which can be interpreted as a mapping from a particle image to the conformation of its imaged particle. While most methods are not trained to produce an interpretable latent space, they often do empirically, and it can be useful to quantify how well a latent space correlates with the underlying distribution of conformations.

Discrete Heterogeneity. To assess discrete heterogeneity in the latent space, we use a commonly employed metric for discrete clustering, the Adjusted Rand Index (ARI). This metric can be seen simply as a classification accuracy invariant to the ordering of the clusters. For the MD dataset, we define two ground truth clusters corresponding to the open and closes states of the protein. The predicted clusters are obtained by running k -means clustering with $k = 2$ on the latent space of a trained model.

Continuous Heterogeneity. To assess continuous heterogeneity in the latent space, we use the following procedure, culminating in a metric based on the Spearman rank correlation:

1. For each pair of particle images, we compute the distance between the atomic models they were generated from, where distance is computed in the space of a collective variable describing the dominant axes of variation.
2. For each pair of particle images, we compute the distance between their embeddings in the latent space of a trained reconstruction model.
3. We compute the Spearman rank correlation between the “true” distances from Step 1, and the “predicted” distances from Step 2. As some of the models encode images nonlinearly, we opt for the rank correlation over the Pearson correlation, the latter of which tests for a linear relationship.

For the MD dataset, we compute this metric separately for the open state images and closed state images, and average the two. For each of the two, we define the collective variable as the first two principal component values of the respective trajectory, where each datapoint of the PCA corresponds to a timestep of the trajectory, and the features for a timestep are $3N$ atomic coordinates (Appendix Figure A1). Since reconstruction models have latent spaces of varying dimensionality, we also apply PCA with two components to these latent spaces. Euclidean distances are computed between pairs of the collective variables, and between pairs of the PCA-transformed latent embeddings.

2.4.2 Reconstruction Quality Metrics

In this section, we propose two metrics for assessing the quality of a distribution of reconstructed volumes, based on the Fourier Shell Correlation. The Fourier Shell Correlation (FSC) is commonly used as a measure of the resolution of a reconstruction. It is defined as the cross-correlation coefficient between two 3D volumes over corresponding shells in Fourier space[25]:

$$\text{FSC}_t(U, V) = \frac{\sum_{s \in S_t} U_s V_s}{\sqrt{\sum_{s \in S_t} |U_s|^2 \sum_{s \in S_t} |V_s|^2}}$$

where U and V are the Fourier transforms of the two volumes, respectively, and S_t is the set of voxels corresponding to resolution t in Fourier space. Typically, an experimental dataset is split into two random halves, a volume is reconstructed from each, an FSC curve is computed over different resolutions, and the resolution is chosen to be its value when the FSC is 0.143 or 0.5, depending on the application [26]. However, in the synthetic data setting with low noise level, the FSC curve often stays high (above 0.5) and we may be interested in FSCs over a range of resolutions. Thus we propose using the area under the FSC curve (AUC FSC) as the basis for our proposed metrics.

Volume-FSC. We define the first metric, Volume-FSC, as follows:

$$\frac{1}{N} \sum_n \max_m \text{AUC}_t [\text{FSC}_t(R_n, G_m)]$$

where $R_n, n \in \{1, \dots, N\}$ is a set of reconstructed volumes and $G_m, m \in \{1, \dots, M\}$

is the set of ground truth volumes. Volume-FSC is the maximum AUC FSC between a reconstructed volume and any ground truth volume, averaged over N representative reconstructions. It asks, for any volume we reconstruct, how well would it match any volume in the ground truth distribution. In practice, we use k -means cluster centers in the latent space, with $k = 10$, as the representative reconstructions.

End-to-end-FSC. We define the second metric, End-to-end-FSC, as follows:

$$\frac{1}{N} \sum_n \text{AUC}_t [\text{FSC}_t(R_n, G_{I_n})]$$

where R_n is a set of reconstructed volumes corresponding to particle images $I_n, n \in \{1, \dots, N\}$, and G_{I_n} is the ground truth volume I_n was generated from. End-to-end-FSC is the AUC FSC between reconstructed volumes and ground truth volumes corresponding to matching conformations, averaged over N particle images representative of different conformations. For a method to score well on this metric, it must both (1) accurately map a particle image to a conformation, and (2) produce a high quality reconstruction of that conformation. In the context of encoder-decoder based methods, End-to-end-FSC jointly assesses encoder and decoder performance, whereas Volume-FSC assesses decoder performance alone. In practice, we uniformly sample 100 images to use as the representative images. Visual explanations of the two metrics are shown in Figure 2.2.

2.5 Benchmarking

2.5.1 Methods

Given the aforementioned datasets and metrics, we aim to benchmark state-of-the-art methods. In this section, we briefly describe the methods tested. We test *fixed pose* methods, which take ground truth poses as input, and *ab initio methods*, which jointly optimize the poses along with the reconstructions.

- 3D Classification uses expectation maximization to reconstruct K discrete volumes [*fixed pose* and *ab initio*] [27]

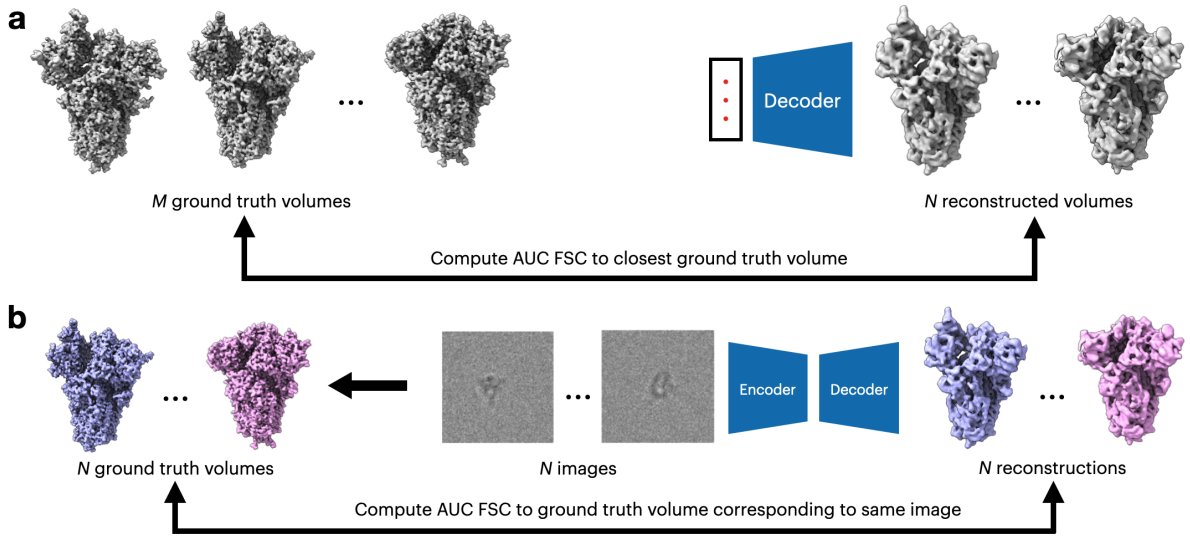


Figure 2.2: **Metrics for evaluation of reconstructions.** (a) *Volume-FSC* assesses reconstruction quality alone by comparing each reconstructed volume with its closest ground truth volume. (b) *End-to-end-FSC* jointly assesses latent inference accuracy and reconstruction quality by comparing pairs of reconstructions and ground truth volumes corresponding to the same particle images.

- 3DVA uncovers a linear subspace in which an image is assumed to be generated from a mean volume plus weighted contributions of variability components [*fixed pose*] [9]
- 3DFlex models volumes as a deep network-parameterized deformation of a canonical volume [*fixed pose*] [10]
- CryoDRGN employs neural fields to implicitly represent a continuous manifold of volumes, allowing for heterogeneous volume generation by conditioning on a latent variable representing particle conformation [*fixed pose* and *ab initio methods*][8] [28]
- DRGN-AI, a method in preparation for publication and built on top of CryoDRGN, introduces autodecoding and a more scalable pose search methodology employing a combination of grid search and gradient descent [*fixed pose* and *ab initio methods*]
- OPUS-DSD, built on top of CryoDRGN, adds 3D convolutional layers and stronger latent space priors [*fixed pose*][29]

In our experiments, for OPUS-DSD, the downsampling ratio is set to 1.0, the template

Method	Volume-FSC	End-to-End-FSC	ARI	Spearman
3D Class (<i>fixed pose</i>)	0.288 (0.025)	0.285 (0.016)	0.999	-
3DVA	0.297 (0.017)	0.299 (0.017)	1.000	0.698
3DFlex	0.220 (0.009)	0.207 (0.010)	0.745	0.704
CryoDRGN (<i>fixed pose</i>)	0.316 (0.013)	0.307 (0.015)	0.987	0.826
DRGN-AI (<i>fixed pose</i>)	0.261 (0.006)	0.254 (0.010)	0.972	0.877
Opus-DSD	0.167 (0.021)	0.169 (0.007)	0.971	0.718
3D Class (<i>ab initio</i>)	0.209 (0.086)	0.251 (0.035)	0.999	-
CryoDRGN (<i>ab initio</i>)	0.157 (0.071)	0.207 (0.052)	0.000	0.060
DRGN-AI (<i>ab initio</i>)	0.257 (0.006)	0.250 (0.010)	0.968	0.710

Table 2.2: **Quantitative benchmarking of MD dataset.** *Volume-FSC* is averaged across K-Means centroids (K=10). *End-to-End-FSC* is averaged across 100 reconstructions. Standard deviation is given in parentheses.

resolution is set to 256, output volumes are padded to size 256, and the model is trained for 15 epochs. For *ab initio* DRGN-AI, the batch size of SGD is reduced to 128. All other hyperparameters are set at their defaults.

2.5.2 Results

Qualitative Results. As shown in Figure 2.3, volumes are generally reconstructed well, with *fixed pose* methods capturing smaller conformational changes, for example, in the tip of the receptor-binding domain. Amongst *ab initio* methods, DRGN-AI produces sensible volumes, while 3D Classification and CryoDRGN produce some blob-like volumes likely corresponding to misaligned images. Both variants of 3D Classification produce fewer than the requested 10 classes. Both the *fixed pose* methods and *ab initio* DRGN-AI are able to separate the open and closed states in the latent space. *Ab initio* CryoDRGN produces uninterpretable clusters in the latent space, once again likely to correspond to misaligned images. Interestingly, within the open and closed clusters, images are roughly ordered by timestep of their respective MD simulation, suggesting that the timestep is correlated with a dominant degree of conformational change.

Quantitative results. As shown in Table 2.2, CryoDRGN achieves the highest Volume-FSC and End-to-end-FSC scores. 3DFlex performs relatively poorly, as its assumption of a deformation of a canonical volume is not appropriate for learning the discrete hetero-

geneity in this dataset. Amongst *ab initio* methods, DRGN-AI outperforms on Volume-FSC, while DRGN-AI and 3D Classification show similar End-to-end-FSCs. For the *ab initio* variants of CryoDRGN and 3D Classification, the End-to-end-FSC is significantly higher than the Volume-FSC, which may be attributed to the presence of poorly reconstructed centroid volumes associated with poorly aligned images, in the case of Volume-FSC. Considering ARI scores, most methods classify images nearly perfectly. 3DVA in particular classifies every image correctly, while *ab initio* CryoDRGN is unable to distinguish the two states at all, likely due to a spurious cluster of poorly aligned images. Likewise, all methods achieve high Spearman correlations except *ab initio* CryoDRGN, with *fixed pose* DRGN-AI achieving the highest. For many methods, there is a considerable difference between the correlations for the open closed states (Appendix Table A1).

2.6 Summary and Future Work

We present a novel benchmark for heterogeneous cryo-EM reconstruction, consisting of three synthetic datasets with challenging forms of heterogeneity, novel metrics for quantitative evaluation of reconstructions, and benchmarking analysis of state-of-the-art methods. We hope that the benchmark suite provides methods developers with a testbed for rigorous algorithmic development, and methods users with a guide to choosing the right methods for their applications.

Ongoing work on this project aims to construct datasets with even more challenging forms of heterogeneity. For compositional heterogeneity, we intend to construct datasets with mixtures of unrelated proteins, or mixtures of a protein with a large degree of variation in subunit composition. For conformational heterogeneity, we intend to construct datasets comprising proteins with highly flexible domains that move freely, rather than being constrained to one or two degrees of freedom. We are also constructing a new molecular dynamics dataset that samples transition states between the open and closed states, using metadynamics simulations from Wieczor et al. [23]. Future work can also include

other challenging aspects of real cryo-EM data, such as non-uniform pose distributions or the presence of “junk” particles.

Additionally, while we employ a simple image formation model consisting of Gaussian noise, more realistic image simulation models can be used. For example, it is possible to simulate the effects of the sample ice, as well as electron beam-induced motion in the sample [30]. Future work can quantify the downstream reconstruction effects of the choice of noise model.

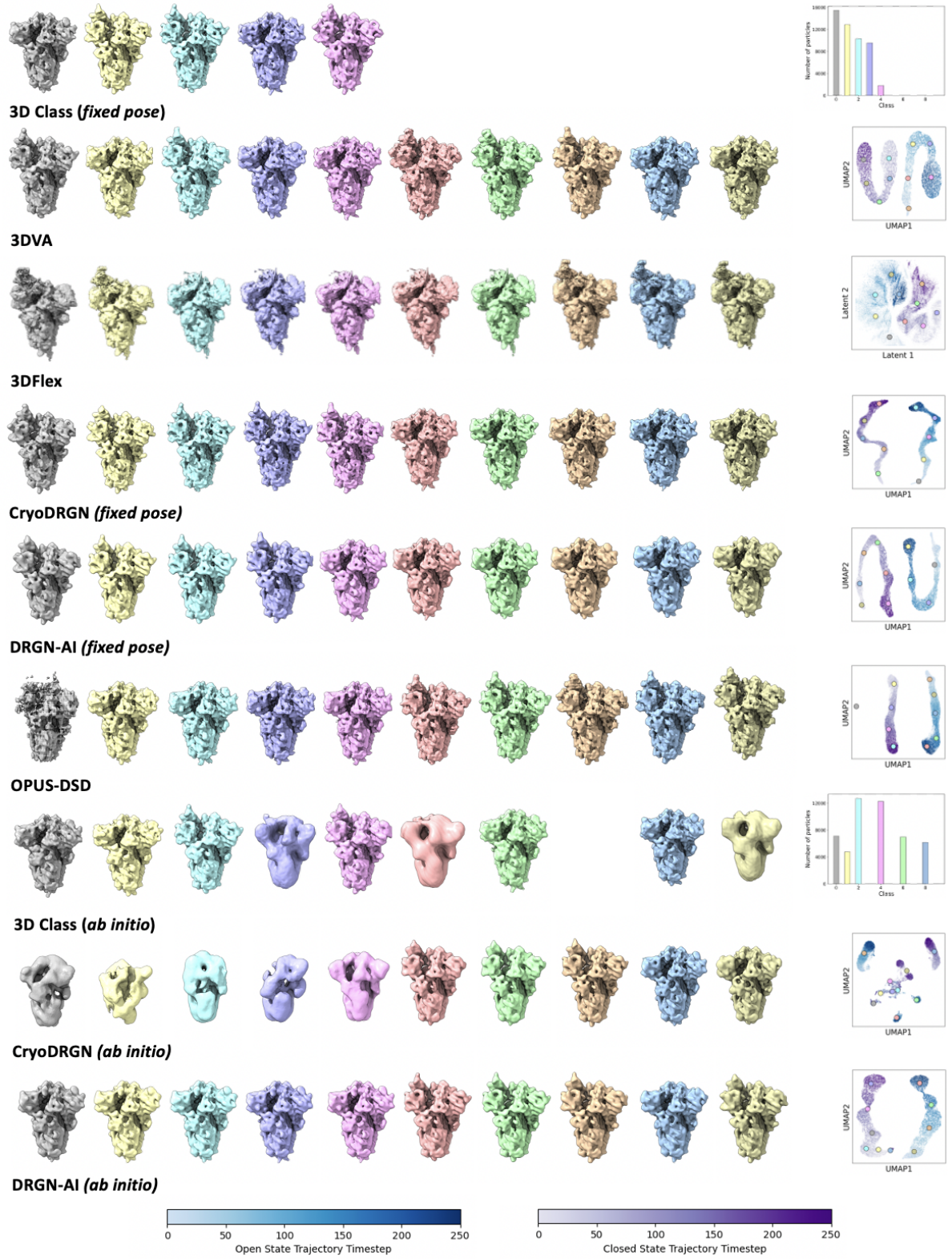


Figure 2.3: **Representative reconstructions and latent space visualizations.** Latent coordinates for each image are colored by state and simulation timestep. Latent coordinates for each reconstruction are marked with a dot of the same color.

Chapter 3

Automated Segmentation of Cryo-electron Tomograms

3.1 Motivation

Cryo-electron tomography (cryo-ET), discussed in Section 1.3, is a powerful technique for visualizing biomolecules and organelles within their cellular contexts. Semantic segmentation is common cryo-ET analysis in which each voxel of a tomographic volume is annotated with a class label corresponding to the cellular structure it belongs to. Bykov et al. use segmentation as part of a workflow to characterize changes in cellular vesicle morphology as a function of position along Golgi membranes [6]. Tao et al. segment synaptic membranes to quantify structural differences between pre-synaptic and post-synaptic neurons [31].

Segmentation workflows are complicated by the fact that tomograms are extremely noisy due to the limited electron dose used in biological TEM, possess anisotropic resolution due to the limited range of tilt angles, and contain small-scale structures that are difficult to distinguish from one another. However, manual segmentation is a laborious process that may take multiple weeks even for a single tomogram, which can be composed of several hundreds of 2D slices, underscoring the need for automated segmentation tools. While such tools have been proposed and work well for certain applications, they can of-

ten segment only one kind of structure, or scale poorly to large-scale datasets. In this chapter, we review the state-of-the-art in cryo-ET segmentation, introduce a user-friendly tool for multiclass segmentation given only sparse manual annotations, and argue for new segmentation paradigms that leverage large-scale data.

3.2 Related Work

3.2.1 Cryo-ET Segmentation

Single-Class Segmentation. TomoSegMemTV uses local gradient information and a tensor voting algorithm to segment all membranes [32]. MemBrainSeg trains a deep neural network on various kinds of tomograms to provide a generalizable pretrained model for segmentation of all membranes [33]. In contrast, we aim to tackle the multiclass segmentation problem, where a user may be interested in defining their own classes for their unique organism.

Multi-Class Segmentation. EMAN2 allows for constructing independent annotations and training a separate network for each class of interest, thus being unable to leverage information across classes [34]. Dragonfly and Amira allows for training with user-defined classes, but both tools are proprietary and support an interactive workflow not amenable to large-scale training and evaluation [35] [36]. The closest tool to our work is DeePiCt, which supports training on 2D or 3D annotations [37]. We supplement this work with another open-source, user-friendly supervised segmentation tool that supports a greater variety of loss functions and data augmentations, in addition to providing benchmarking results on unexplored datasets.

3.2.2 Diffusion Models for Segmentation

Synthetic Data Generation. DiffInfinite generates large histopathology images by first diffusing a segmentation mask, then diffusing image patches conditioned on the mask [38]. SatSynth jointly generates satellite images and segmentation masks by training a diffusion model on the channel-wise concatenation of image-mask pairs, with some added tricks

for stability in learning the discrete-valued masks [39]. DiffuMask uses text-image cross-attention weights in a Stable Diffusion model to localize class information and generate pixel-level segmentation masks [40] [41].

Supervised Segmentation. Rahman et al., Wolleb et al., and Wu et al. each propose using diffusion models as a supervised segmentation architecture, where the generative process allows for sampling of multiple plausible outputs and thus uncertainty quantification [42] [43] [44].

Unsupervised Representation Learning. Baranchuk et al. train an unconditional denoising diffusion probabilistic model, aggregate intermediate features across layers and diffusion timesteps, and train a lightweight segmentation classifier on the resultant per-pixel features for a subset of annotated data [45] [46]. DiffSeg extracts attention maps from Stable Diffusion, uses a novel algorithm to aggregate the maps across layers, and produces a set of object proposals through an iterative merging technique [40] [47].

3.3 Methods

3.3.1 Supervised Learning

We employ a standard supervised segmentation approach in which a 2D U-Net model, with 4 blocks in the contracting and expanding paths, takes 2D tomogram patches as input and outputs pseudo-probabilities over K possible class labels (Figure 3.3a) [48]. The user has control over the following parameters:

- Preprocessing options, including CTF deconvolution and simple image filters such as histogram equalization
- Input construction options, such as random or uniform patch selection from tomogram slices, as well as the ability to include adjacent slices as context (“2.5D” inputs)
- Different loss functions, including simple losses like cross-entropy, a membrane-specific loss Surface DICE as proposed in [33], and supervision of the intermediate

U-Net layers

- Different data augmentations, including simple transforms like Gaussian noise and geometric operators, as well as the tomogram-specific missing wedge augmentation and Fourier amplitude matching proposed in [33]

We next describe some of the cryo-ET specific training options and show examples in Figure B1.

Surface DICE Loss. Here, the predicted and ground truth segmentation masks are eroded to produce thin “skeletonized” segmentations for comparison. In the cryo-ET setting, pixels along the borders of membranes often have ambiguous true class labels. Thus, this loss forces the model to focus on accurate prediction of morphologies over pixel-accurate prediction.

Missing Wedge Augmentation. Tomograms suffer from a missing wedge artifact, in which the limited series of tilt angles (typically -60 to +60) leads to decreased information content in some areas of the Fourier spectrum after backprojection of the tilt images. We further impose this corruption on the 2D patches by masking 2D wedges in their Fourier spectra.

Fourier Amplitude Matching. Different tomograms have different patterns in their Fourier spectra. To emulate 2D patches from one tomogram in the style of another, their Fourier spectra can be corrupted. As proposed in [33], we create a 1D sequence half the dimension of the patch via a random walk, convert the 1D sequence to a 2D radial kernel, and multiply it with the Fourier transform of the patch.

3.3.2 Unsupervised Learning

The prevailing paradigm for segmentation of cryo-ET data is supervised learning on some manually annotated slices. We posit that while such approaches work well in certain settings, automated segmentation may be improved by leveraging the vast amounts of unannotated slices in a dataset. Diffusion models have been shown to be useful in segmentation tasks (Section 3.2.2). Here, we explore the use of diffusion models as an

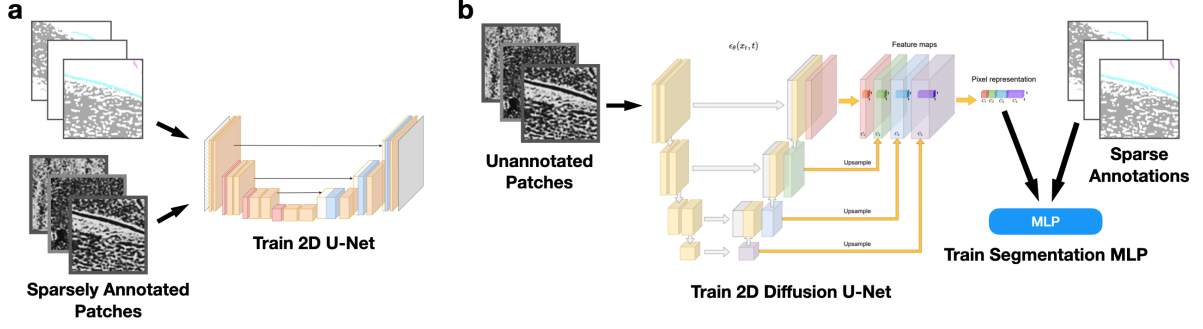


Figure 3.1: **Cryo-ET segmentation approaches.** (a) A U-Net model is trained in a supervised fashion on sparsely annotated subset of the dataset. (b) A diffusion model is trained on the entire dataset. A lightweight segmentation model is then trained on the sparsely annotated subset using unsupervised representations from the diffusion model. Adapted from Baranchuk et al. [46].

unsupervised representation learner for cryo-ET segmentation.

Proof-of-Concept on Volume EM Data. Prior to tackling the diffusion of cryo-ET data, which is noisy and lesser in amount, we desire a proof-of-concept that recent diffusion models can be applied on biological imaging. Volume EM (VEM) is a modality in which an entire volumetric sample is imaged under an electron microscope slice by slice [49]. We use a VEM dataset of the entire *Platynereis dumerilii* organism, which contains 11416 slices of dimension 27499×25916 [50]. We extract cellular regions and uniformly crop a subset of the slices into patches of dimension 512×512 , creating a dataset with roughly 50k images. We then train four state-of-the-art diffusion models (with default hyperparameters) on this data: Stable Diffusion via the LoRA finetuning paradigm [40] [51], Stable Diffusion via the DreamBooth finetuning paradigm [52], an unconditional denoising diffusion probabilistic model (DDPM) [45], and an unconditional latent diffusion model (LDM) with a KL-VAE autoencoder trained from scratch [40]. Results are shown in Appendix Figure B2 and Table B1.

Diffusion-based Segmentation of Cryo-ET Data. From the experiments on VEM data, we identify DDPM as the most promising choice of diffusion model for our biological data. We extract patches from our cryo-ET datasets (described in Section 3.4.1) and train DDPM models on them. We then re-implement DiffSeg, a recent work which uses the intermediate representations of the trained diffusion model for segmentation

[47]. The input images are noised corresponding to timesteps $t=50$ and $t=100$, which are both early in the forward diffusion process, and thus are likely to contain strong semantic information, unlike late timesteps in which the image would be nearly pure noise. The intermediate features of the U-Net of the trained diffusion model are extracted and upsampled to the input image dimensions. These feature maps are then concatenated across all layers and across both timesteps. This results in a 4608-length feature vector for each pixel. A lightweight, 1-layer MLP with 128 hidden units is trained on the per-pixel features to output pseudo-probabilities over the K possible class labels. We emphasize that the diffusion model is trained on the entire dataset of unannotated patches, while the segmentation network is trained only on the sparse annotated subset of data that is used to train the purely supervised approach described in Section 3.3.1.

3.3.3 Metrics

Segmentation Metrics. The main metric used to quantify segmentation accuracy is the DICE score, a simple measure of overlap between the predicted and ground truth segmentation masks [53]. The DICE score is defined as $\frac{2|X \cap Y|}{|X| + |Y|}$, where X and Y are the predicted and ground truth segmentations. For supervised segmentation results on single tomograms, we also report the Surface DICE score (one minus the Surface DICE loss described in Section 3.3.1), as well as the Hausdorff-95 distance, which is the maximal distance in pixel space between a pixel in the predicted segmentation and its closest pixel in the ground truth annotation, after excluding the top 5th percentile of such distances [54].

Diffusion Metrics. To assess the quality of diffusion samples, we report the Frechet Inception Distance (FID) between 10k real images and 10k generated images, where FID is a Wasserstein distance between Gaussians fit to InceptionV3 features of both sets of images [55]. We also report the Kernel Inception Distance (KID), which relaxes the Gaussian assumption and is unbiased to the number of samples [56]. We note that as InceptionV3 is not trained on biological images, its features are not the ideal image embeddings in our domain [57]. However, a correlation is observed between the FID/KID

values and the visual quality of the generated samples, suggesting that these metrics suffice for a rough comparison of different diffusion models.

3.4 Results

3.4.1 Datasets

CZII Mitochondria. We download a dataset of tomograms containing mammalian mitochondria from the Chan Zuckering Imaging Institute CryoET Data Portal [58]. These include corresponding ground truth segmentation masks generated by manually correcting and assigning classes to a TomoSegMemTV automated membrane segmentation [32]. Four tomograms are excluded due to misaligned annotations, one tomogram is excluded due to annotation quality, and one tomogram is excluded as it is binned to a lower resolution than the others, leaving a dataset of 28 tomograms. We keep two class labels, corresponding to the inner mitochondrial membrane and outer mitochondrial membrane, while treating pixels of the endoplasmic reticulum class as background due to inconsistent annotations. To assess the effect of ground truth annotation uncertainty, for one tomogram, we create a second ground truth segmentation by predicting all membranes with MemBrainSeg [33], then manually cleaning the segmented pixels and assigning class labels in Napari [59].

PolNet Synthetic. PolNet is a tool for the flexible generation of synthetic tomograms [60]. It first builds a 3D model of the cell from membrane-like parametric surfaces, actin-like and microtubule-like spherical arrangements, and atomic protein models from the PDB, after which it simulates tilt series acquisition followed by tomographic reconstruction. We run the simulator with different random seeds (and all other hyperparameters at default) to generate 16 synthetic tomograms composed of 250 slices of dimension 1024×1024 . A benefit of the synthetic approach is that the underlying 3D model also serves as a perfect ground truth segmentation. We merge the membrane classes and protein classes, respectively, such that there are 4 classes of interest: membranes, actin, microtubules, and proteins.

3.4.2 Single Tomogram Supervised Learning

For one tomogram of the CZII Mitochondria dataset, we conduct extensive hyperparameter tuning and ablations, including testing both variants of the ground truth annotation as described in Section 3.4.1. We train supervised models on 3 annotated slices and evaluate them on the remaining 184 slices. The best-performing model is trained with patches of size 256x256, cross-entropy loss, 2.5D inputs (11 channels in the z-dimension), random flip augmentation, and CTF deconvolution with the tomogram’s pixel size and target defocus as input parameters.

As shown in Table 3.1, the model achieves a DICE of 0.893 when trained on the original annotations, and 0.920 when trained on the improved annotations from MemBrainSeg. In either case, these values indicate strong performance, while also suggesting that higher quality training annotations can make a positive impact on performance. The model achieves a Surface DICE of 0.877 when trained on the original annotations, and 0.952 when trained on the improved annotations. When combined with visual inspection of the annotations and segmentations, this suggests that the gains in performance from the MemBrainSeg annotations come from its coverage of some missed membranes areas in the original annotations, as opposed to being due to mere annotation differences at the boundaries of membranes.

Ablations demonstrate that the 2.5D inputs are important, with DICE drops of 0.047 and 0.030 for the original and improved annotations, respectively. Removing data augmentation and CTF deconvolution also result in modest performance drops across both annotation sources. As shown in Figure 3.2, visually, 2.5D inputs and CTF deconvolution result in lesser segmentation of noise. The former likely encourages the model to segment pixels that appear consistent across adjacent slices, while the latter likely reduces the delocalization of membrane signal.

Performance is higher for the outer mitochondrial membrane compared to the inner mitochondrial membrane for both training annotations, though the gap is more pronounced in the original annotations. The improved annotations result in a higher overall performance mostly due to higher performance for the inner mitochondrial membrane. A

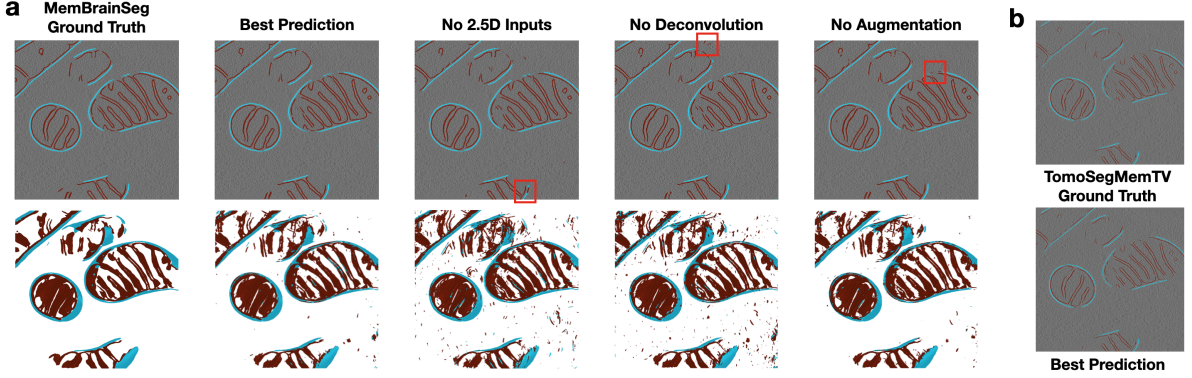


Figure 3.2: **Supervised segmentation visualizations on one mitochondrial tomogram.** (a) Ground truth and predictions with ablations for the improved MemBrainSeg annotations. Areas where model ablations degrade performance are marked with red boxes. (b) Original TomoSegMemTV annotations and corresponding prediction of the best-performing model.

Model	MemBrainSeg Annotations			TomoSegMemTV Annotations		
	DICE $\uparrow$	Surface DICE $\uparrow$	Hausdorff-95 $\downarrow$	DICE $\uparrow$	Surface DICE $\uparrow$	Hausdorff-95 $\downarrow$
Best	0.920	0.952	1.00	0.893	0.877	1.00
No 2.5D Inputs	0.890	0.934	1.41	0.846	0.826	1.29
No Deconvolution	0.912	0.941	1.31	0.873	0.856	1.00
No Augmentation	0.906	0.939	1.52	0.880	0.862	1.13

Table 3.1: **Supervised segmentation performance on one mitochondrial tomogram.** Performance metrics and ablations are presented for one tomogram of the CZII Mitochondria dataset, using both sets of training annotations.

key source of error across all models and annotation sources appears to be misclassification between the two membrane types. Performance for each class is given in Appendix Table B2.

3.4.3 Generalization

We split each of the two datasets into 4 random groups (7 per group for CZII Mitochondria and 4 per group for PolNet Synthetic). We perform 4-fold cross-validation in which we train a model on one group of tomograms and evaluate on all groups. For each training tomogram, we use 3 annotated slices for CZII Mitochondria and 4 annotated slices for PolNet Synthetic. We emphasize the sparseness of the training annotations - for PolNet Synthetic, for example, we train on 4 tomograms $\times$ 4 slices = 16 slices, whereas the total dataset consists of 16 tomograms $\times$ 250 slices = 4000 slices. The CZII Mitochondria

Dataset	Model	DICE - All	DICE - Train	DICE - Test
CZII Mitochondria	Self Supervised	0.742 (0.093)	-	-
	K-fold Supervised	0.605 (0.220)	0.815 (0.044)	0.543 (0.213)
	K-fold Unsupervised	0.166 (0.131)	0.297 (0.133)	0.128 (0.103)
PolNet Synthetic	Self Supervised	0.557 (0.059)	-	-
	K-fold Supervised	0.583 (0.062)	0.595 (0.058)	0.579 (0.062)
	K-fold Unsupervised	0.379 (0.076)	0.385 (0.072)	0.378 (0.077)

Table 3.2: **Generalization results.** Average DICE Scores for generalization experiments on both datasets. *Self Supervised* is a baseline in which a separate model is trained and tested on each tomogram individually. Standard deviations are indicated in parentheses.

tomograms consist of between 120 and 500 slices. In addition, since we aim to probe generalization performance from seen to unseen tomograms, we also train a separate model on each tomogram as a baseline for comparison.

Single tomogram models are trained with the settings described in Section 3.4.2. For generalization experiments, supervised models are trained with patches of size 256×256 , cross-entropy loss, 2.5D inputs (11 channels in the z-dimension) and data augmentation (flips, elastic transform, cutouts, rotations, shearing, and scaling). We find that CTF deconvolution and more complex transforms like Fourier spectrum matching lead to higher performance for some tomograms and lower for others. Thus, we opt for simple settings, though future work should explore optimal hyperparameter settings in the multi-tomogram case. Unsupervised models are trained on the same image patches as the supervised models, without any additional training strategies.

As shown in Table 3.2, on the CZII Mitochondria dataset, the single-tomogram baseline achieves an average DICE of 0.742 while the 4-fold supervised generalization yields a DICE of 0.605. There is a large train-test gap - the DICE for tomograms within the training set is 0.815 but only 0.543 for the unseen set. Note that the former is higher than the single-tomogram baseline, indicating that when the model has seen a few slices of a tomogram, seeing slices from other tomograms as well can boost performance, but when the model has not seen a tomogram, it performs much worse. There is also higher variance amongst the unseen tomograms, with a standard deviation of 0.213 in DICE, compared to 0.044 for the training tomograms. In contrast, on the PolNet Synthetic

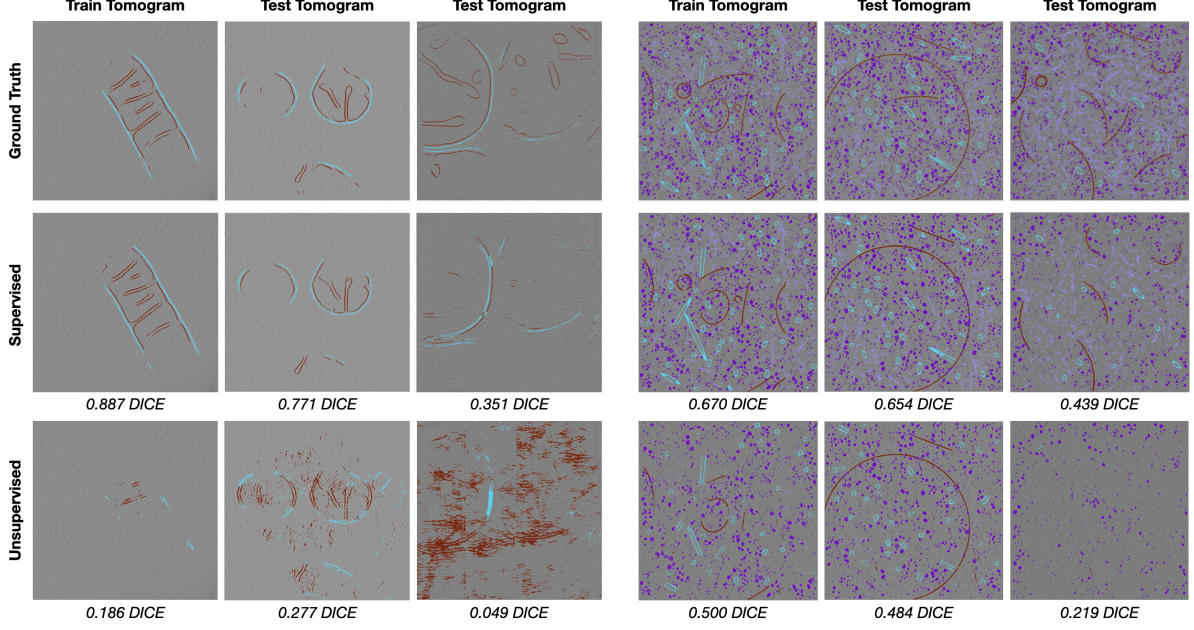


Figure 3.3: **Generalization results.** Ground truth, supervised model predictions, and unsupervised predictions for both datasets. Shown for each dataset is one training tomogram, one test tomogram on which the models perform relatively well, and one test tomogram on which the models perform relatively poorly.

dataset, the single-tomogram baseline achieves an average DICE of 0.557 while the 4-fold supervised generalization yields a higher DICE of 0.583. This rise in performance occurs across both training and test tomograms, with average scores of 0.595 and 0.579, respectively. Visually, there is less structural variation in the synthetic tomograms compared to the real mitochondria tomograms, suggesting that the model might leverage shared information better on the former. We also observe that amongst synthetic tomograms, performance is positively correlated with signal-to-noise ratio (Appendix Figure B3).

The unsupervised model produces poorer performance, scoring only 0.166 on the CZII Mitochondria dataset and 0.379 on the PolNet Synthetic dataset. Once again, there is a large train-test gap for the CZII Mitochondria dataset, with 0.297 DICE on training tomograms against 0.128 DICE on unseen tomograms, whereas this gap nearly disappears for the PolNet Synthetic dataset, with 0.385 DICE on training tomograms against 0.378 on unseen tomograms. Looking at class-wise scores reveals that the model performs similarly on both classes of membranes in the CZII Mitochondria dataset, but that there large inter-class variation on the PolNet Synthetic dataset. In particular, the

model achieves 0.523 DICE on proteins (against a single-tomogram baseline of 0.643), yet only a 0.057 DICE on actin (against a single-tomogram baseline of 0.409). Class-wise scores are shown in Tables B3 and B4. Improvements to the approach may come from an improved diffusion model (sample quality metrics shown in Table B5), or an improved segmentation model.

3.5 Summary and Future Work

We present a user-friendly tool for supervised training and segmentation of cryo-electron tomograms, in addition to benchmark analysis on real and synthetic datasets. Additionally, we present preliminary results on unsupervised approaches that learn from unlabelled data. Results show that while methods perform well in segmenting a tomogram when trained on a few annotated slices of that tomogram, they often generalize poorly to unseen tomograms containing the same structures. This underscores the need for further research into methods that can scalably segment large cryo-ET datasets with minimal supervision.

Ongoing work aims to push the limits of the unsupervised approach explored here. In addition, an observation is that manual annotations are often ambiguous or erroneous, as seen in the CZII mitochondria dataset [58]. Future approaches may benefit from incorporating techniques that learn from noisy labels or output uncertainties over the pixel-wise predictions.

Chapter 4

Summary

In this work, we presented new tools to aid in the analysis of cryo-electron microscopy (cryo-EM) and cryo-electron tomography (cryo-ET) data. First, we introduced CryoBench, a suite of novel synthetic datasets and metrics for quantitative evaluation of cryo-EM reconstruction methods. Then, we explored the semantic segmentation task in cryo-ET, presenting supervised and unsupervised baselines on benchmark datasets.

As the abundance and quality of cryo-EM and cryo-ET data grow, there is an increasing need for novel algorithms that can uncover the wealth of biological information present in these data. In coming years, we aim to push the limits of new methods that can reliably recover structural ensembles from highly heterogeneous cryo-EM datasets. Likewise, we aim to turn cryo-ET segmentation from a visual tool into a quantitative tool, which will require scalable algorithms that can process datasets containing hundreds of tomograms. Finally, there are numerous exciting research directions not discussed here, such as analysis workflows connecting cryo-EM and cryo-ET. We hope that these future developments in the field accelerate the pace and breadth of biological discovery.

Appendix A: Supplemental Information for Chapter 2

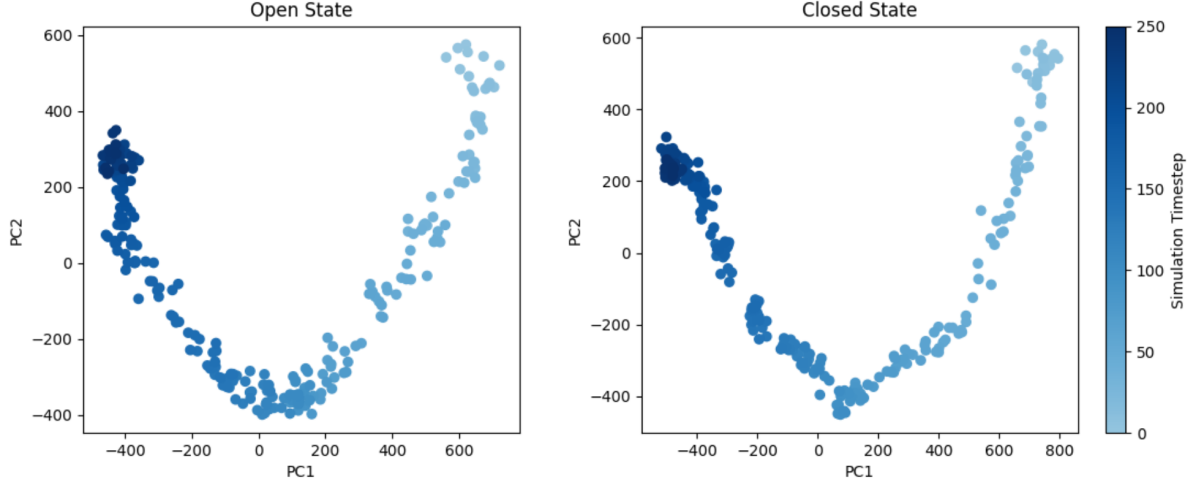


Figure A1: **Collective variables for MD conformations.** PCA is applied on the open and closed state MD trajectories, separately, to define 2D collective variables for the conformations in each.

Model	Open	Closed
3D Class (<i>fixed pose</i>)	-	-
3DVA	0.628	0.768
3DFlex	0.756	0.651
CryoDRGN (<i>fixed pose</i>)	0.837	0.816
DRGN-AI (<i>fixed pose</i>)	0.831	0.922
Opus-DSD	0.837	0.599
3D Class (<i>ab initio</i>)	-	-
CryoDRGN (<i>ab initio</i>)	0.069	0.050
DRGN-AI (<i>ab initio</i>)	0.748	0.672

Table A1: **Spearman rank correlation for open and closed MD states separately.**

Appendix B: Supplemental Information for Chapter 3

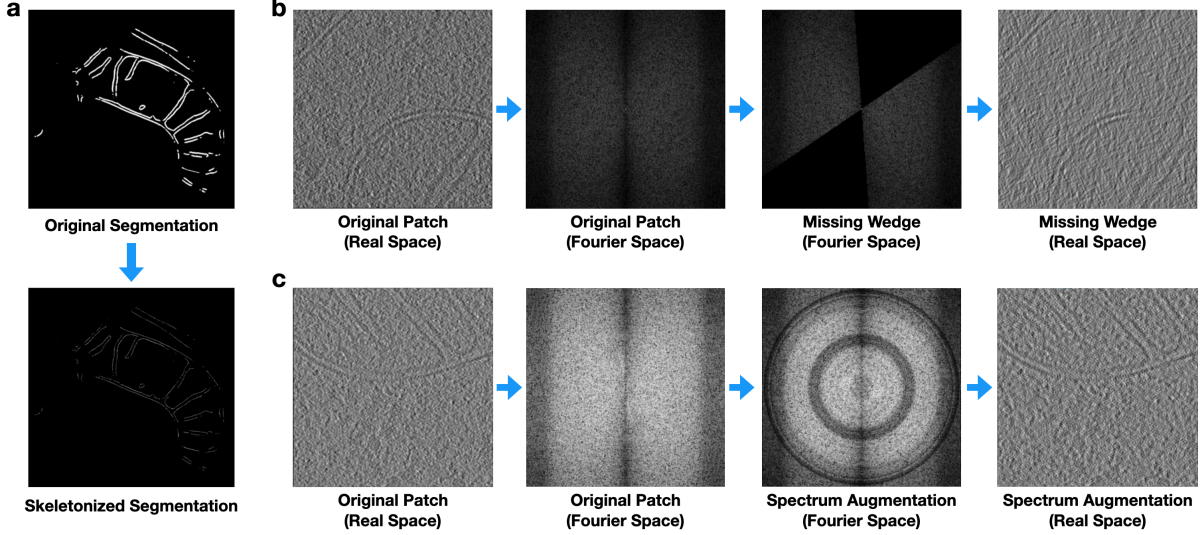


Figure B1: **Training options in the supervised segmentation tool.** (a) Surface DICE loss uses “skeletonized” ground truth and predicted segmentations to emphasize membrane morphology. (b) Missing wedge augmentation artificially imposes the cryo-ET missing wedge artifact in the Fourier space of images. (c) Fourier spectrum augmentation emulates images from one tomogram in the style of others by scaling Fourier components.

Model	FID ↓	KID ↓
LoRA-finetuned Stable Diffusion	205.17	0.195 (0.016)
DreamBooth-finetuned Stable Diffusion	131.04	0.143 (0.016)
DDPM	96.63	0.098 (0.010)
Unconditional Latent Diffusion	141.88	0.169 (0.017)

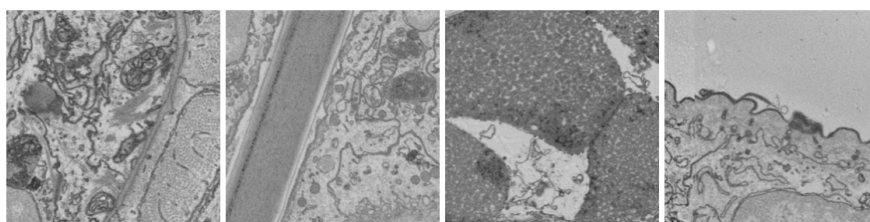
Table B1: **Sample quality metrics for diffusion models trained on VEM data.** 10k real images and 10k generated images were used to compute FID and KID.

Model	MemBrainSeg Annotations			TomoSegMemTV Annotations		
	Inner Membrane	Outer Membrane	Overall	Inner Membrane	Outer Membrane	Overall
Best	0.917	0.927	0.920	0.884	0.912	0.893
No 2.5D Inputs	0.886	0.902	0.890	0.830	0.882	0.846
No Deconvolution	0.909	0.921	0.912	0.864	0.894	0.873
No Augmentation	0.905	0.907	0.906	0.874	0.894	0.880

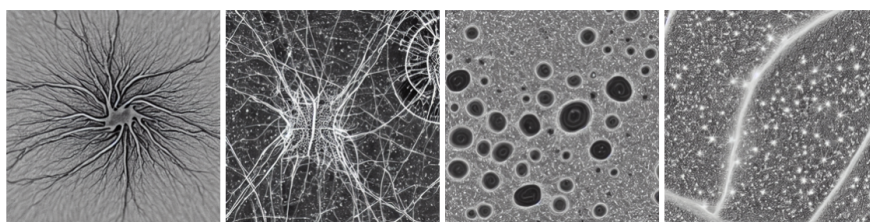
Table B2: **Class-wise single tomogram results.** Average DICE Scores per class for ablations on a single tomogram from the CZII Mitochondria dataset.

Model	Inner Membrane	Outer Membrane	Overall
Self Supervised	0.726 (0.108)	0.768 (0.098)	0.742 (0.093)
K-fold Supervised	0.597 (0.222)	0.618 (0.228)	0.605 (0.220)
K-fold Unsupervised	0.145 (0.140)	0.183 (0.144)	0.166 (0.131)

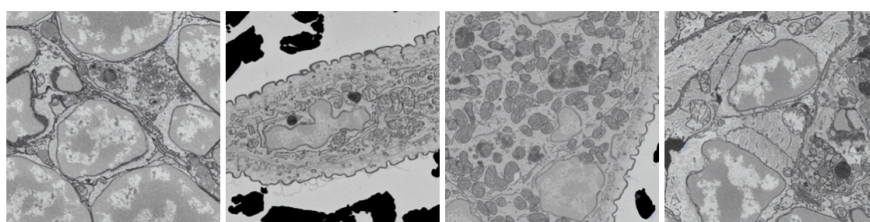
Table B3: **Class-wise generalization results on CZII Mitochondria.** Average DICE Scores per class for generalization experiments on the CZII Mitochondria dataset. Standard deviations are indicated in parentheses.



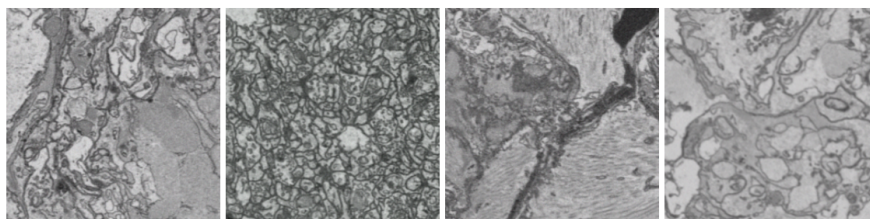
Real Images



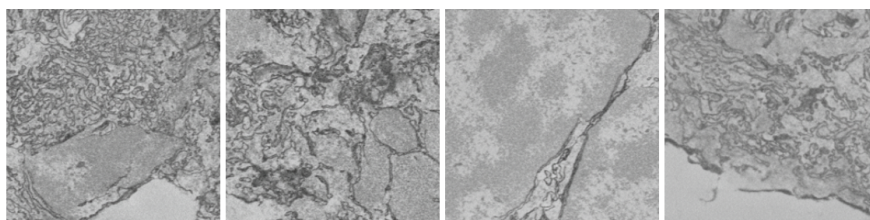
LoRA-finetuned Stable Diffusion
(Prompt: "An electron microscopy image")



DreamBooth-finetuned Stable Diffusion
(Prompt: "sks electron microscopy image")



Denoising Diffusion Probabilistic Model [DDPM]



Latent Diffusion
(KL-VAE autoencoder)

Figure B2: Sample generations from diffusion models trained on VEM data.

Model	Membrane	Microtubule	Actin	Protein	Overall
Self Supervised	0.495 (0.100)	0.557 (0.090)	0.409 (0.065)	0.643 (0.044)	0.557 (0.059)
K-fold Supervised	0.560 (0.095)	0.530 (0.088)	0.470 (0.064)	0.659 (0.050)	0.583 (0.062)
K-fold Unsupervised	0.272 (0.110)	0.250 (0.117)	0.057 (0.030)	0.523 (0.061)	0.379 (0.076)

Table B4: **Class-wise generalization results on PolNet Synthetic.** Average DICE Scores per class for generalization experiments on the PolNet Synthetic dataset. Standard deviations are indicated in parentheses.

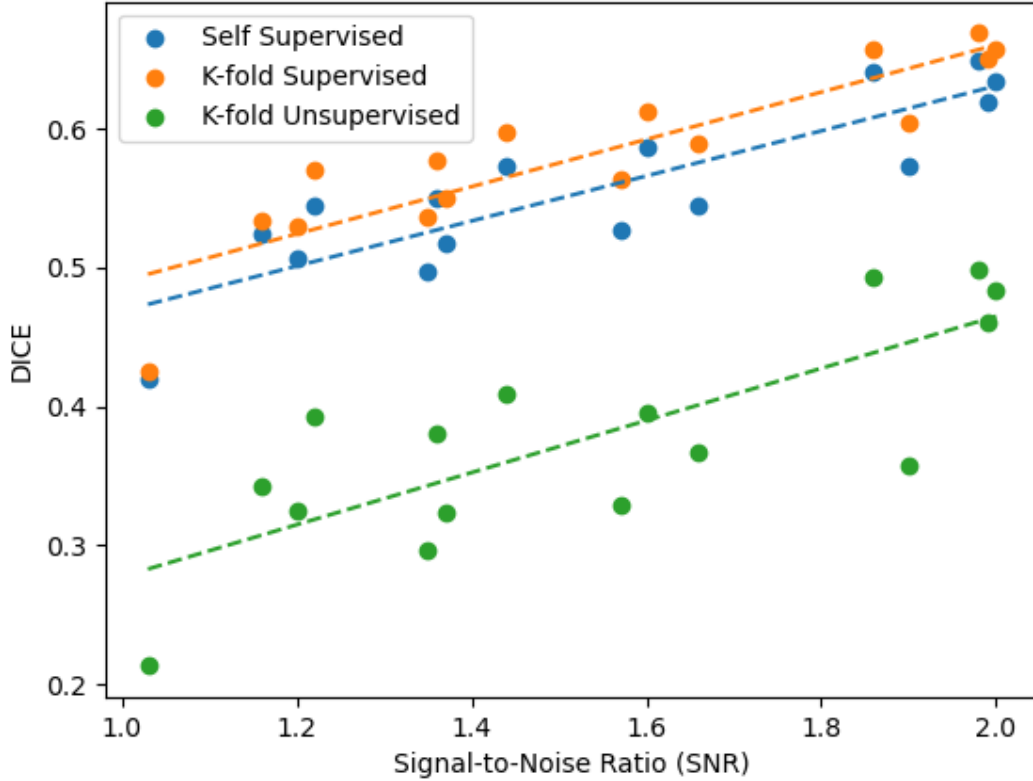


Figure B3: **Performance vs noise level for each tomogram in the PolNet Synthetic dataset.** For the K-fold generalization experiments, DICE scores are averaged over the K experiments. Dotted lines indicate linear regression lines.

Original Patch Size	Input Patch Size	FID ↓	KID ↓
128	128	47.49	0.043 (0.009)
256	128	78.70	0.093 (0.016)
256	256	21.24	0.010 (0.004)

Table B5: **Sample quality metrics for DDPM trained on patches of CZII Mitochondria tomograms.** 10k real images and 10k generated images were used to compute FID and KID. Note that all images are resized to 299x299 when passed to the InceptionV3 network.

Bibliography

- [1] Mahamid J. Nogales, E. Bridging structural and cell biology with cryo-electron microscopy. *Nature*, 628:47–56, 2024.
- [2] Chari A. Ciferri C. et al. Renaud, JP. Cryo-em in drug discovery: achievements, limitations and prospects. *Nat Rev Drug Discov*, 17:471–492, 2018.
- [3] Tortorici MA Wall A McGuire AT Veerler D. Walls AC, Park YJ. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 181(2):281–292, 2020.
- [4] Brown A Fernandez IS Hanssen E Condron M Tan YH Baum J Scheres SH Wong W, Bai XC. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *Elife*, 3, 2014.
- [5] Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. *FEBS Letters*, 594(20):3243–3261, 2020.
- [6] Yury S Bykov, Miroslava Schaffer, Svetlana O Dodonova, Sahradha Albert, Jürgen M Plitzko, Wolfgang Baumeister, Benjamin D Engel, and John AG Briggs. The structure of the copi coat determined within the cell. *eLife*, 6:e32493, nov 2017.
- [7] Helen R Saibil. Cryo-em in molecular and cellular biology. *Mol Cell*, 82(2):274–284, 2022.
- [8] Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, and Bonnie Berger. Reconstructing continuous distributions of 3d protein structure from cryo-em images. In *International Conference on Learning Representations*, 2020.

- [9] Fleet D. Punjani, A. 3d variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-em. *Journal of Structural Biology*, 213, 2021.
- [10] Fleet D.J. Punjani, A. 3dflex: determining structure and motion of flexible proteins from cryo-em. *Nat Methods*, 20, 2023.
- [11] Poitevin F Zhong ED Miolane N Donnat C, Levy A. Deep generative modeling for volume reconstruction in cryo-electron microscopy. *J Struct Biol*, 214, 2022.
- [12] Heterogeneity in cryo-electron microscopy. <https://www.simonsfoundation.org/flatiron/center-for-computational-biology/structural-and-molecular-biophysics-collaboration/heterogeneity-in-cryo-electron-microscopy/>. Accessed: 2024-04-29.
- [13] David J Fleet Robert A McLeod RJ Dwayne Miller Jane Y Howe Bin Shi, Kevin Zhang. Deep generative priors for biomolecular 3d heterogeneous reconstruction from cryo-em projections. *Journal of Structural Biology*, 216, 2024.
- [14] Erica Ollmann Saphire, Paul W. H. I. Parren, Ralph Pantophlet, Michael B. Zwick, Garrett M. Morris, Pauline M. Rudd, Raymond A. Dwek, Robyn L. Stanfield, Dennis R. Burton, and Ian A. Wilson. Crystal structure of a neutralizing human igg against hiv-1: A template for vaccine design. *Science*, 293(5532):1155–1159, 2001.
- [15] Gilad Ofek, Min Tang, Anna Sambor, Hermann Kattinger, John R. Mascola, Richard Wyatt, and Peter D. Kwong. Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2f5 in complex with its gp41 epitope. *Journal of Virology*, 78(19):10724–10737, 2004.
- [16] Elena Gustchina, Mi Li, Rodolfo Ghirlando, Peter Schuck, John M. Louis, Jason Pierson, Prashant Rao, Sriram Subramaniam, Alla Gustchina, G. Marius Clore, and Alexander Wlodawer. Complexes of neutralizing and non-neutralizing affinity matured fabs with a mimetic of the internal trimeric coiled-coil of hiv-1 gp41. *PLOS ONE*, 8(11):1–15, 11 2013.

- [17] Shuo Du, Yunlong Cao, Qinyu Zhu, Pin Yu, Feifei Qi, Guopeng Wang, Xiaoxia Du, Linlin Bao, Wei Deng, Hua Zhu, Jiangning Liu, Jianhui Nie, Yinghui Zheng, Haoyu Liang, Ruixue Liu, Shuran Gong, Hua Xu, Ayijiang Yisimayi, Qi Lv, Bo Wang, Runsheng He, Yunlin Han, Wenjie Zhao, Yali Bai, Yajin Qu, Xiang Gao, Chenggong Ji, Qisheng Wang, Ning Gao, Weijin Huang, Youchun Wang, X. Sunney Xie, Xiaodong Su, Junyu Xiao, and Chuan Qin. Structurally resolved sars-cov-2 antibody shows high efficacy in severely infected hamsters and provides a potent cocktail pairing strategy. *Cell*, 183(4):1013–1023.e13, 2020.
- [18] Johannes F. Scheid, Christopher O. Barnes, Basak Eraslan, Andrew Hudak, Jennifer R. Keefe, Lisa A. Cosimi, Eric M. Brown, Frauke Muecksch, Yiska Weisblum, Shuting Zhang, Toni Delorey, Ann E. Woolley, Fadi Ghanous, Sung-Moo Park, Devan Phillips, Betsabeh Tusi, Kathryn E. Huey-Tubman, Alexander A. Cohen, Priyanthi N.P. Gnanapragasam, Kara Rzasa, Theodora Hatziioanno, Michael A. Durney, Xiebin Gu, Takuya Tada, Nathaniel R. Landau, Anthony P. West, Orit Rozenblatt-Rosen, Michael S. Seaman, Lindsey R. Baden, Daniel B. Graham, Jacques Deguine, Paul D. Bieniasz, Aviv Regev, Deborah Hung, Pamela J. Bjorkman, and Ramnik J. Xavier. B cell genomics behind cross-neutralization of sars-cov-2 variants and sars-cov. *Cell*, 184(12):3205–3221.e24, 2021.
- [19] Structural basis of spike rbm-specific human antibodies counteracting broad sars-cov-2 variants. *Commun Biol*, 6(395), 2023.
- [20] Molecular dynamics simulations related to sars-cov-2. In *D. E. Shaw Research Technical Data*, 2020.
- [21] Jeon et al. Cryobench: Datasets and benchmarks for the heterogeneous reconstruction task in cryo-em. In *In submission*, 2024.
- [22] Dror RO Hollingsworth SA. Molecular dynamics simulation for all. *Neuron*, 99, 2018.

- [23] Orozco M Cossio P Wieczór M, Tang PK. Omicron mutations increase interdomain interactions and reduce epitope exposure in the sars-cov-2 spike. *iScience*, 26, 2023.
- [24] Elaine C. Meng, Thomas D. Goddard, Eric F. Pettersen, Greg S. Couch, Zach J. Pearson, John H. Morris, and Thomas E. Ferrin. Ucsf chimeraX: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023.
- [25] Baumeister W Saxton WO. The correlation averaging of a regularly arranged bacterial cell envelope protein. *J Microsc.*, 127, 1982.
- [26] Chen S Scheres SH. Prevention of overfitting in cryo-em structure determination. *Nat Methods*, 9, 2012.
- [27] Rubinstein J. Fleet D. et al. Punjani, A. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nat Methods*, 14, 2017.
- [28] Ellen D. Zhong, Adam Lerer, Joseph H. Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4046–4055, 2021.
- [29] Ni F. Wang Q. et al. Luo, Z. Opus-dsd: deep structural disentanglement for cryo-em single-particle analysis. *Nat Methods*, 20, 2023.
- [30] Grigorieff N Himes B. Cryo-tem simulations of amorphous radiation-sensitive samples using multislice wave propagation. *IUCrJ*, 8, 2021.
- [31] Rong Sun Bin Zhang Lei Qi Sakar Shivakoti Chong-Li Tian Peijun Zhang Pak-Ming Lau Z. Hong Zhou Guo-Qiang Bi Chang-Lu Tao, Yun-Tao Liu. Differentiation and characterization of excitatory and inhibitory synapses by cryo-electron tomography and correlative microscopy. *Journal of Neuroscience*, 38, 2018.
- [32] Shoh Asano Vladan Lucic Jose-Jesus Fernandez Antonio Martinez-Sanchez, Inmaculada Garcia. Robust membrane detection based on tensor voting for electron tomography. *J Struct Biol*, 186, 2014.

- [33] Lorenz Lamm, Simon Zufferey, Ricardo D. Righetto, Wojciech Wietrzynski, Kevin A. Yamauchi, Alister Burt, Ye Liu, Hanyi Zhang, Antonio Martinez-Sanchez, Sebastian Ziegler, Fabian Isensee, Julia A. Schnabel, Benjamin D. Engel, and Tingying Peng. Membrain v2: an end-to-end tool for the analysis of membranes in cryo-electron tomography. *bioRxiv*, 2024.
- [34] Sun SY, Jonasch D, He CY, Schmid MF, Chiu W, Ludtke SJ, Chen M, Dai W. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nat Methods*, 14, 2017.
- [35] Dragonfly. In *Comet Technologies Canada Inc*, 2022.
- [36] Amira. In *ThermoFisher Scientific*, 2024.
- [37] Goetz S.K. Mattausch A. et al. de Teresa-Trueba, I. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nat Methods*, 20, 2023.
- [38] Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderrick Murray-Smith, Ahmed Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, Frederick Klauschen, Bruno Sanguinetti, and Luis Oala. Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology, 2023.
- [39] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation, 2024.
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [41] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Dif-fumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models, 2024.

- [42] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models, 2023.
- [43] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles, 2021.
- [44] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model, 2023.
- [45] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.
- [46] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models, 2022.
- [47] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion, 2024.
- [48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [49] Bosch C. Bullen A. et al. Collinson, L.M. Volume em: a quiet revolution takes shape. *Nat Methods*, 20, 2023.
- [50] Hernando M Vergara, Constantin Pape, Kimberly I Meechan, Valentyna Zinchenko, Christel Genoud, Adrian A Wanner, Kevin Nzumbi Mutemi, Benjamin Titze, Rachel M Templin, Paola Y Bertucci, Oleg Simakov, Wiebke Dürichen, Pedro Machado, Emily L Savage, Lothar Schermelleh, Yannick Schwab, Rainer W Friedrich, Anna Kreshuk, Christian Tischer, and Detlev Arendt. Whole-body integration of gene expression and single-cell morphology. *Cell*, 184(18):4819—4837.e22, September 2021.

- [51] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.
- [53] Sørensen–dice coefficient. https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient. Accessed: 2024-04-29.
- [54] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.
- [55] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.
- [56] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021.
- [57] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.
- [58] Daniel Fuentes R. Luke Wiseman-Danielle A. Grotjahn Benjamin A. Barad, Michaela Medina. Quantifying organellar ultrastructure in cryo-electron tomography using a surface morphometrics pipeline. *J Cell Biol*, 222, 2023.
- [59] Jannis Ahlers, Daniel Althviz Moré, Oren Amsalem, Ashley Anderson, Grzegorz Bokota, Peter Boone, Jordão Bragantini, Genevieve Buckley, Alister Burt, Matthias Bussonnier, Ahmet Can Solak, Clément Caporal, Draga Doncila Pop, Kira Evans, Jeremy Freeman, Lorenzo Gaifas, Christoph Gohlke, Kabilar Gunalan, Hagai Har-Gil, Mark Harfouche, Kyle I. S. Harrington, Volker Hilsenstein, Katherine Hutchings, Talley Lambert, Jessy Lauer, Gregor Lichtner, Ziyang Liu, Lucy Liu, Alan

Lowe, Luca Marconato, Sean Martin, Abigail McGovern, Lukasz Migas, Nadalyn Miller, Hector Muñoz, Jan-Hendrik Müller, Christopher Nauroth-Kreß, Juan Nunez-Iglesias, Constantin Pape, Kim Pevey, Gonzalo Peña-Castellanos, Andrea Pierré, Jaime Rodríguez-Guerra, David Ross, Loic Royer, Craig T. Russell, Gabriel Selzer, Paul Smith, Peter Sobolewski, Konstantin Sofiuk, Nicholas Sofroniew, David Stansby, Andrew Sweet, Wouter-Michiel Vierdag, Pam Wadhwa, Melissa Weber Mendonça, Jonas Windhager, Philip Winston, and Kevin Yamauchi. napari: a multi-dimensional image viewer for Python, July 2023.

- [60] Antonio Martinez-Sanchez, Lorenz Lamm, Marion Jasnin, and Harold Phelippeau. Simulating the cellular context in synthetic datasets for cryo-electron tomography. *bioRxiv*, 2024.